Sharing is caring: Addressing shared issues and challenges in hate speech research

Paasch-Colberg, Sünje; Strippel, Christian; Emmer, Martin; Trebbe, Joachim

Digital
Communication
Research.de

Recommended citation: Paasch-Colberg, S., Strippel, C., Emmer, M., & Trebbe, J. (2023). Sharing is caring: Addressing shared issues and challenges in hate speech research. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 11–22). Digital Communication Research. https://doi.org/10.48541/dcr.v12.1

Abstract: This book is the result of a conference that could not take place. It is a collection of 26 texts that address and discuss the latest developments in international hate speech research from a wide range of disciplinary perspectives. This includes case studies from Brazil, Lebanon, Poland, Nigeria, and India, theoretical introductions to the concepts of hate speech, dangerous speech, incivility, toxicity, extreme speech, and dark participation, as well as reflections on methodological challenges such as scraping, annotation, datafication, implicity, explainability, and machine learning. As such, it provides a much-needed forum for cross-national and cross-disciplinary conversations in what is currently a very vibrant field of research.

DOI 10.48541/dcr.v12.1

Sünje Paasch-Colberg, Christian Strippel,
Martin Emmer & Joachim Trebbe

Sharing is Caring

Addressing shared issues and challenges in hate speech research

1 Introduction

This book is in some way an unplanned outcome of a research project that we worked on in the past five years.[1] When we started in October 2017, online hate speech had been an increasingly important issue in both public and academia for quite some time already. However, our project coincided with a socially and politically turbulent time, which challenged hate speech research and called for an increased exchange in the field. For example, the Network Enforcement Act came into force in Germany at that time. This law not only caused debate about how to identify criminal content in the volatile interactive spaces of the Internet and about who should be responsible for regulating these spaces, but it has also been

[1] The interdisciplinary research project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners" was funded by the German Federal Ministry of Education and Research [grant number: 01UG1735AX]. It brought together communication scholars from Freie Universität Berlin, computer scientists from the Berliner Hochschule für Technik, and computer linguists from VICO Research & Consulting.

used to justify the introduction of restrictive social media laws in autocratic states and flawed democracies. Thus, it renewed questions about contextual factors in our thinking about norms and boundaries in public debates.

Other examples that strongly affected our and others' research were Facebook's decision to restrict its API after the Cambridge Analytica scandal came to light in early 2018, and the General Data Protection Regulation (GDPR) of the European Union, which unsettled many blog operators, and eventually led to the closure of their comment sections.

To respond to these developments and the implications they had for our research, we invited a group of colleagues working on similar topics to a workshop at the Berlin Weizenbaum Institute in 2019 to share experiences with common theoretical, conceptual, and methodological issues in our field of research.[2] We discussed questions around data collection, protection and exchange, identification and classification of norm-transgressive user-generated content, as well as data analysis and automation. One important outcome of this workshop was the realization that considering perspectives from different political and cultural contexts, as well as from different academic disciplines, is crucial to better understand hate speech as a global and multifaceted phenomenon. Furthermore, the exchange confirmed how important debates around theoretical concepts and definitions are for the growing and transdisciplinary field of research on hate speech.

With the aim to further address these points together with a broader group of people, we planned an international and interdisciplinary conference on hate speech analysis for mid March 2020 in Berlin. In addition to a few invited presentations, our main idea for this conference was to provide space and opportunities for in-depth discussions and exchange among the participants. The large number of registrations we received from scholars from many different countries and disciplines showed that there is indeed a great interest in such discursive formats within the community of hate speech researchers. Unfortunately, the conference had to be canceled a few weeks before it was scheduled due to the onset of the COVID-19 pandemic.

[2] The participants of the workshop were (in alphabetic order): Arndt Allhorn, Chris Biemann, Svenja Boberg, Ines Engelmann, Katharina Esau, Annett Heft, Dominique Heinbach, Jakob Jünger, Tim König, Constanze Kuechler, Sebastian Kuehn, Laura Laugwitz, Wiebke Loosen, Alexander Löser, Hanna Marzinkowski, Teresa Naab, Pablo Porten-Cheé, Cornelius Puschmann, Liane Reiners, Susanne Reinhardt, Diana Rieger, Julian Risch, Tim Schatto-Eckrodt, Anke Stoll, Betty van Aken, and Marc Ziegele.

