

Social Influencer:in: Mit wirksamer Gesprächsführung gegen Hass im Netz

Impressum

März 2026

Zürcher Hochschule für Angewandte Wissenschaften

Departement Soziale Arbeit
Institut für Delinquenz und
Kriminalprävention
Pfingstweidstrasse 96
Postfach
CH-8037 Zürich
www.zhaw.ch/sozialarbeit

Zürcher Hochschule für Angewandte Wissenschaften

School of Engineering
Center for Artificial Intelligence
Technikumstrasse 71
8400 Winterthur
www.zhaw.ch/de/engineering

Kontakt

ZHAW Soziale Arbeit
Institut für Delinquenz und
Kriminalprävention
Projektleitung: Judith Bühler
Kontakt: judith.buehler@zhaw.ch

Autor:innen

Judith Bühler
Kyra Jetzer
Johanna Seiwald
Pius von Däniken
Maria Kamenowski
Dr. Katja Girschik
Prof. Dr. Dirk Baier
Laurent Sedano
Lars Schmid
Dr. Don Tuggener
Patrick Giedemann

Kurzübersicht

Digitale Plattformen sind heute zentrale Orte gesellschaftlicher Diskussionen. Viele politische Themen, gesellschaftliche Konflikte oder aktuelle Ereignisse werden in Kommentarspalten von Onlinezeitungen und auf Social Media diskutiert.

Diese Diskussionsräume bieten zwar Möglichkeiten für Austausch und Meinungsvielfalt, sind jedoch häufig auch von toxischen Kommunikationsformen und Hate Speech geprägt. Abwertende Aussagen, pauschale Zuschreibungen oder diskriminierende Kommentare können Diskussionen eskalieren lassen und zu einer Verschlechterung der Diskussionskultur beitragen.

Solche Kommentare können dabei nicht nur einzelne Personen betreffen, sondern auch auf mitlesende Nutzer:innen wirken und gesellschaftliche sowie politische Debatten beeinflussen.



Was ist Hate Speech?

Der Begriff Hate Speech bezeichnet kommunikative Äußerungen, die Personen oder Gruppen aufgrund bestimmter Merkmale abwerten, angreifen oder entmenslichen. Sie kann unterschiedliche Formen annehmen – von offenen Beleidigungen über stereotype Zuschreibungen bis hin zu subtileren Formen der Abwertung.

Dazu gehören beispielsweise Angriffe aufgrund von

- Herkunft, Ethnizität oder Religion
- Geschlecht oder sexueller Orientierung
- politischer Zugehörigkeit oder sozialem Status
- körperlicher Merkmale

Toxic Speech

Neben Hate Speech kommt in Online-Diskussionen häufig auch toxische Kommunikation vor. Dabei handelt es sich um Kommentare, die zwar nicht direkt diskriminierend sind, aber die Diskussion durch aggressive, druckausübende oder respektlose Ausdrucksweisen vergiften.

Counter Speech

Eine mögliche Reaktion auf Hate Speech ist Counter Speech. Darunter versteht man kommunikative Gegenrede, die diskriminierenden oder toxischen Aussagen widerspricht und alternative Perspektiven sichtbar macht.

Schon gewusst?

10815

analysierte
Kommentare

9431

Counter Speech
Interventionen

886

untersuchte
Diskussionsverläufe

hate

Teilprojekt 1 | Methodengeleitete Counter Speech in digitalen Diskursräumen

Das erste Teilprojekt untersucht, wie digitale Soziale Arbeit zur Deeskalation toxischer Online-Diskurse beitragen kann. Ziel war es, Hate Speech nicht nur zu dokumentieren, sondern aktiv und professionell darauf zu reagieren.

Hierfür wurden Social Influencer:innen rekrutiert, geschult und systematisch begleitet. Sie intervenierten auf Social Media sowie in Kommentarspalten von Onlinezeitungen. Im Zentrum stand Counter Speech, der Diskursqualität stärken, Empathie fördern und Polarisierung entgegenwirken sollte.

Counter Speech Methoden im Überblick

Tabelle 1: Kommunikationsmethoden für Counter Speech

Empathie	Alternative Narrative	DiQu – Fokus auf Diskussionsqualität
Ziel: Perspektivübernahme ermöglichen	Ziel: Konstruktive Gegenperspektiven stärken	Ziel: Kooperative Konfliktbearbeitung
Ansatz: Person wird eingeladen, sich in Betroffene hineinzuversetzen.	Ansatz: Positive Perspektiven zu Vielfalt, Integration und Zusammenleben aufzeigen.	Ansatz: • Begrüssung • Problem beschreiben • Intention benennen • Verabschiedung
Wirklogik: Deeskalation durch emotionale Reflexion.	Wirklogik: Freund-Feind-Schema wird durch differenzierte Sichtweisen entgegengewirkt.	Wirklogik: Verbesserung der Gesprächskultur.

