Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

EXPLORING SURVEY INSTRUMENTS IN ONLINE HATE SPEECH RESEARCH: A COMPREHENSIVE SCOPING REVIEW

Živa Šubelj
Faculty of Social Sciences, University of Ljubljana

Vasja Vehovar
Faculty of Social Sciences, University of Ljubljana

Avis Wisteria
Faculty of Social Sciences, University of Ljubljana

Andraž Petrovčič
Faculty of Social Sciences, University of Ljubljana

Jošt Bartol
Faculty of Social Sciences, University of Ljubljana

Introduction

Over the last decade, online hate speech (OHS) has emerged as a subject of active debate among a wide range of scholars as well as governmental and private entities. The notion of hate speech, otherwise rooted in legal frameworks, in the context of the European legislation system typically denotes extreme negative communication towards minorities and other protected groups, thereby being a specific exception to the freedom of speech (European Court for Human Rights, 2023). With the ever-evolving technological and social landscape, OHS is increasingly studied across various scientific fields, including computer science, criminology, communication, psychology, and educational research (see Waqas et al., 2019). With the advancements in machine learning technology, many scholars focused on researching new methods to identify OHS (Jahan & Oussalah, 2023; Poletto et al., 2021; Tontodimamma et al., 2021) and limit its spread (Masud et al., 2022; Qian et al., 2019). Additional impetus to the OHS debate was given in 2016, when global social media corporations and the European Commission signed a Code of Conduct to counter illegal OHS, binding IT companies to

Suggested Citation (APA): Šubelj, Ž., Vehovar, V., Wisteria, A., Petrovčič, A., & Bartol, J. (2024, October). *Exploring survey instruments in online hate speech research: A comprehensive scoping review*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

speedily review and remove potential hateful content (European Commission, 2016). Furthermore, the passing of the Digital Services Act in EU countries made hate speech issues formally nested in a broader legal and institutional frame (European Commission, 2020).

Nevertheless, scholars have only recently begun to systematically explore how individuals perceive and experience OHS as well as what kind of strategies they use for countering it (Bliuc et al., 2018). While such insights can be obtained with various methods – including qualitative methods and data mining techniques – this study focuses on survey measures, widely used research means for understanding large-scale social phenomena. High-quality survey measurement can bring insights into individuals' OHS perceptions and experiences, which is of great importance for the design and implementation of strategies aimed at informing the public about OHS and restraining its expansion. Importantly, educational and informational initiatives have proved effective in developing anti-hate critical thinking and fighting OHS (Müller & Lopez-Sanchez, 2021; Woo & Cho, 2023).

A scoping review of existing survey measures assessing individuals' perceptions and experiences with OHS will provide several original insights with practical implications. First, no such review has yet been conducted. Multiple studies reviewing prior scholarly literature looked at existing definitions of hate speech, with a focus on different online environments (Ermida, 2023; Hietanen & Eddebo, 2023; Kansok-Dusche et al., 2023; Papcunová et al., 2023; Sellars, 2016). However, most reviews were narrative or otherwise limited in scope. In addition, scholars conducting reviews of existing academic research on OHS emphasized that certain groups, such as OHS perpetrators, have not been given enough attention (Tontodimamma et al., 2021) and advocated for a greater focus on quantitative research methods (Bliuc et al., 2018; Castaño-Pulgarín et al., 2021; Kearns et al., 2023). Second, the current assessment approaches and collected data are fragmented (Kansok-Dusche et al., 2023). Providing an overview of existing survey measures can lead to the informed and expedient development of comprehensive and reliable instruments, that could be used in the future to gather valid and comparable data. Additionally, mapping the researched topics within the OHS field as well as population groups will help identify gaps in the literature and provide directions for future scholarly endeavours.

Therefore, we systematically collected and assessed existing academic papers to answer the following research questions (RQs):

RQ1: Which population groups are studied using OHS survey measures?
RQ2: Which OHS topics are covered by survey measures?
RQ3: Which kind of survey measures are used to assess OHS?

Methods and Results

A systematic approach has been undertaken to map existing academic evidence, following the PRISMA-ScR guidelines (Tricco et al., 2018). The search was executed

using the Web of Science and Scopus databases, alongside the bibliographic harvester of the University of Ljubljana (DiKUL).

A total of 725 articles were initially identified for review. After removing duplicates, 370 articles were screened by two independent reviewers based on their titles and abstracts. A total of 65 records (comprising 67 different studies), which included 309 survey measures related to OHS, were selected for full-text review. During this phase, data were extracted, including demographic information, survey admission type, and key characteristics of each survey measure. In addition, an inductive thematic analysis was conducted to categorize each survey measure based on the OHS aspect it addresses.

Concerning RQ1, preliminary findings indicate that the highest percentage of OHS survey-based studies focused on student and young adult populations (34%). Nearly half of the studies employed non-probability sampling methods. 70% disseminated the questionnaire online. Over 60% of the studies investigated OHS in general Internet contexts, followed by 36% focused on social media. Regarding minorities, most survey measures addressed OHS in general, followed by 26% focused on discrimination related to nationality, ethnicity, or migrant status.