Since we were determined that it was important to provide a forum for research-related discussions amongst hate speech scholars, we decided to organize this volume and reached out to a number of scholars, who had registered for our then-canceled conference, as well as colleagues from the closer environment of our research project to contribute. To do justice to all the discussions we have missed in the panels and coffee breaks of the conference, we asked these colleagues for short programmatic papers that question current research threads, point out new ways, and give impulses for future research. In addition, we invited texts that respond to these papers as well as discuss and contextualize them in relation to each other.

To our great pleasure, almost all colleagues accepted our invitation, and those who did liked the assignment, confirming that they too see a need for this kind of exchange. As a result, we could realize an even more diverse authorship and hopefully have a bigger outreach than the conference would have been able to. We are excited that, with a total of 26 chapters, we can now cover a wide range of topics that contribute to the field of hate speech research by (1) focusing on recent research and policy developments in countries that are less visible in literature, (2) discussing the multiplicity of theoretical concepts, definitions, and measurements, and (3) presenting new approaches of interdisciplinary research and machine learning that come with new questions, challenges, and implications.

2 Political perspectives: Current issues and developments

The first section of this volume opens with contributions dedicated to the foundations of hate speech research. One of these foundations is that the assessment of speech as hate speech is context-dependent, for example, with respect to the legal and political framework in which the public discourse takes place. This fact comes with issues of generalizability and comparability of findings and touches concerns of specific biases in international hate speech research. In particular, the issue of a Western bias of contemporary social research also manifests in the field of hate speech research (Matamoros-Fernández & Farkas, 2021). As in many other research areas, much more resources go into research on hate speech in the US, Europe or East Asian countries than in countries of the Global South. For this reason, we aimed to include perspectives from Non-Western researchers into this volume to have a better picture of global hate speech research.

However, context is not the only cause of blind spots in hate speech research. Insufficient definition, conceptualization and operationalization of the phenomenon in question also contribute to this issue. Hate speech legislation or automated text analysis software often simply work on the basis of a binary "hate / no hate" logic, which does not reflect the various shades on the continuum of problematic and disruptive speech. Thus, some authors in this section aim to advance our understanding of hate speech and its variants from different perspectives, providing theoretical conceptualizations or recommendations for more thorough methodological approaches.

As a start, *Afonso de Albuquerque* and *Marcelo Alves* analyze the specific conditions under which the Bolsonaro family in Brazil managed to build a social media-based ecosystem that combined strategies of disinformation, fake accounts and hate speech to support Jair Bolsonaros finally successful campaign for presidency. In their comprehensive account of the situation in Brazil, the authors highlight both national peculiarities and general tendencies of the evolution of hate speech in the context of political campaigns.

Zahera Harb adds the perspective of Lebanon, a country strongly impacted by severe confrontations of ethnically-defined political groups. Using the events around the explosions in the Beirut harbor, she widens the perspective to the role of journalists in the distribution of hate speech in society. In her study, she shows that in Lebanon many journalists do not have a differentiated understanding of hate speech and often spread hate messages amongst (legitimate) criticism of politicians. The difficult political situation of the country, which is mirrored in public discourse, requires a very thorough definition and understanding of hate speech and its consequences.

Using a feminist campaign in Poland as an example, *Dagmara Szczepańska* and *Marta Marchlewska* are exploring the boundaries between hate speech and offensive and vulgar language as means to attract attention and start a discourse in society. From their national background, they contribute to the debate about a context-sensitive definition of hate speech. It is not an expression or a term per se that constitutes hate speech, so their argument, but whether it is used—as in the example of the All-Poland Women's Strike—to point towards abuse and raise awareness for a societal problem or to attack an isolated group aiming at degrading their dignity and incite violence against them.

Anna Litvinenko connects to the preceding texts problematizing contextual factors by providing a theoretical categorization of different levels of context. Opening up a spectrum between situational and sociocultural contexts, she refers to the problem of a too simple black-and-white understanding of hate speech, which is not only part of many scientific approaches but also of current legislation. Such shortcomings can seriously harm anti-hate speech measures, for example by negatively affecting free speech, which is why she argues in favor of more context-sensitive approaches both in science and regulation.

Issues with regulatory interventions against hate speech are also in the focus of *Tomiwa Ilori*. In his example and from a legal perspective, the practical conflict between the prevention of hate speech and the violation of freedom of expression becomes apparent. Referring to the Nigerian context, but also including the wider approach of the African Commission on Human and People's Rights, he discusses alternative approaches to countering hate speech while preserving citizens' right to free speech.

