ICT20-015 - Transparent Automated Content Moderation (TACo)

Abstract

Online political discussions are increasingly perceived as negative, aggressive, and toxic. This is a worry, because exposure to toxic talk undermines trust and fosters cynicism, leading to a polarized society. Defining what could be considered “toxic” is therefore one of the most pressing challenges for researchers today, because such a definition may be used to develop (semi-) automated content moderation systems that ensure healthy political conversations on a global scale. However, the available research on toxic talk and content moderation is elite-driven and imposes top-down definitions of what is “good” or “bad” on users. This has resulted in biased content moderation models, and it has damaged the reputation of those who have implemented them. More importantly, however, a top-down approach removes agency from citizens in a time when many already feel they have too little influence on their daily information intake. Therefore, TACo proposes a novel user-centric approach towards automated content moderation. We conduct qualitative and exploratory social science research to learn what citizens themselves want, when it comes to toxic talk and content moderation. Then, we develop moderation scenarios based on this knowledge, testing for usefulness and reliability of models. Finally, we test whether what people “want” truly has beneficial effects for them: we conduct experiments that test the effects of these models on citizens’ political trust, engagement, and well-being.

Scientific disciplines:
Political communication (50%) | Data science (50%)

Keywords:
content moderators, user comments, incivility, toxic talk, citizen engagement

Principal Investigator:	Sophie Lecheler
Institution:	University of Vienna
Co-Principal Investigator(s):	Allan Hanbury (TU Wien)

v.l.n.r. Sophie Lecheler, Allan Hanbury ©Foto Schoerg

Status: Ongoing (01.09.2021 - 30.04.2026)
GrantID: 10.47379/ICT20015

Further links to the persons involved and to the project can be found under
https://www.gmbh.wwtf.at/funding/programmes/ict/ICT20-015/