Vom Einzelimpuls zum Dialog

Im Projektverlauf wurden zwei Interventionsformen erprobt. Im ersten Projektjahr erfolgten einmalige Interventionen auf toxische Kommentare. Im zweiten Projektjahr wurde das Vorgehen erweitert und dialogische Mehrfachinterventionen getestet.

Ablauf Einmalinterventionen



Ergänzungen bei Mehrfachinterventionen

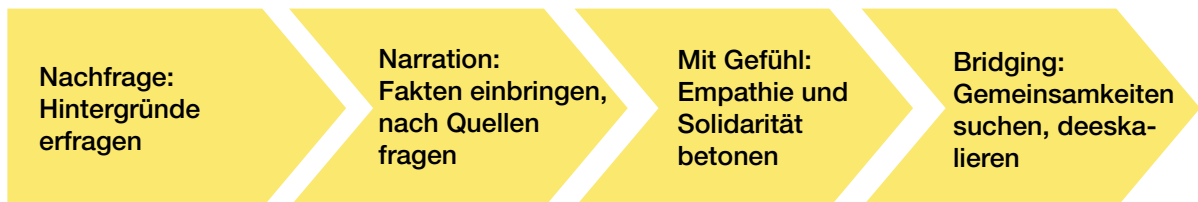


Abbildung 1: Ablauf Einmalinterventionen inkl. Ergänzungen bei Mehrfachinterventionen

Erkenntnisse

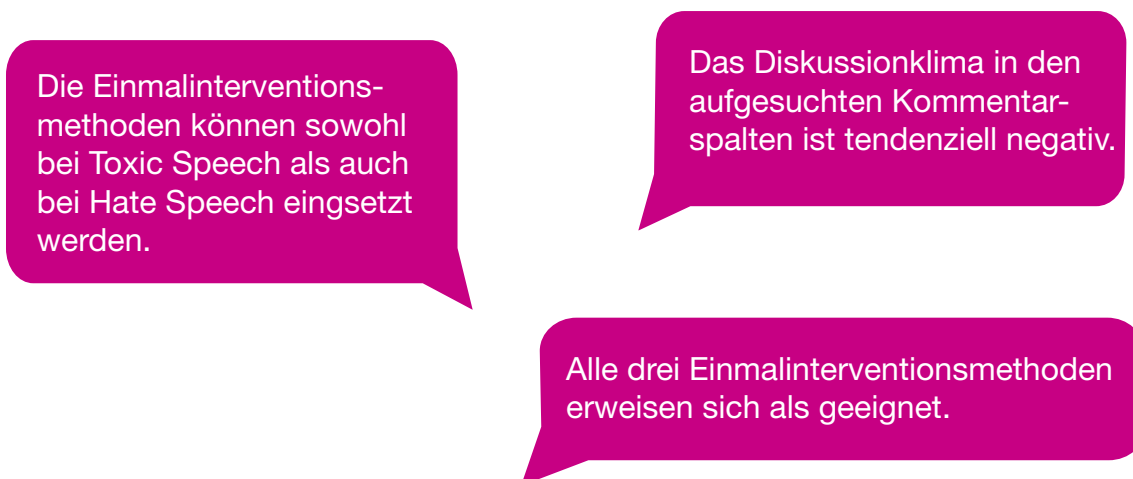


Abbildung 2: Zentrale Erkenntnisse aus Teilprojekt 1

Das Teilprojekt zeigt, dass digitale Soziale Arbeit zur Verbesserung öffentlicher Diskurse beitragen kann. Professionalisierter Counter Speech ist keine spontane Reaktion, sondern eine strukturierte und methodisch fundierte Intervention. Sie eröffnet neue Handlungsperspektiven für Soziale Arbeit im digitalen Raum.

Teilprojekt 2 | Technische Unterstützung durch die Browser Extension Hyssop

Das zweite Teilprojekt untersuchte, wie technische Werkzeuge Social Influencer:innen bei der Bearbeitung von Hate Speech unterstützen können.