In relation to RQ2, the 309 identified survey measures addressed different aspects of OHS. Through inductive thematic analysis, each question was categorized into one of 11 topics. The largest proportion of questions (21%) addressed individuals' reactions and coping strategies when faced with OHS. Over 18% of the questions examined perceptions of hate speech, while 16% focused on exposure to OHS and 13% on victimization. Topics of self-perpetration, others' perpetration, and combating hate speech each accounted for 5–10% of the questions. Fewer than 5% of the questions explored each of the following topics: consequences of OHS, free speech vs hate speech, legislation, and definitions of hate speech.

With respect to RQ3, 83% of survey measures were close-ended questions, and 4% were close-ended questions with an open-ended option “Other”. Half of the identified measures were ordinal scales (Likert-type questions (n=73) or Likert scales (n=71)), followed by 31% of questions with categorical response options (single- (n=73) or multiple-choice questions (n=22)) and three open-ended questions. However, no scale was consistently employed in a standardized way, resulting in fragmentation in question typology. Nevertheless, for each of the 11 topics, the most common and typical survey questions will be identified in the next phase.

Conclusion

Preliminary findings reveal a disproportionate focus on students and young adults, leaving older population groups under-researched, even though older adults express high levels of concerns regarding OHS (Pacheco, 2024) and online ageism increased during the pandemic (Levy et al., 2022). Additionally, the predominant use of non-

representative sampling methods in the reviewed studies raises concerns about the generalizability of the findings. Broad research scopes are predominant; in order to better understand real-world experiences, scholars should aim for more detailed research scopes, including studying OHS in specific online contexts and within specific groups. In terms of topics, existing survey measures mostly cover the topics of direct exposure/perpetration of OHS, neglecting the whole hate speech ecosystem; what comes before and after the dissemination of hate speech online. In terms of question types, ordinal scales are most common, but the lack of standardized scales results in fragmented findings that are difficult to compare. Future research should prioritize examining general and older populations and focus on their understanding of hate speech and how it can be mitigated. Standardized survey measures for OHS should be developed and systematically evaluated to improve the reliability and comparability of future research findings.

References

Bliuc, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*, 75–86. https://doi.org/10.1016/j.chb.2018.05.026

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*. https://doi.org/10.1016/j.avb.2021.101608

Ermida, I. (2023). Distinguishing online hate speech from aggressive speech: A Five-factor annotation model. In I. Ermida (Ed.), *Hate Speech in Social Media: Linguistic Approaches* (pp. 35–75). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-38248-2_2

European Commission. (2016). *Code of Conduct on countering illegal hate speech online*. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

European Commission. (2020). *Executive summary of the impact assessment report. Accompanying the document: Proposal for a regulation of the European Parliament and of The Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*. https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act

European Court for Human Rights. (2023). *Hate speech*. https://www.echr.coe.int/documents/d/echr/fs_hate_speech_eng

Hietanen, M., & Eddebo, J. (2023). Towards a definition of hate speech—With a focus on online contexts. *Journal of Communication Inquiry*, *47*(4), 440–458. https://doi.org/10.1177/01968599221124309

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*. https://doi.org/10.1016/j.neucom.2023.126232

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *TRAUMA, VIOLENCE, & ABUSE*, *24*(4), 2598–2615. https://doi.org/10.1177/15248380221108070

Kearns, C., Sinclair, G., Black, J., Doidge, M., Fletcher, T., Kilvington, D., Liston, K., Lynn, T., & Rosati, P. (2023). A scoping review of research on online hate and sport. *Communication and Sport*, *11*(2), 402–430. https://doi.org/10.1177/21674795221132728

Levy, S. R., Lytle, A., & Macdonald, J. (2022). The worldwide ageism crisis. *Journal of Social Issues*, *78*(4), 743–768. https://doi.org/10.1111/josi.12568

Masud, S., Bedi, M., Khan, M. A., Akhtar, M. S., & Chakraborty, T. (2022). Proactively reducing the hate intensity of online posts via hate speech normalization. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3524–3534. https://doi.org/10.1145/3534678.3539161

Müller, A., & Lopez-Sanchez, M. (2021). Countering negative effects of hate speech in a multi-agent society. *Frontiers in Artificial Intelligence and Applications*, *339*, 103–112. https://doi.org/10.3233/FAIA210122

Pacheco, E. (2024). *Older adults' safety and security online: A post-pandemic exploration of attitudes and behaviors*. ArXiv. https://doi.org/10.48550/arxiv.2403.09208

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., Moro, R., Pikuliak, M., Šimko, M., & Adamkovič, M. (2023). Hate speech operationalization: A preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems*, *9*, 2827–2842. https://doi.org/10.1007/s40747-021-00561-0

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation, 55*(2), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, *China*, 4755–4764. https://doi.org/10.18653/v1/d19-1482

Sellars, A. (2016). *Defining hate speech*. https://ssrn.com/abstract=2882244

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*(1), 157–179. https://doi.org/10.1007/s11192-020-03737-6

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–486. https://doi.org/10.7326/M18-0850

Waqas, A., Salminen, J., Jung, S., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS ONE*, *14*(9). https://doi.org/10.1371/journal.pone.0222194

Woo, H., & Cho, Y. Y. (2023). Fighting hate and hate speech: Raising anti-hate awareness through critical analysis of popular cultural texts on an undergraduate course. *Societies*, *13*(11). https://doi.org/10.3390/soc13110240