Wie Algorithmen lernen, fair zu sein

Wenn im Netz bestimmte Gruppen kaum vertreten sind, kann das zur Folge haben, dass Künstliche Intelligenz (KI) Vorurteile übernimmt, die gegenüber diesen Gruppen bestehen. An der IT-Universität Kopenhagen hat man dieses Problem erkannt. Man versucht nun, Algorithmen so zu programmieren, dass sie niemanden diskriminieren. Doch das ist nicht so leicht.

Manuskript

SPRECHERIN:
Je **diverser**, desto besser: Das sollte für jede Gruppe gelten, die einen Lernd**atensatz** für ein KI-Training **zusammenstellt**. Denn: Kommt eine **Sichtweise** nicht vor, entstehen Fehler. Das berichten auch die Forschungsgruppen der IT-Universität Kopenhagen.

LEON DERCZYNSKI (Forscher):
Hassrede zu erkennen, ist schwierig, denn manchmal ist es nur eine **Minderheit**, eine bestimmte Gruppe von Menschen, die sie erkennen kann. Wir haben beobachtet, dass es Fälle gab, wo **sich** alle **einig waren** – alle bis auf eine Person. Zum Beispiel gab es eine **Muslima**, und sie hat manche Worte erkannt, mit denen immer **abfällig** über Muslime gesprochen wurde. Die anderen kannten die gar nicht.

SPRECHERIN:
Ist eine Gruppe nicht **repräsentiert**, entwickelt die KI **Vorurteile**. Dazu kommt: Eine KI lernt ständig weiter. Wenn also im Netz einige Gruppen weniger **Wortmeldungen** haben als der Rest, denkt die KI, deren Meinungen seien Minderheitsmeinungen.

LEON DERCZYNSKI:
Zum Beispiel in den **virtuellen** Räumen, in denen es viel **Frauenfeindlichkeit** gibt, **melden sich** Frauen natürlich weniger **zu Wort**. Die KI denkt dann, dass **Ansichten** wie **geschlechtergerechte** Bezahlung **extremistisch** seien, weil diese Ansichten eben dort kaum vorkommen.

SPRECHERIN:
Deswegen arbeitet eine Forschungsgruppe der IT-Universität an einer KI, die alle Menschen online fair behandelt. Ein Algorithmus soll Hassrede im Netz erkennen, ohne dabei eine bestimmte Gruppe zu diskriminieren. Doch das ist nicht leicht: Inhalte im Netz sind oft nicht **eindeutig**. **Schlagworte** für das Training dieser KI zu erstellen, ist schwierig.

EMIL LYSDAHL FAHRENHOLTZ (Student):
Ein Beispiel könnte hier sein: „Ich hasse alle Dänen.“ Wir haben hier eine Liste mit **Identitäten**. Dänen, englische Menschen – das kann alles Mögliche sein. Und dann **vergeben** wir verschiedene **Kategorien**: Ist das **verletzend** gegenüber einer bestimmten Identität? Und in welche Richtung geht es? Ist es allgemein? Oder geht es um eine bestimmte Person? Um mich?

SPRECHERIN:
Bis der Algorithmus fertig entwickelt ist, wird es noch einige Monate dauern. Firmen wie Facebook könnten in Zukunft **von** solchen **Modellen profitieren**.

Glossar

Algorithmus, Algorithmen (m.) – eine Reihe von Vorschriften und Befehlen, damit ein Computer bestimmte Probleme lösen kann (hier auch: die künstliche Intelligenz)

fair (aus dem Englischen) – hier: so, dass alle die gleichen Chancen haben

Künstliche Intelligenz (f., nur Singular) – ein Programm, das selbstständig Dinge erkennen, lernen und entscheiden kann (Abkürzung: KI)

jemanden diskriminieren – jemanden schlecht behandeln, weil er anders ist

divers – hier: vielfältig; unterschiedlich; so, dass auch Menschen aus Minderheiten vertreten sind

Datensatz, -sätze (m.) – eine Gruppe von Daten, die inhaltlich zusammengehören

etwas zusammen|stellen – hier: Teile für etwas suchen und miteinander kombinieren

Sichtweise, -n (f.) – die Meinung; die Art, wie man über etwas denkt

Hassrede, -n (f.) – Reden oder Texte, mit denen besonders im Internet Hass gegen bestimmte Menschen oder Gruppen verbreitet werden soll

Minderheit, -en (f) – hier: eine Gruppe, die anders ist als die meisten Menschen in einem Land, z. B. weil sie eine andere Religion hat oder eine andere Sprache spricht

sich einig sein – der gleichen Meinung sein

Muslim, -e/Muslima, -s – eine Person, die den Islam als Religion hat

abfällig – so, dass man seine schlechte Meinung über jemanden/etwas ausdrückt

repräsentiert – so, dass jemand/etwas vertreten ist

Vorurteil, -e (n.) – eine meist negative Meinung über jemanden/etwas, ohne jemanden/etwas richtig zu kennen

Wortmeldung, -en (f.) – die schriftliche oder mündliche Äußerung einer Person; der Redebeitrag

virtuell – so, dass etwas nur am Computer oder im Internet existiert

Frauenfeindlichkeit (f., nur Singular) – die Tatsache, dass man schlecht über Frauen denkt und/oder ihnen schaden will

sich zu Wort melden – sich schriftlich oder mündliche äußern

Ansicht, -en (f.) – hier: die Meinung

geschlechtergerecht – so, dass alle Geschlechter gleichermaßen vertreten sind bzw. gleich behandelt werden

extremistisch – radikal; so, dass man eine extreme Position vertritt

eindeutig – ohne Zweifel; ganz klar

Schlagwort, -e (n.) – hier: ein Stichwort oder ein einzelner Begriff

Identität, -en (f.) – hier: das innere Wesen von jemanden; das, was jemanden als Person ausmacht

etwas vergeben – hier: einer Sache ein bestimmtes Merkmal zuordnen

Kategorie, -n (f.) – der Bereich; eine Gruppe von Dingen oder Personen mit bestimmten gemeinsamen Merkmalen

verletzend – hier: so, dass jemand beleidigt wird

Modell, -e (n.) – hier: die Idee; das Konzept

von etwas profitieren – einen Vorteil durch etwas haben

Autorin/Autor: Katja Liersch, Philipp Reichert