Dafür wurde die Browser Extension Hyssop entwickelt. Sie half dabei, problematische Kommentare zu identifizieren, Interventionen zu dokumentieren und – im späteren Projektverlauf – Vorschläge für Counter Speech zu generieren.

Die Extension fungierte damit als technisches Assistenzsystem für digitale Interventionsarbeit.

Funktionen von Hyssop

Hate Speech Classifier

Eine KI-Analysekomponente bewertete Kommentare automatisch auf ihre Toxizität und hob potenziell problematische Beiträge hervor.

Annotation

Kommentare konnten direkt markiert und kategorisiert werden. Diese Annotationen dienten als Grundlage für die spätere Auswertung.

Formulierung von Counter Speech

Ab 2025 generierte die Extension mittels KI Vorschläge für Counter Speech auf Basis der verwendeten Methoden (Empathie, Alternative Narrative, DiQu). Diese konnten übernommen, angepasst oder neu formuliert werden.

Speicherung von Interaktionen

Zur Analyse dokumentierte Hyssop zentrale Elemente jeder Intervention:

- ursprünglicher Kommentar
- Annotation der Influencer:in
- gewählte Methode
- verfasster Counter Speech

Erkenntnisse

1 Technische Unterstützung erleichtert die Arbeit

Die Browser Extension half dabei, toxische Kommentare schneller zu identifizieren und Interventionen systematisch zu dokumentieren.

2 Automatische Klassifikation funktioniert – mit Einschränkungen

Das Klassifikationsmodell konnte Kommentare grundsätzlich zwischen Toxic, Hate und nicht relevant unterscheiden.

3 Schweizer Dialekt bleibt Herausforderung

Ein erheblicher Teil der Kommentare war im Schweizer Dialekt verfasst, was die automatische Analyse erschwerte.

4 KI bleibt Unterstützung

Die generierten Counter Speech-Vorschläge wurden meist stark überarbeitet – menschliche Kontextkompetenz bleibt entscheidend.

