Gesellschaftliche Stereotype in Sprachmodellen: Herausforderungen und Stand der Forschung

Software mit Künstlicher Intelligenz kommt in zahlreichen Anwendungen zum Einsatz. Die Stereotype der Gesellschaft, die in den Modellen enthalten sein können, bergen jedoch Herausforderungen.

Mascha Kurpicz-Briki

Software mit Künstlicher Intelligenz ist spätestens seit der Veröffentlichung des Chatbots ChatGPT in aller Munde. Diese Art von Software basiert auf dem Konzept des maschinellen Lernens, bei welchem der Computer anhand von Beispielen «lernt» und sich dabei auf grosse Mengen von Daten stützt. Maschinelles Lernen ist jedoch nicht nur für Chatbots wichtig. Verschiedene Arten von Daten wie Bilder, strukturierte Daten (vergleichbar mit einer Tabelle in Excel), Texte oder Videos können so verarbeitet werden und eine ganze Reihe von Anwendungen ermöglichen. Es können zum Beispiel Daten in verschiedene Gruppen eingeteilt werden – das nennt man Klassifikation oder *Clustering* – oder Vorhersagen für die Zukunft basierend auf historischen Daten gemacht werden.

Alle diese Anwendungen des maschinellen Lernens haben gemeinsam, dass sie typischerweise auf einer grossen Menge von Daten basieren. Diese Daten werden meistens durch unsere Gesellschaft geschaffen und bilden daher ganz verschiedene Aspekte davon ab. Während es oftmals sehr nützliche Beziehungen in den Daten gibt, die eben genau dieses maschinelle Lernen ermöglichen, so enthalten sie auch unerwünschte Rollenbilder und Stereotypen.

In den letzten Jahren wurde eine ganze Reihe von Fällen bekannt, wie solche ungewünschten Beziehungen durch KI-Anwendungen wiedergegeben wurden oder gar zu Diskriminierung geführt haben. Ein Beispiel war die Rekrutierung bei einer grossen Firma der amerikanischen Tech-Branche. Die KI sollte diese dabei unterstützen, im Rekrutierungsprozess eine Vorselektion vorzunehmen. Doch das System lernte, männliche Kandidierende zu bevorzugen: Die Künstliche Intelligenz wurde nämlich auf der Basis der ehemaligen Bewerbungen trainiert, welche in der Vergangenheit mehrheitlich von Männern eingereicht wurden.[1] Andere bekannte Beispiele gibt es im Bereich der Bilderkennung. Eine Recherche der BBC zeigte, dass Frauen mit dunklerer Haut mit doppelt so grosser Wahrscheinlichkeit wie hellhäutige Männer in der online Passregistrierung die Meldung bekamen, dass ihre Fotos nicht den britischen Passregeln entsprechen.[2] Auch in Sprachmodellen können Stereotype oder unerwünschte Beziehungen abgebildet sein.

Wie kommt es zu Stereotypen in Sprachmodellen?

Sprachmodelle bilden die Grundlage für Dialogsysteme wie ChatGPT oder Google Bard, kommen aber auch in anderen Bereichen zum Einsatz, wie beispielsweise der automatischen Textanalyse, der maschinellen Übersetzung oder bei Suchmaschinen. Diese Modelle werden mit einer grossen Menge von Texten trainiert und lernen so, wie verschiedene Wörter in einer bestimmten Sprache zusammenhängen.

Eine mögliche Variante, solche Modelle zu trainieren, ist das sogenannte *Masked Language Modelling*.[3] Dabei werden in einer grossen Textmenge einzelne Wörter abgedeckt und die KI wird darauf trainiert, diese fehlenden Wörter vorherzusagen. Das abgedeckte Wort aus den Trainingsdaten ist bekannt und wird daher als richtige Antwort gewertet. Je nachdem, welche Textbeispiele in den Trainingsdaten vorkommen, haben diese eine Auswirkung auf die resultierenden Modelle. Für dieses Training wird eine so unvorstellbar grosse Textmenge benötigt, dass diese nicht mehr manuell kontrollierbar ist. Diese Texte wurden durch unsere Gesellschaft produziert und widerspiegeln diese. Das bedeutet, dass nicht mehr sichergestellt werden kann, dass keine unerwünschten, falschen oder sonst ungeeigneten Beispieldaten in einem Datensatz enthalten sind.

Mascha Kurpicz-Briki

«Die Forschung hat an verschiedenen Beispielen gezeigt, dass bekannte gesellschaftliche Vorurteile in Sprachmodellen messbar sind.»

«Je nach verwendetem Adjektiv wird der Satz eher männlich oder weiblich übersetzt.»

Was sind die Auswirkungen von Stereotypen in Modellen?