A crucial field of fighting hate speech, both promising and potentially harmful, is the subject of *Sana Ahmad*, who takes a closer look at the internal content moderation policies of social media platforms. While many of us still hope that platforms sorting out negative content may be a solution for hate speech, disinformation and other sorts of content, her study on content moderation workers and sub-contractors in India puts the spotlight on moderation processes and working conditions as relevant contextual factors in the ecosystem of anti-hate speech actors and strategies. Connecting to the organizational layer of context outlined by Anna Litvinenko before, working conditions and power relations appear as important factors for the effectiveness of anti-hate speech measures.

The first section concludes with a text by *Christian Schemer* and *Liane Reiners*. Written as a response to the articles above, their contribution focuses on questions of comparability of hate speech studies from a basic, methodological perspective. The two authors discuss various aspects of concepts like the core term "hate speech", sampling and operationalization. As contexts of research are always quite different by nature, they argue that functional equivalence should be the goal in comparative hate speech research. However, they do not focus on comparative research alone but on hate speech research in general, which needs to produce findings that can be interpreted across studies to produce progress in our understanding of the phenomenon.

3 Theoretical perspectives: Terms, concepts and definitions

Taking up the question of which concepts we should work with in our research, the second part of this volume is devoted to the multiplicity of terms and definitions in the field of hate speech research and its neighboring strands. There are two main motivations behind this focus: First, we have a growing set of concepts competing in the broader field, but only little discussion of the implications and issues related to this inflation of terms and definitions (Sellars, 2016, p. 4). Accordingly, we see the need for a broader conversation about the theoretical and empirical contributions of each concept. How can we balance the demand for comparability of research with the need for specification and focus?

Second, we see not only a growing number of concepts but also a sort of camp formation in terms of who works with which of these concepts. For example, in a recent review paper on racism and hate speech in social media, Matamoros-Fernández and Farkas (2021) note "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'" (p. 216). Based on this finding, they detect a "terminological divide in the field" (p. 212). And indeed, our observation as editors of this volume is a similar one: Conceptual issues were discussed quite passionately between the authors in the course of the mutual reviews. There is clearly a need for more in-depth discussion here.

Our collection of texts on different concepts can hopefully be a start for this discussion, especially since it does not cover all of them. That said, our hope is that it initiates a more intense and informed conversation and helps building bridges in the process. We think academia has a special responsibility to address conceptual and definitional issues, given the fact that hate speech is also the subject of intense public debate.

We start with the "hate speech" concept as it is prominently included in this book's title, and also because we work with this concept in our own research as well. Nevertheless, we asked *Liriam Sponholz* to write a plea for this concept, since she is a renowned expert in this regard. In the first text of this section, she elaborates on the origins of the hate speech term in critical race theory, which has already embedded the consideration of social inequalities and power asymmetries in the definition of the term. According to this understanding, hate speech is defined as a symbolic attack against historically or systematically marginalized groups and their (supposed) members. Against this background, she then discusses the issues of concept stretching, concept shrinking and conceptual inflation in the recent literature and their consequences for academia, politics, and society.

Lena Frischlich discusses the specific fallouts of hate speech from a social psychological perspective, similarly concluding that hate speech cannot be understood without taking into account pre-existing power structures and resource inequalities. In the second part of the text, she discusses the psychological research on perpetrators of hate speech and derives valuable insights for preventive measures.

With the concept of "dangerous speech," *Susan Benesch* contributes a perspective that also focuses on the harm of speech acts but draws on empirically observed patterns in public speech in the run-up to genocides and mass violence in different parts of the world and historical periods. Specific to the concept is the observation that speech acts have a cumulative effect on people through repetition and that different contextual factors play a role in assessing the (gradual) dangerousness of a speech act.

Marike Bormann and *Marc Ziegele* argue for the concept of (political) "incivility," which is rooted in social theory (e.g., deliberation theory and politeness theories) and has a long research tradition. The two authors discuss current challenges of the research strand related to the inconsistency of definitions and measures, the reliability of incivility measurement, and normative implications. Moreover, they offer a multidimensional model of political incivility that integrates different strands of incivility research and encompasses violations of five different norms of communication.

