Hate Speech (Hate Speech/Incivility)

AUTHOR
Katharina Esau

KEYWORDS
hate speech, incivility, stereotyping, racism, se

BRIEF DESCRIPTION
The variable hate speech is an indicator used to describe communication that expresses and/or promotes hatred towards others (Erjavec & Kovačič, 2012; Rosenfeld, 2012; Ziegele, Koehler, & Weber, 2018). A second element is that hate speech is directed against others on the basis of their ethnic or national origin, religion, gender, disability, sexual orientation or political conviction (Erjavec & Kovačič, 2012; Rosenfeld, 2012; Waseem & Hovy, 2016) and typically uses terms to denigrate, degrade and threaten others (Döring & Mohseni, 2020; Gagliardone, Gal, Alves, & Martínez, 2015). Hate speech and incivility are often used synonymously as hateful speech is considered part of incivility (Ziegele et al., 2018).

FIELD OF APPLICATION
Hate speech (see also incivility) has become an issue of growing concern both in public and academic discourses on user-generated online communication.

REFERENCES/COMBINATION WITH OTHER METHODS OF DATA COLLECTION:
Hate speech is examined through content analysis and can be combined with comparative or experimental designs (Muddiman, 2017; Oz, Zheng, & Chen, 2017; Rowe, 2015). In addition, content analyses can be accompanied by interviews or surveys, for example to validate the results of the content analysis (Erjavec & Kovačič, 2012).

EXAMPLE STUDIES
Research question/research interest: Previous studies have been interested in the extent of hate speech in online communication (e.g. in one specific online discussion, in discussions on a specific topic or discussions on a specific platform or different platforms in comparatively) (Döring & Mohseni, 2020; Poole, Giraud, & Quincey, 2020; Waseem & Hovy, 2016).

Object of analysis: Previous studies have investigated hate speech in user comments for example on news websites, social media platforms (e.g. Twitter) and social live streaming services (e.g. YouTube, YouNow).

Level of analysis: Most manual content analysis studies measure hate speech on the level of a message, for example on the level of user comments. On a higher level of analysis, the level of hate speech for a whole discussion thread or online platform could be measured or estimated. On a lower level of analysis hate speech can be measured on the level of utterances, sentences or words which are the preferred levels of analysis in automated content analyses.

Table 1. *Previous manual and automated content analysis studies and measures of hate speech.*

Example study (type of content analysis)	Construct	Dimensions/variables	Explanation/ example	Reliability
Waseem & Hovy (2016) (automated content analysis)	hate speech	sexist or racial slur	–	–
		attack of a minority	–	–
		silencing of a minority	–	–
		criticizing of a minority without argument or straw man argument	–	–
		promotion of hate speech or violent crime	–	–
		misrepresentation of truth or see-	–	–
		problematic hash tags. e.g.	–	–
		negative stereotypes of a minority	–	–
		defending xenophobia or sexism	–	–
		user name that is offensive, as per the previous criteria	–	–
		hate speech	–	κ = .84

Example study (type of content analysis)	Construct	Dimensions/variables	Explanation/ example	Reliability
Döring & Mohseni (2020) (manual content analysis)	hate speech	explicitly or aggressively sexual hate	e. g. "are you single, and can I lick you?"	κ = .74; PA = .99
		racist or sexist hate	e.g. "this is why ignorant whores like you belong in the fucking kitchen", "oh my god that accent sounds like crappy American"	κ = .66; PA = .99
		hate speech	–	κ = .70

Note: Previous studies used different inter-coder reliability statistics; κ = Cohen's Kappa; PA = percentage agreement.

More coded variables with definitions used in the study Döring & Mohseni (2020) are available under: https://osf.io/da8tw/

REFERENCES

Döring, N., & Mohseni, M. R. (2020). Gendered hate speech in YouTube and YouNow comments: Results of two content analyses. SCM Studies in Communication and Media, 9(1), 62–88. https://doi.org/10.5771/2192-4007-2020-1-62

Erjavec, K., & Kovačič, M. P. (2012). "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. Mass Communication and Society, 15(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Gagliardone, I., Gal, D., Alves, T., & Martínez, G. (2015). Countering online hate speech. UNESCO Series on Internet Freedom. Retrieved from http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

Muddiman, A. (2017). : Personal and public levels of political incivility. International Journal of Communication, 11, 3182–3202.

Oz, M., Zheng, P., & Chen, G. M. (2017). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. New Media & Society, 20(9), 3400–3419. https://doi.org/10.1177/1461444817749516

Poole, E., Giraud, E. H., & Quincey, E. de (2020). Tactical interventions in online hate speech: The case of #stopIslam. New Media & Society, 146144482090331. https://doi.org/10.1177/1461444820903319

Rosenfeld, M. (2012). Hate Speech in Constitutional Jurisprudence. In M. Herz & P. Molnar (Eds.), The Content and Context of Hate Speech (pp. 242–289). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.018

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. Information, Communication & Society, 18(2), 121–138. https://doi.org/10.1080/1369118X.2014.940365

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In J. Andreas, E. Choi, & A. Lazaridou (Chairs), Proceedings of the NAACL Student Research Workshop.

Ziegele, M., Koehler, C., & Weber, M. (2018). Socially Destructive? Effects of Negative and Hateful User Comments on Readers' Donation Behavior toward Refugees and Homeless Persons. Journal of Broadcasting & Electronic Media, 62(4), 636–653. https://doi.org/10.1080/08838151.2018.1532430