Die Trainingsdaten haben Auswirkungen auf die Anwendungen, welche die Sprachmodelle anschliessend verwenden. Die Forschung hat an verschiedenen Beispielen gezeigt, dass bekannte gesellschaftliche Vorurteile in Sprachmodellen messbar sind.[4] Auch bei oft verwendeten Tools zur maschinellen Übersetzung lassen sich solche Rollenbilder beobachten, beispielsweise wenn das Geschlecht der genannten Person in einer Sprache nicht eindeutig definiert ist. Übersetzen wir beispielsweise «the student» von Englisch nach Deutsch, so gibt es auf Deutsch zwei Möglichkeiten: «der Schüler» oder «die Schülerin». Das System muss daher eine Vorhersage machen, wenn nur ein Ergebnis angezeigt werden soll. Zum Zeitpunkt des Verfassens dieses Artikels ging ein bekanntes Tool der maschinellen Übersetzung mit folgenden Beispielen so um: «The smart student» wurde mit «der kluge Schüler», «the ugly student» hingegen mit «die hässliche Schülerin» übersetzt. Je nach verwendetem Adjektiv wird der Satz eher männlich oder weiblich übersetzt. Ähnlich verhält es sich bei Berufen: «I am a professor» wird mit «Ich bin ein Professor», «I am a medical assistant» jedoch mit «Ich bin eine medizinische Assistentin» übersetzt. Die maschinelle Übersetzung ist nur eine Art, wie Sprachmodelle angewendet werden. Das Problem betrifft auch andere Arten von Modellen und die Reproduktion von gesellschaftlichen Stereotypen kann beispielsweise auch bei der Generierung von Bildern beobachtet werden.

Herausforderungen und Forschungsfragen

Die Frage der Stereotype in KI-Modellen und insbesondere Sprachmodellen ist Bestandteil aktueller Forschung. Es ist eine grosse Herausforderung, eine technische Lösung für dieses Problem zu entwickeln, denn viele Dimensionen kommen zusammen. Einerseits sind unterschiedliche Arten von Daten betroffen (z. B. Bilder, strukturierte Daten, Texte) und verschiedene Charakteristiken (Alter, Herkunft, Geschlecht, etc.) können darin ganz unterschiedlich zum Ausdruck kommen. Andererseits können die Stereotype auch intersektional sein, wenn Personen zwei oder mehr Kategorien angehören, beispielsweise ältere homosexuelle Personen oder Women of Color.[5]

Das Projekt *BIAS: Mitigating Diversity Biases of AI in the Labor Market*[6] (2022–2026) ist ein europäisches Forschungsprojekt aus dem Horizon Europe Programm, welches durch das Staatssekretariat für Bildung, Forschung und Innovation SBFI mitfinanziert wird. Es setzt sich mit der Frage auseinander, wie sich solche Stereotype in Sprachmodellen und KI-Anwendungen auf dem Arbeitsmarkt auswirken. In einer ersten Phase wurden in verschiedenen europäischen Ländern inklusive der Schweiz Interviews mit 70 Personen aus den Bereichen Personalmanagement und KI-Entwicklung durchgeführt.[7] Die meisten Befragten hatten grundsätzlich eine positive Haltung gegenüber dem Einsatz von KI-Technologien im Rekrutierungsprozess und bei der Personalselektion. Es bestanden jedoch grössere Bedenken, was den Einsatz der Technologie beim Personalmanagement betrifft.

Die Forschung im technischen Bereich des Projekts BIAS setzt sich unter anderem mit der Frage auseinander, wie die in Sprachmodellen vorhandenen Stereotype messbar gemacht und anschliessend reduziert werden können. Das Projekt widmet sich insbesondere den europäischen Sprachen und basierend darauf den verschiedenen regionalen Stereotypen aus den beteiligten Ländern. Ausserdem versucht das Projekt, eine Lücke zu schliessen, was die Berücksichtigung von Intersektionalität bei der Untersuchung von Stereotypen in KI-Modellen betrifft.

Kommen wir auf das Beispiel der Rollenbilder und Stereotype in Sprachmodellen zurück. Diese können nämlich auch dann ein Risiko für Anwendungen sein, die solche Sprachmodelle als Grundlage verwenden, wenn relevante Daten wie Name oder Geschlecht explizit aus dem Entscheidungsprozess ausgeschlossen werden. So kann es sein, dass indirekte Beziehungen bestehen zwischen den verschiedenen Wörtern und vermeintlich unauffällige Wörter wie Interessen oder Hobbys durch das Modell einem Geschlecht zugeordnet werden. Wie sich dies auf Empfehlungen oder generierte Texte von KI-Software auswirkt, ist Teil der technischen Forschung im Projekt BIAS.

