Die Presse
Automatische Gegenrede für Respekt im Netz
Die Presse/Österreich Morgen | Seite 31 | 20. Juni 2020
Auflage: 79.632 | Reichweite: 317.000
Tunnel23

Automatische Gegenrede für Respekt im Netz

Menschenrechte. Ein Wiener Forscherteam erstellt Algorithmen, die gegen Rassismus in sozialen Medien aktiv werden sollen: Mit beruhigenden Meldungen, die Hassposter und Mitlesende zum Nachdenken bringen.

VON VERONIKA SCHMIDT

Die Gegenrede ist ein wichtiges Instrument gegen Diskriminierung: Doch die Gegenredner fühlen sich oft in der Minderheit. [Reuters]

Rassismus und Diskriminierung sollte niemand aushalten müssen", sagte Justizministerin Alma Zadić, die selbst zur Zielscheibe von Hasspostings wurde, im Jänner zur „Presse". Das Phänomen „Hass im Netz" betrifft viele, und das nicht erst seit den jüngsten Entwicklungen: Allein die von der Antidiskriminierungsstelle Steiermark entwickelte App „BanHate" erfasste 5500 Fälle von Hasspostings seit 2017 in Österreich.

Ein Projekt des Ludwig-Boltzmann-Instituts (LBI) für Menschenrechte, gefördert von der Stadt Wien, erstellt nun eine Basis dafür, dass Rassismus im Internet automatisch erkannt wird und dass solche Meldungen nicht unkommentiert weiteren Hass schüren. Das Konzept heißt „Gegenrede" (Counter Speech) und wird als Strategie gegen Angriffe auf sozialen Medien schon lang den Betroffenen beigebracht.

„Das Ziel von ‚Gegenrede' ist weniger ein Umstimmen derjenigen, die Hass verbreiten, als eine Bewusstseinsbildung: Man versucht, Menschen zum Nachdenken zu bringen, was sie mit ihren Hasspostings auslösen. Und dabei den vielen Mitlesenden, die sich nicht aktiv einbringen, zu zeigen, dass das nicht okay ist", erklärt Barbara Liegl, die das Projekt Counter-Bot und die Abteilung für Asyl, Anti-Diskriminierung und Diversität am LBI für Menschenrechte leitet – und eine der Geschäftsführerinnen von Zara (Zivilcourage und Anti-Rassismus-Arbeit) ist.

Kalmieren ist ermüdend

Rainer Alexandrowicz vom Institut für Psychologie der Uni Klagenfurt fügt hinzu: „Gegenrede soll die ungehinderte Abfolge von Meldungen, die sich immer weiter zu übertreffen versuchen, kalmierend unterbrechen." Aus Experimenten und Verhaltenstrainings ist bekannt, dass das Einwerfen von beruhigenden Meldungen hassproduzierende Dynamik durchbrechen kann. Das gilt sowohl offline als auch online. „Auch bei Fällen, die nicht strafrechtlich relevant sind, ist die Gegenrede ein sehr wichtiges Instrument, um auf gesamtgesellschaftlicher Ebene Respekt und Nichtdiskriminierung auszubreiten", sagt Liegl.

Ihr Team, an dem auch das Institut für Sprachwissenschaft der Uni Wien und die IT-Entwickler von Tunnel23 beteiligt sind, will nun künstliche Intelligenz dafür einsetzen, dass respektvolle Gegenrede im Netz automatisch geschieht: durch Algorithmen, die man Bots nennt, also digital erstellte Postings in sozialen Medien, die nicht von einem Menschen eingetippt werden. „In Trainings hat sich gezeigt, dass Menschen, die sich zur Wehr setzen und der Hassrede entgegentreten, viel schneller ermüden als jene, die Hass produzieren", sagt Liegl. Auch wenn die Gegenredner in einem Setting in der Mehrzahl sind, haben diese Personen meist das Gefühl, gegenüber den Menschen, die Hass produzieren, in der Minderheit zu sein.

„Die Gegenrede ist ermüdend, sodass viele einfach aufgeben." Der Psychologe Alexandrowicz erklärt, dass hier ein heuristisches Prinzip wirksam ist, bei dem die Wahrnehmung der negativen Meldungen überwiegt, auch wenn die positiven Stimmen real mehr sind: „Das ist, wie wenn Sie nach der zweiten roten Ampel denken, dass Sie eine rote Welle haben, obwohl Sie davor schon mehr grüne Ampeln durchfahren haben."

Das Computerprogramm Counter-Bot soll also den Druck von Menschen nehmen, die sich im Internet für Respekt und Antidiskriminierung einsetzen. Die Software müsste sich in Diskussionen einklinken, etwa mit Texten wie „Ihre Äußerungen sind rassistisch. Ich finde das schlimm, hören Sie auf damit!" – oder Fotos und Videos als Antwort auf Hass posten. „Bei der Entwicklung müssen wir überlegen, welche Strategien zu welchem Ziel führen sollen", sagt Liegl. Soll der Counter-Bot Personen unterstützen, die Hasspostings ausgesetzt sind, oder soll er die Person, die Hass verbreitet, dazu bewegen aufzuhören? „All diese komplexen Vorgänge müssen in die künstliche Intelligenz integriert werden", so Liegl.

Signalwörter und Sarkasmus

Derzeit arbeiten die Forscher an der Erstellung von Signalwörterlisten, mit denen das System trainiert werden kann, in Zukunft rassistische Äußerungen zu erkennen: nicht nur solche gegen schwarze Menschen, sondern breit gefächert von antisemtisch, antimuslimisch bis antiziganistisch. „Dabei ist Sarkasmus eine besondere Schwierigkeit", sagt Alexandrowicz. Auch Emojis, Jugendslang und Dialekte stellen die beteiligten Sprachwissenschaftler und Statistiker noch vor Herausforderungen, wenn Twitter-Postings analysiert werden.

„Wichtig ist auch, dass die künstliche Intelligenz sich nicht in eine Richtung entwickelt, die wir selbst nicht mehr steuern können", betont Liegl. D. h. schon bei der Planung soll verhindert werden, dass ein Counter-Bot irgendwann selbst die Hassposter rassistisch beschimpft oder gegen gut gemeinte Äußerungen vorgeht.

IN ZAHLEN

1070 **Meldungen** rassistischer Diskriminierung (von 1950 Meldungen, die 2019 bei Zara eingingen), bezogen sich auf Rassismus im Internet. Davon konnten 65 Prozent nicht strafrechtlich verfolgt werden.

87 **Prozent** der rassistischen Fälle, die während der strengen Ausgangsbeschränkungen im März und April gemeldet wurden, passierten online.