With the concept of "toxicity," *Julian Risch* presents a quite different perspective. The concept originated in computer science and application-oriented, industry-led research in the area of automated user comment classification (and hiding/removal). It focuses on the impact of user comments in online discussions and on the goal of ensuring that no users are pushed out of these discussions. Similar to incivility, toxicity is a comparatively broad concept that can encompass various subcategories and can be adapted to the specific needs of the potential users of a classification solution.

In her text on "extreme speech," *Sahana Udupa* introduces a critical perspective on digital practices that departs from established definitions of hate speech and mis-, dis- or malinformation but calls for a holistic, culturally and historically sensitive approach to these practices. Rather than replacing existing concepts, extreme speech research aims to add new perspectives to hate speech research and considers ambivalences in the context of (political and economic) power relations, colonialism, and socio-technological transformations. Therefore, the framework emphasizes the need to balance the close contextualization of immediate contexts with a deep contextualization of underlying historical and colonial continuities.

The text by *Thorsten Quandt* and *Johanna Klapproth* revises the umbrella concept of "dark participation," introduced by the first author in 2018. This concept offers a systematization of various forms of negative or destructive user participation on the Internet along the main categories of actor, reasoning, object/target, audience, and process. However, the original article was also motivated as a commentary on the prevailing, one-sided focus of research (and, thus, also of this volume) on such negative aspects and as a call for more integrative theorizing and research. In their text for this volume, the two authors now reemphasize this motivation, discuss the resulting conceptual limitations of the dark participation model, and summarize the reactions and recommendations of the research community that followed the original publication.

Gina M. Masullo concludes the second part of this book with a text calling for a new approach to incivility research, which can also be read with regard to other concepts. In this text, Masullo pleads for addressing the specific forms of incivility, rather than continuing to treat it "as a monolith." In particular, she points to the need for multidimensional approaches that take into account the different theoretical underpinnings of incivility and allow for more specific research questions to be asked, for example, regarding the harmfulness of certain forms of incivility or contextual factors. She further identifies three research areas that need more research: the impact of online incivility on marginalized social groups and the protection of these groups, the role and power of social media platforms in regulating online incivility, and the dynamics between incivility and other forms of problematic online communication such as mis- and disinformation.

4 Methodological perspectives: Operationalization, automation and data

The third section of this volume focuses on methodological issues in the context of hate speech research. As in any other field, valid and reliable methods are key to scientific evidence on hate speech, especially because this field of research brings together different disciplinary perspectives and methodological standpoints. As an object of academic research, hate speech in social media is not conventional media content but rather a form of applied language sitting in the ambivalent space between interpersonal and public communication, shaped by social interactions, algorithmic decision-making, business models and design decisions of platform companies. Given the fast evolving possibilities for the collection and analysis of (big) data, empirical hate speech research not only demands for new theoretical models of public spheres and social discourse but also has to solve challenges of accessing, archiving, sharing and analyzing data.

The section opens with a text by *Babak Bahador*, who presents an approach to monitoring hate speech that he and his team have used to analyze U.S. media. Starting from a critique of common hate speech definitions, he introduces an hate speech intensity scale that ranges from "disagreement" to "death." He justifies the necessity of such an early warning system, which also includes weaker forms of antagonistic criticism, by pointing out that "[o]nce more extreme hate speech takes hold, it could also be a sign that it is too late to implement more peaceful preventative actions."

Salla-Maaria Laaksonen provides valuable insights into lessons learnt in a use case for automated hate speech detection. She describes which compromises and simplifications are necessary to develop and apply a successful machine learning model for the identification of hate speech and emphasizes the importance of human training and monitoring. In her use case, contextual factors regarding the message, the author and the public impact of the postings increased the model quality and its lifetime.

Christian Baden discusses the numerous challenges of language for machine-assisted hate speech detection. For example, changes in language can be used metaphorically and ironically and thus mask insults and hate. In addition, the expansion of classification models through contextual data could lead to more ambiguity and evasive language use by those who use hate speech. It is a kind of arms race. The methods are refined but still cannot overcome the evolving social abysses behind animosity and hate.

Besides ambiguity and irony, implicity is another major challenge for identifying hate speech. Falling back on a corpus from their research project "Decoding Antisemitism," *Matthias J. Becker* and *Hagen Troschke* present examples of implicit statements that contain antisemitic stereotypes and prejudices but that are not clear at first glance. They distinguish three areas of knowledge that help to extrapolate the implicit, and eventually identify those forms of antisemitism that are often disguised. In order to secure one's own interpretations in this context, the authors give concrete examples of "how implicitness can be realized at the different levels and how these levels can interact."