Eine weitere wichtige Komponente des Forschungsprojekts ist der Einbezug von Stakeholdergruppen aus verschiedenen Bereichen wie HR, Informatik, Recht und der breiten Öffentlichkeit. Interessierte Personen können sich für verschiedene Arbeitsgruppen, sogenannte *National Labs*[8], anmelden. Dadurch soll das Bewusstsein in der Gesellschaft für diese Problematik entwickelt und über den Verlauf des Projekts informiert werden.

Ausblick und Empfehlungen

Das Problem von Stereotypen in Sprachmodellen und anderen KI-Anwendungen ist eine grosse technische Herausforderung und von der Forschung noch nicht abschliessend gelöst. Trotzdem wird KI in ganz unterschiedlichen Anwendungsfällen eingesetzt. Eine Berücksichtigung dieser Problematik bei der Planung, der Entwicklung, dem Einkauf und dem Einsatz solcher Tools ist unverzichtbar, setzt jedoch kritisches Bewusstsein für das Risiko voraus. Je nach Art einer Anwendung oder eines Anwendungsfalles kann die beschriebene Problematik mehr oder weniger relevant sein.

KI-Technologien bringen gewisse Risiken mit sich, wie etwa die hier beschriebenen Stereotype, haben jedoch auch grosses Potenzial für verschiedene Anwendungsfälle, wenn sie verantwortungsvoll eingesetzt werden. Aufgrund der Grenzen der Technologie ist es bei sensiblen Anwendungsfällen wichtig, die Menschen ins Zentrum zu stellen. Die Technologie muss als wirksames Werkzeug eingesetzt werden, um Menschen zu unterstützen, anstatt anzustreben, sie zu ersetzen.

Prof. Dr. Mascha Kurpicz-Briki forscht an der Berner Fachhochschule im Bereich KI und Sprachmodelle. Sie ist Autorin des Buchs «More than a Chatbot: Language Models Demystified».

«Die Technologie muss als wirksames Werkzeug eingesetzt werden, um Menschen zu unterstützen, anstatt anzustreben, sie zu ersetzen.»

Anmerkungen

1 Dastin, Jeffrey: Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of data and analytics, 2022, S. 296–299.

2 Ahmed, Maryam: UK passport photo checker shows bias against dark-skinned women. BBC News, 08.10.2020. www.bbc.com/news/technology-54349538 (abgerufen am: 05.05.2024).

3 Kurpicz-Briki, Mascha: More than a Chatbot: Language Models Demystified. Springer, 2023.

4 Siehe z.B.: Caliskan, Aylin / Bryson, Joanna J. / Narayanan, Arvind: Semantics derived automatically from language corpora contain human-like biases. In: Science 356, 2017, S. 183–186. Oder: Guo, Wei / Caliskan, Aylin: Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, S. 122–133.

5 Søraa, Roger: AI for Diversity. CRC Press, 2023.

6 Projekt BIAS: https://www.biasproject.eu/ (abgerufen am: 25.06.2024).

7 Rigotti, Carlotta et al.: Mitigating Diversity Biases of AI in the Labor Market. In: EWAF'23: European Workshop on Algorithmic Fairness. Winterthur, Switzerland, 2023. https://ceur-ws.org/Vol-3442 (abgerufen am: 05.07.2024).

8 Weitere Informationen zu den National Labs: www.biasproject.eu/nationallabs.

Abstracts

Les stéréotypes sociaux dans les modèles langagiers : défis et état de la recherche

Les logiciels d'intelligence artificielle tels que les modèles langagiers identifient beaucoup de liens utiles entre les mots, mais ils peuvent aussi véhiculer des éléments indésirables, comme des stéréotypes sur les rôles de genre. **Mascha Kurpicz-Briki,** professeure à la Haute école spécialisée bernoise, explique en quoi cela peut être problématique pour les applications utilisant ce genre de modèles. Le projet de recherche européen BIAS étudie comment ces préjugés se retrouvent dans les applications d'IA utilisées sur le marché du travail et ce qui peut être fait pour y remédier.

Stereotipi sociali nei modelli linguistici: sfide e stato della ricerca

I modelli linguistici basati sull'intelligenza artificiale riconoscono molte relazioni utili tra le parole, ma possono anche contenere elementi indesiderati come stereotipi di ruoli sociali. In questo articolo, la professoressa **Mascha Kurpicz-Briki** della Berner Fachhochschule spiega perché ciò possa costituire un problema per le applicazioni che utilizzano tali modelli. Il progetto di ricerca europeo BIAS sta esaminando in quale modo questi bias si riflettano nelle applicazioni basate sull'intelligenza artificiale per il mercato del lavoro e cosa può essere fatto per contrastarlo.