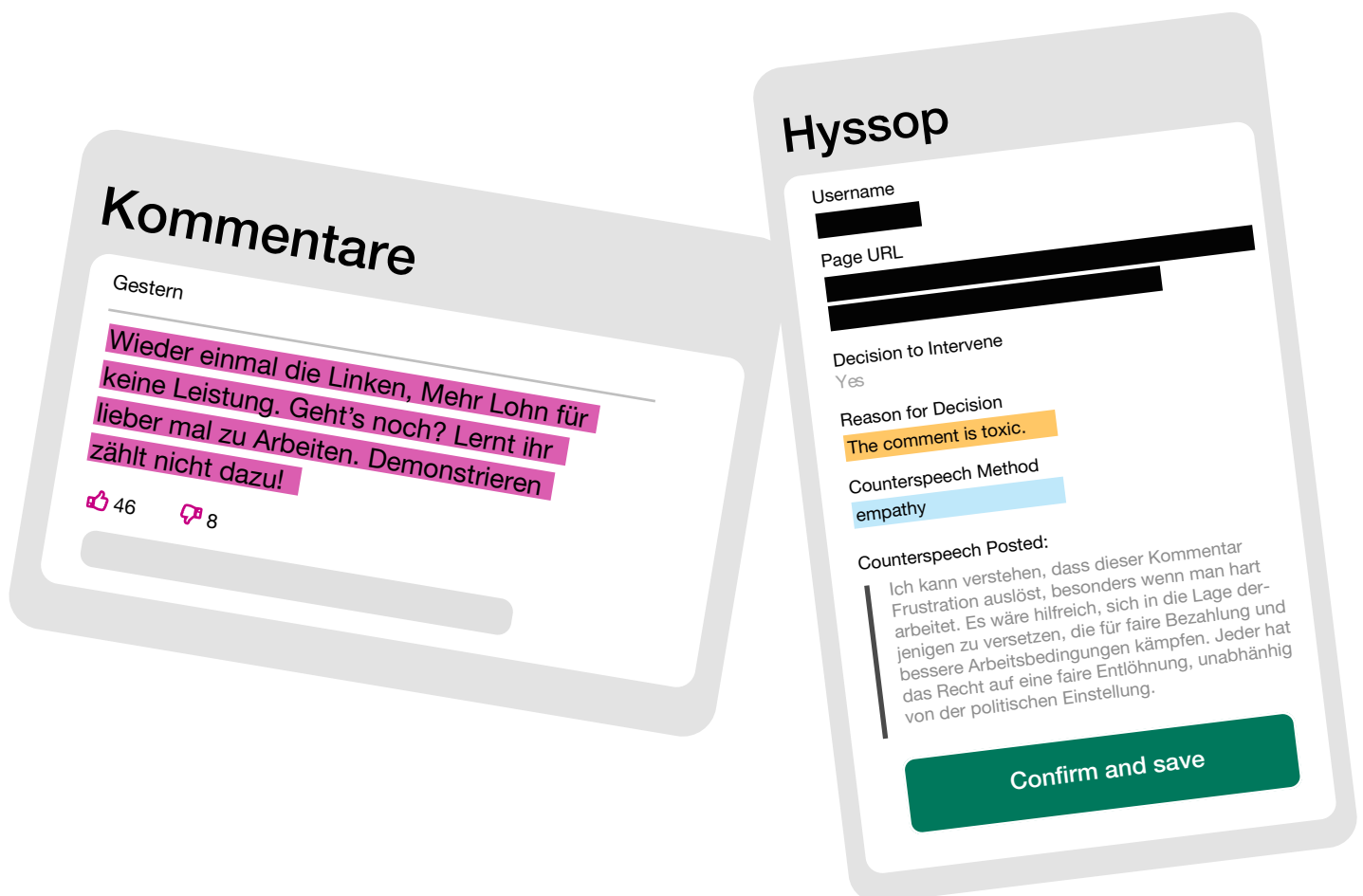


Abbildung 3: Hate Speech Classifier in Hyssop

Teilprojekt 3 | Wissenschaftliche Begleitforschung

Das dritte Teilprojekt untersuchte das Projekt aus wissenschaftlicher Perspektive. Im Zentrum standen die Wirkung der eingesetzten Kommunikationsmethoden, die Erfahrungen der Social Influencer:innen sowie die Analyse von Profilen, die toxische oder hasserfüllte Kommentare verbreiteten.

Die Begleitforschung verfolgte das Ziel, systematisch zu untersuchen, unter welchen Bedingungen Counter Speech wirkt, welche Unterstützung Social Influencer:innen benötigen und welche Erkenntnisse sich über hassverbreitende Profile gewinnen lassen.

3 Teilmodule im Überblick

Effekte der Kommunikationsmethoden

Untersucht wurde, wie sich die drei Methoden aus Teilprojekt 1, ergänzt durch Mehrfachinterventionen, auf den weiteren Diskussionsverlauf auswirkten. Dabei lag der Fokus nicht primär auf den hassäussernden Personen selbst, sondern auch auf den Mitlesenden und der Frage, ob sich die Diskussionsqualität durch Counter Speech verändern lässt.

Erfahrungen der Social Influencer:innen

Mittels qualitativer Interviews und Schulungsevaluationen wurde erhoben, wie die Social Influencer:innen ihre Tätigkeit erlebten. Im Fokus standen:

- Erfahrungen mit den Kommunikationsmethoden
- Umgang mit Hasskommentaren in Bezug auf Belastung und Psychohygiene
- technische Unterstützung
- Schulungen und Begleitangebote

Analyse hassverbreitender Profile

Im zweiten Projektjahr wurden öffentlich zugängliche Profilinformationen von Accounts ausgewertet, mit denen im Rahmen von Einmal- oder Mehrfachinterventionen interagiert wurde. Ziel war es, explorativ zu prüfen, welche Profilm Merkmale sichtbar sind und welche Erkenntnisse sich daraus ableiten lassen.

Erkenntnisse

Counter Speech wirkt punktuell

Interventionen können Wirkung entfalten.

Wirkung hängt vom Kontext ab

Methode, Plattform und Diskussionsthema beeinflussen die Effekte.

Begleitung ist zentral

Schulung, Feedback und Austausch stärkten die Social Influencer:innen bei ihrer Tätigkeit.

Profile bleiben nur teilweise sichtbar

Öffentlich zugängliche Profildaten erlauben nur begrenzte Rückschlüsse.

Counter Speech kann einfach erlernt werden

Die drei Eimalinterventionen sind unterschiedlich anspruchsvoll. Sie können mit etwas Übung einfach erlernt werden.

Counter Speech hat seinen Preis

Inhaltlich und technisch anspruchsvolle Arbeit bedarf ausreichender zeitlicher, personeller und finanzieller Ressourcen. Der Bedarf ist nicht zu unterschätzen.

KI hat Potenzial in der Formulierung von Counter Speech

Der von KI vorformulierte Counter Speech hat viel Verbesserungspotenzial. Dies kann nun trainiert werden.