"Machines do not decide hate speech" is the title and claim of the text by *Jae Yeon Kim*. The author understands the establishment of what counts as hate speech as a negotiation process between social groups based on norms. Transparency and debate on the applied definitions of hate speech must therefore precede the model-building process. Accordingly, he argues that persons and groups affected by hate speech need to be included into the process, which would make it both more accurate and democratic.

Anke Stoll critically comments machine learning as part of the artificial intelligence hype. In a kind of recipe, she shows how, in four simple steps, a phony classifier can be trained to deliver seemingly outstanding results that are nothing but artifacts. In this context, she discusses potential pitfalls and flaws of machine learning models and shows how not to proceed if we aim for meaningful results.

In the next text, *Laura Laugwitz* demonstrates how validity as a major quality criterion for empirical studies can be applied to automated content analyses. She explains various supervised text classification methods and shows that the functional descriptions of these models are not suitable for an assessment of validity in the empirical sense. Following an interdisciplinary approach, she pleads for closer cooperation between computer science and communication science to develop such criteria.

From a legal perspective, *Paddy Leerssen, Amélie Heldt* and *Matthias C. Kettemann* look at the accessibility of social media data for researchers in Europe. There are many laws that make access difficult and some regulations that should make it easier to get data from platforms. Privacy, freedom of information, data protection and copyright are rights and areas of law that partly overlap and can make scientific access to platform data difficult. Finally, the authors call for a clear and unambiguous framework for scientific data access.

Jakob Jünger takes a look at social media data from a hermeneutic perspective. Data collection here is an uncertain process that requires many interpretative decisions and therefore has a great influence on the later research results. The selection and availability of data, access restrictions, the systematics of the websites as well as the archiving of the data show the tension between creativity and standardization that we as researchers face and that we have to dissolve thoughtfully.

Paula Fortuna, Juan Soler-Company and *Leo Wanner* discuss challenges for both building and comparing annotation datasets. Studies in the context of abusive language research have shown the importance of such data for machine learning models, a lack of common understandings in this context, and the presence of bias and artifacts in recognition and evaluation. Against this background, the authors provide guidelines to address the most pressing issues in a step-by-step guideline to improve the quality of annotated datasets.

In their response to the texts in this third section, *Jaime Lee Kirtz* and *Zeerak Talat* reflect on the various methodological challenges that each step of hate speech research faces, providing a broader orientation for each text of this section they discuss. In this context, they attach particular importance to social issues that need to be addressed in future research on hate speech detection.

Taken together, the third part of this book critically reflects the diversity and heterogeneity of methodological perspectives on machine-based models for the detection of linguistic constructs in social media. Against the background of these contributions, we think that the field of hate speech research is unlikely to succeed without true interdisciplinary exchange, discussions and collaboration. With this volume, we hope to contribute to such a project, and to stimulate first steps toward building bridges between disciplines, theoretical perspectives, and methods.

Last but not least, we would like to thank all authors of this volume for their excellent contributions, the rich discussions during the review process, and for their infinite patience with us editors. From our point of view, the experiment of a discursive collection of texts on the various challenges and future perspectives of hate speech research was more than successful. Perhaps it can even serve as a model for other research fields that are considering similar endeavors. To you, the reader, we wish an exciting and insightful read.

Sünje Paasch-Colberg is a Research Associate at the German Centre for Integration and Migration Research (DeZIM) in Berlin, Germany. https://orcid.org/0000-0002-0771-9646

Christian Strippel is research unit lead of the Weizenbaum Panel and the Methods Lab at the Weizenbaum Institute for the Networked Society in Berlin, Germany. https://orcid.org/0000-0002-7465-4918

Martin Emmer is Professor for Media and Communication Studies at Freie Universität Berlin and Principal Investigator at the Weizenbaum Institute for the Networked Society in Berlin, Germany. https://orcid.org/0000-0002-0722-132X

Joachim Trebbe is Professor for Media and Communication Studies at Freie Universität Berlin, Germany.

References

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.org/10.1177/152747642098223

Sellars, A. F. (2016). Defining Hate Speech. Research Publication No. 2016–20. Berkman Klein Center. https://doi.org/10.2139/ssrn.2882244