Monitoring hate speech and the limits of current definition

Bahador, Babak

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Bahador, B. (2023). Monitoring hate speech and the limits of current definition. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 291-298). Berlin https://doi.org/10.48541/dcr.v12.17

Digital
Communication
Research.de

Recommended citation: Bahador, B. (2023). Monitoring hate speech and the limits of current definition. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 291–298). Digital Communication Research. https://doi.org/10.48541/dcr.v12.17

Abstract: Current definitions of hate speech are inadequate as the basis for monitoring hate speech targeted at groups. First, they do not capture escalating group-targeted negative speech which can be a precursor to more extreme forms of hate speech such as dehumanization, demonization, and incitement to violence. While not hate speech, such negative speech is an early warning that could be helpful for a hate speech monitoring system to track, as responses and interventions, especially to the offline harms of hate speech, can take time to operationalize. Second, current definitions of hate speech do not capture hateful rhetoric aimed at groups not traditionally included in hate speech definitions (those without immutable qualities), such as groups targeted for hate based on profession-based identity like journalists. This chapter presents some suggestions for addressing these issues, including a hate speech intensity scale.

DOI 10.48541/dcr.v12.17

Babak Bahador

Monitoring Hate Speech and the Limits of Current Definition

1 Introduction

Monitoring hate speech in order to prevent offline harms is a laudable goal that has proven largely elusive to date. This is due to a number of factors including the limits of current definitions, knowing exactly when such speech triggers offline harms, tracking hate speech in real time and creating and implementing effective interventions. This article is primarily focused on the first issue regarding how current definitions of hate speech can limit effective hate speech monitoring. The article begins by examining how hate speech is typically currently defined and some limitations that this poses for monitoring due to the restricted scope of the groups and language included. The article argues that these limits mean that some non-traditonal types of groups targeted with hate are excluded, even though they could also become victims of hate-based violence. Furthermore, it argues that hate speech monitoring should include language escalating to traditional hate speech content so early warning signs can be detected and action taken earlier. Once more extreme hate speech takes hold, it could also be a sign that it is too late to implement more peaceful preventative actions. Finally, the article introduces an hate speech intensity scale that includes early warning categories for hate speech monitoring.

2 The limits of defining hate speech

In its most blantant manifestations, hate speech is communication aimed at groups of people to dehumanize, demonize or incite violence against them. If hate against a particular out-group (the group targeted for hate speech) is successfully sold to an in-group (the group the hate speech attempts to persuade), then all members of the out-group are viewed as a negative stereotype, losing their individuality and humanity. In such scenarios, which are often driven by falsehoods and exaggerated fears about the out-group, retribution against all members of the out-group is justified as they all represent the same threat to the in-group (Bahador, 2012).

Most definitions of hate speech limit its targets to groups that hold immutable qualities such as a particular race, nationality, religion, ethnicity, gender, age bracket or sexual orientation (see Sponholz in this volume). However, in research that measured hate aimed at groups more broadly, findings showed that groups with immutable qualities were less frequently targeted versus other types of groups not usually included in hate speech definitions (Bahador, Kerchner et al., 2019). The first of these types of groups can be classified as professions and industries[1], with journalists and the media sector a primary and leading example. While professions and employment in particular industries is by choice (so not immutable), they nonetheless are groups that are distinguishable and a growing target of hate speech. The concern here is less about harassment of journalists for particular content they produce, but about attacks based on group identity, which makes it similar to other groups with immutable qualities.

Hate speech against journalists has grown notably over recent years, not least because journalists act as a check on authoritarian power, which has been a growing trend worldwide (Sulzberger, 2019). This is particularly the case for female journalists, who are under unprecedented levels of attack online and are targeted both for their gender and profession (The Guardian, 2021).

Another notable group typology frequently targeted for hateful rhetoric is foreign countries. While this is related to the traditional immutable category of nationality (e.g., Chinese), there are often negative references to the country

1 This definition excludes professions and industries that engage in violent or other malicious behavior.

itself (e.g., China) that can increase negative public sentiment towards the country (Brewer et al., 2003), and most importantly, its people and those associated with them. For example, references to China as the actor responsible for the Covid-19 global pandemic by certain political leaders is considered to be a key factor that led to a spike in hate crimes against Asians in the United States including American citizens from Asian descent (BBC News, 2021). Those political leaders did not mention anything about Chinese people or Americans of Asian descent. They mentioned China, and this is excluded from traditional definitions of hate speech (as it is a country, not race or nationality), showing that the current definitions are inadequate for addressing a serious problem.

While it is certainly appropriate to criticize foreign governments, those in influential positions, such as journalists and politicians, often inadvertently refer to the country without sufficiently delineating that their critique is of the government and its actions and policies, not the people. To avoid such conflation and its potential negative effects, those in power need to be precise and only refer to the foreign governments, government institutions and agencies or political leaders and not the country as a whole. When criticising states such as the United States or Russia, the criticism should be against the government specifically, for example, "the Russian government attacked," not "Russia attacked"; or "the U.S. military bombed this," not "the U.S. bombed this." The latter cases build hate towards people associated with the country; the former offer legitimate criticism of the government which should be subject to scrutiny.

