ICT20-016 - Young People Against Online Hate: Computer-assisted Strategies for Facilitating Citizen-generated Counter Speech

Abstract

Background

Dealing with online hate, cyberbullying and digital violence poses a significant challenge to society. Adolescents are particularly affected due to their intensive internet usage. They often respond to such attacks with avoidance or withdrawal, despite their considerable potential for intervention, particularly by means of promoting counter speech. This form of digital civic courage often fails due to a lack of skills, particularly in light of platform -specific dynamics. In this study, sociologists and computer scientists collaborated with young people to develop computer -based strategies to promote digital civic courage .

Methods

The following methodological approaches were employed : (i) automated, computer -based data collection from various case studies to capture typical manifestations of online hate and counter speech on platforms such as TikTok ; (ii) quantitative and qualitative content analys i s as well as machine learning techniques to identify specific patterns of counter speech and use these for the computational detection of counter speech among adolescents ; (iii) participatory research methods to co -design requirements for a digital counter -speech tool with young people in a way that ensure s its acceptance and actual use .

Main Findings

The study highlights the challenges of addressing digital violence and platform -specific obstacles. The practicability of counter speech is significantly impeded by the fact that online hate takes subtle forms in both text and visual aspects , and is continuously transforming. Consequently, counter speech is difficult to identify. At the same time, adolescents are exposed to varying degrees of online hate , depending on platform algorithms, user behavior, and community interactions. Those who are willing to reject hate and demonstrate civic courage often encounter little exposure to hate in their own “bubble,” which is characterized by positive content and like -minded users. When adolescents deliberately search for negative content to respond with counter speech, algorithmic logic often results in increased and unwanted exposure to hateful content. The absence of counter speech among adolescents does not primarily stem from a lack of empathy; rather, empathetic individuals often have few opportunities to intervene or, when they do, face even greater exposure to hate. At the same time, the lack of digital civic courage among adolescents can be mainly explained by the perceived limitations of digital spaces: young people fail to achieve the visibility, reach, and impact required for effective counter speech, leading to a sense of powerlessness. In our study, we developed a supportive and playful counter -speech recommendation tool that allows to mark hate, prompt s responses, and suggest s counter speech that is perceived as attractive and motivating .

Conclusions

The findings show that platform -specific algorithms that prioritize polarizing content and limit the visibility of counter speech play a decisive role in reducing the effectiveness of counter speech and hindering constructive engagement. It is essential that platform providers ensure the implementation of digital tools that actively support online civic courage.

Wiener Wissenschafts-,
Forschungs- und Technologiefonds
Vienna Science and Technology Fund

Further information on the results in this article on the website of FH St. Pölten.

Scientific disciplines:
Sociology of youth (50%) | Machine learning (50%)

Keywords:
Hate speech, counter speech, online civil courage, information retrieval, multimodal analysis

Principal Investigator:	Ulrike Zartler
Institution:	University of Vienna
Co-Principal Investigator(s):	Matthias Zeppelzauer (University of Applied Sciences St. Pölten) Christiane Atzmüller (University of Vienna)



v.l.n.r. Ulrike Zartler ©citronenrot, Matthias Zeppelzauer ©FHSanktPölten, Christiane Atzmüller ©Gilbert Handler

Status: Completed (15.09.2021 – 14.09.2025)
GrantID: 10.47379/ICT20016

Further links to the persons involved and to the project can be found under
https://www.wwtf.at/funding/programmes/ict/ICT20-016/