In this research, four different hate-speech group typologies are distinguished: 1) immutable qualities (traditional hate speech groups), 2) professions and industries, 3) countries, and 4) "other," which captures other groups of people who are targeted for hate speech but otherwise excluded from the other three categories. These include groups such as "the elite," which are generally excluded in traditional definitions.[2] If one wants to capture the full breadth of hate-speech group targets, this more extensive approach will capture more hate speech.

2 Terms such as the "the elite" often refer to a variety of other groups but may mean different groups to different audiences. For example, in a survey of Americans, conservatives often considered the elite to be cultural, political and academic elite such as actors, politicians and professors, while liberal Americans thought elites were economic, industrial and financial elites (Bahador, Entman & Knüpfer, 2019).

3 Early warning to hate speech

When considering what type messaging should be considered hate speech, most definitions include language that attemts to demonize, dehumanize or incite violence against groups. Dehumanization is a tactic that depicts groups as less than human and usually involves associating them with sub-human creatures such as rats and cockroaches or non-human forms such as garbage or dirt. Alternatively, groups can be demonized, in which they become threatening super-human creatures such as monsters and demons, or equated to fatal threats such as cancer or a virus. When presented this way, the elimination of such groups becomes beneficial and desirable, as removing them takes away a perceived threat to the in-group (Dower, 1986; Keen, 1991; Carruthers, 2011; Bahador 2015). Calls to attack, harm or kill groups—often the same ones that were dehumanized and demonized—is incitement, which is another central type of hate speech. Incitement is often the most extreme type of hate speech content. Even in the United States, in which hate speech is generally protected on free speech grounds under the First Amendment of the Constitution, it is still a crime to incite "immiment lawless action" if likely to occur within a short period of time (Tucker, 2015).

As with the previous concern over limiting hate speech groups to only ones with immutable qualities (and missing other groups that are also subject to hate), it is also problematic to restrict hate speech definitions to only the most severe types (dehumanization, demonization and incitement) if other speech builds up hate and disdain towards groups. Hate in the context of hate speech, after all, can be defined as a human emotion that is triggered through exposure to a particular type of information. When it emerges, this emotion involves an enduring dislike for a group, a loss of empathy for them, and a desire for harm against them (Waltman & Mattheis, 2017). However, there is no reason to assume that other types of negative speech against groups also do not create hate. At its root, hate against groups begins when an us-them dynamic is created and a different group is differentiated from your own and negative actions and characteristics are allocated to them, coming over time to define the group and all its members as a negative stereotype. But even if negative words towards groups such as insults do not constitute hate speech, it is an early warning that should be addressed before it becomes acceptable and builds tolerance for more extreme forms of speech. In any hate speech monitoring system, it is important to have early warning and

not just activate the system when it has become dangerous. As such, incorporating language that can be considered as early warning on the road to hate speech should be an important consideration of any monitoring system.

4 Hate speech intensity scale

To monitor hate speech in a way that incorporates both a broader set of group categories and an early warning element, a hate speech intensity scale is proposed, demonstrating the escalating nature of hate speech content. To make it easy to follow, colors, numbers, titles, descriptions and examples are provided in Table 1. Furthermore, a distinction is made between how groups are characterized (referred to as "rhetoric") and "responses" or actions the in-group is recommended to take against the out-group.

The hate speech intensity scale has six categories. Categories 1 to 3 are referred to as "early warning." Category 4 is dehumanization/demonization, and categories 5 and 6 are associations with or calls for violence (#5) and death (#6). The following section goes through the 6 categories in more detail.

In this scale, the first early warning category involves disagreement with groups. While there is nothing wrong with disagreement in principle, and it can be argued that it is essential to democracy, the exercise does involve the creation of an us versus them dynamic, with "them" being viewed collectively in a negative light. Also, there is likely some misinformation involved in such a claim against a group, as rarely will an entire group hold similar views and beliefs. Collectivizing their views or beliefs, therefore, is likely to miss important differences amongst group members. By itself, such rhetoric is likely not hateful, and thus, the green color designation indicating it is safe to proceed (as per a traffic light). However, it is something to start monitoring for the purposes of a monitoring system.

The second early warning category involves language that blames a group for particular negative actions, often carried out by one or a few members. However, there is a tendency to blame the entire out-group in such scenarios for the negative actions of a few. This category includes non-violent negative actions, such as claims that the group stole or withdrew from a positive event. When such alleged actions are ambiguous on the use of violence (e.g., they stopped them) or use of non-violent

negative metaphors, they fit in this category (if unambiguous on violence, it's classified a 5). Responses involve non-violent actions the in-group should do towards the out-group, such as voting them out or protesting against them.

Table 1: Hate speech intensity scale

	Title	Description	Examples
	6. Death	Rhetoric includes literal killing by group. Responses include the literal death/elimination of a group.	Killed, annihilate, destroy
	5. Violence	Rhetoric includes infliction of physical harm or metaphoric/ aspirational physical harm or death. Responses include calls for literal violence or metaphoric/aspirational physical harm or death.	Punched, raped, starved, torturing, mugging
	4. Demonizing and Dehumanizing	Rhetoric includes sub-human and superhuman characteristics. There are no responses for #4.	Rat, monkey, Nazi, demon, cancer, monster
	3. Negative character	Rhetoric includes non-violent characterizations and insults. There are no responses for #3.	Stupid, thief, aggressor, fake, crazy
	2. Negative actions	Rhetoric includes negative non-violent actions associated with the group. Responses include non-violent actions including metaphors.	Threatened, stole, outrageous act, poor treatment, alienate
	1. Disagreement	Rhetoric includes disagreeing at the idea/belief level. Responses include challenging claims, ideas, beliefs, or trying to change their view.	False, incorrect, wrong, challenge, persuade, change minds

The third early warning typology includes negative characterizations or insults. This is worse than just negative non-violent actions, as it makes an intrinsic claim about the group as opposed to a one-off action claim. As this category is not action oriented (unlike #1, 2, 5 and 6), there are no responses. The fourth category

is also the second typology and can be considered an extreme form of negative characterization involving dehumanization and/or demonization. Like the third category, there are no responses in this category.

The fifth and sixth categories are part of the third and most intense typology, involving violent actions and death. The fifth category refers to literal violence allocated to the out-group either in their past, present or future. This also includes metaphorical or aspirational violence that is either nonlethal or lethal. Responses call for literal non-lethal violence towards the out-group such as assaulting them. The sixth category involves rhetorically referring to the out-group as killers (past, present and future). Responses call for our side to kill the out-group.

To monitor hate speech effectively, it is important to not miss any notable groups targeted for hate (such as journalists), even if they don't fit into traditional hate speech definitions. Furthermore, it is critical to see how hate speech builds up before it starts to be more harmful with stronger rhetoric. To this end, the hate speech intensity scale offers a tool that could be operationalized to monitor hate speech. In early experimentation using this tool to monitor hate speech from leading media personalities in the U.S., we found that about half of all hate speech is against journalists and the media (Bahador, Kerchner et al., 2019). We also found few examples of more extreme speech (#4, 5 and 6 on the scale) representing less than 5% of all cases using this scale. However, #2 and 3 (negative actions and characteristics allocated to groups) were prevalent, accounting the vast majority of cases.

Babak Bahador is Research Professor at the School of Media and Public Affairs (SMPA) at George Washington University, USA, and a Senior Fellow at the University of Canterbury in New Zealand. https://orcid.org/0000-0001-7872-9764

References

Bahador, B. (2012). Rehumanizing enemy images: Media framing from war to peace. In K. V. Korostelina (Ed.), *Forming a culture of peace: Reframing narratives of intergroup relations, equity and justice* (pp. 195–211). Palgrave Macmillan.

Bahador, B. (2015). The media and deconstruction of the enemy image. In V. Hawkins & L. Hoffmann (Eds.), *Communication and peace: Mapping an emerging field* (pp. 120–132). Routledge.

Bahador, B., Kerchner, D., Bacon, L., & Menas, A. (2019). Monitoring hate speech in the US Media. Working Paper. Media and Peacebuilding Project. George Washington University. https://cpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/8/846/files/2019/03/Monitoring-Hate-Speech-in-the-US-Media-3_22-z0h5kk.pdf

Bahador, B., Entman, R., & Knüpfer C. (2019). Who's elite and how the answer matters to politics. *Political Communication, 36*(1), 195–202. https://doi.org/10.1080/10584609.2018.1548412

BBC News (2021, May 21). Covid 'hate crimes' against Asian Americans on rise. https://www.bbc.com/news/world-us-canada-56218684

Brewer, P. R., Joseph, G., & Willnat, L. (2003). Priming or framing: Media influence on attitudes toward foreign countries. *International Communication Gazette, 65*(6), 493–508. https://doi.org/10.1177/0016549203065006005

Carruthers, S. (2011). *The media and war*. Palgrave MacMillan.

Dower, J. W. (1986). *War without mercy: Race and power in the Pacific War*. Pantheon Books.

Keen, S. (1991). *Faces of the enemy: Reflections on the hostile imagination*. Harper & Row.

Sulzberger, A. G. (2019, September 23). The growing threat to journalism around the world. *New York Times*. https://www.nytimes.com/2019/09/23/opinion/press-freedom-arthur-sulzberger.html

The Guardian (2021, May 9). The Guardian view on online abuse of female journalists: a problem for all. *Editorial*. https://www.theguardian.com/commentisfree/2021/may/09/the-guardian-view-on-online-abuse-of-female-journalists-a-problem-for-all

Tucker, E. (2015, December 31). How federal law draws a line between freedom of speech and hate crimes. *PBS News Hour*. https://www.pbs.org/newshour/nation/how-federal-law-draws-a-line-between-free-speech-and-hate-crimes

Waltman, M. S., & Mattheis, A. (2017). Understanding hate speech. In *Oxford encyclopedia of communication*. https://doi.org/10.1093/acrefore/9780190228613.013.422