Wie neue Werkzeuge Bias in KI-Systemen erkennen und abschwächen

Von Mascha Kurpicz-Briki (BFH Technik & Informatik) | 0 Kommentare

Voreingenommenheit in der KI ist eine grosse Herausforderung für die heutigen Technologien der künstlichen Intelligenz (KI). Die im Horizon Europe-Projekt BIAS entwickelte Proof-of-Concept-Technologie adressiert dieses Problem. Die ersten Ergebnisse wurden kürzlich auf dem EWAF'23: European Workshop on Algorithmic Fairness von BFH-Forscher*innen in Zusammenarbeit mit Partner*innen der Universität Leiden vorgestellt.

In den letzten Jahren haben viele Beispiele das Problem der Voreingenommenheit bei der Anwendung von künstlicher Intelligenz aufgezeigt. Voreingenommenheit bei der Rekrutierung von Frauen durch KI [2], rassistische Chatbots [3] oder Passkontrollsysteme, die Vorurteile gegenüber Women of Color aufdecken [1], sind nur einige Beispiele dafür, was passieren kann, wenn Voreingenommenheit in Trainingsdaten oder -modellen für maschinelles Lernen enthalten ist.

Es gibt viele Herausforderungen, um Vorurteilen zu erkennen und zu entschärfen. KI-Technologien kommen in vielen verschiedenen Formen vor und erfordern daher unterschiedliche Trainingsdaten wie Videos, Bilder, Texte oder strukturierte Daten. Unterschiedliche Trainingsdaten erfordern möglicherweise eine unterschiedliche Erkennung und Reduzieren von Verzerrungen. Auch können die Stereotypen der Gesellschaft, die zu einer solchen Voreingenommenheit in den Daten führen, auf sehr unterschiedliche Attribute wie Geschlecht, Herkunft, Alter und viele andere persönliche Merkmale ausgerichtet sein. Ausserdem kann die Verzerrung intersektional sein. Die Voreingenommenheit kann auch in verschiedenen Phasen des Entwicklungsprozesses eingeführt werden, z. B. in den Trainingsdaten, die eine historische Voreingenommenheit der Gesellschaft widerspiegeln, oder in späteren Phasen des Prozesses bei der Anwendung oder Entwicklung der Technologie. Schliesslich ist auch die Definition von Fairness eine Herausforderung. Was fair ist oder nicht, kann für verschiedene Menschen unterschiedliche Dinge bedeuten und erfordert daher zusätzliche Untersuchungen für jeden spezifischen Anwendungsfall.

KI auf dem Arbeitsmarkt

KI-Technologien werden zunehmend auch im Kontext des Arbeitsmarktes eingesetzt. Die Fragen von Fairness und Voreingenommenheit in diesem Zusammenhang werden in dem von Horizon Europe geförderten Projekt «BIAS» untersucht. In einer ersten Runde der Feldforschung wurden 70 Personalmanager*innen und KI-Entwickler*innen in verschiedenen europäischen Ländern befragt [4]. Im Allgemeinen standen die Teilnehmer*innen dem Einsatz von KI-Anwendungen, die den Einstellungs- und Auswahlprozess unterstützen, positiv gegenüber, einige äusserten jedoch Bedenken hinsichtlich der Verwaltung von Mitarbeiter*innen unter Einbeziehung von KI. Die Teilnehmer*innen forderten auch die Annahme von Massnahmen zur Abschwächung von Vorurteilen in Bezug auf die Vielfalt in diesem Zusammenhang, wobei der Schwerpunkt auf geschlechtsspezifischen Vorurteilen lag.

Ziel des BIAS-Projekts ist die Einführung von Methoden zur Erkennung und Abschwächung von Verzerrungen in KI-Anwendungen, insbesondere in Sprachmodellen, sowie die Entwicklung einer fairen Entscheidungsfindung im Kontext von HR-Anwendungen. Zu diesem Zweck wird im Rahmen des Projekts eine Proof-of-Concept-Technologie, der Debiaser, entwickelt.

Der Debiaser

Der Debiaser besteht aus drei verschiedenen Komponenten, wie in Abbildung 1 dargestellt.

Abbildung 1: Die verschiedenen Komponenten des Debiaser.

Zunächst wird untersucht, wie eine faire Rekrutierung mit Hilfe von *Case Based Reasoning* durchgeführt werden kann. Dies erfordert eine anwendungsfallspezifische Definition von Fairness, die sicherstellt, dass ähnliche Kandidaten auf ähnliche Weise behandelt werden. Der Kern des Case-Based Reasoning ist die Idee, ähnliche Probleme durch mit bereits eingesetzten Lösungen zu beheben, die zuvor auf ähnliche Probleme angewandt wurden und manuell von Menschen kuratiert wurden.

Der Debiaser untersucht darüber hinaus, wie Anwendungen aus dem Bereich der natürlichen Sprachverarbeitung (NLP) in diesem Zusammenhang verzerrt werden können. Auf der einen Seite betrachtet er die textbasierte Entscheidungsfindung. Ziel ist es, zu erklären, wie Entscheidungen getroffen werden, und somit potenzielle Verzerrungen in den Trainingsdaten aufzudecken und Methoden vorzuschlagen, um diese Verzerrungen abzuschwächen.

Schliesslich untersucht der Debiaser, wie sich gesellschaftliche Stereotypen in Sprachmodellen widerspiegeln. Sprachmodelle sind die Modelle hinter Anwendungen wie ChatGPT oder Bard. Sie kodieren die Beziehungen zwischen Wörtern in mathematischen Vektoren, so dass Berechnungen durchgeführt werden können, um zu ermitteln, ob beispielsweise zwei Wörter miteinander verwandt sind. Hier kommt die Verzerrung ins Spiel. Es hat sich gezeigt, dass es eine Verzerrung bei Wörtern aus den Bereichen Beruf/Familie und männliche/weibliche Vornamen gibt [5]. Insbesondere gibt es in der Forschung Hinweise darauf, dass diese Arten von gesellschaftlichen Stereotypen, die in Sprachmodellen kodiert sind, von der Sprache und dem kulturellen Kontext abhängen [6]. Der Debiaser zielt darauf ab, die Verzerrungen in Sprachmodellen verschiedener europäischer Sprachen zu quantifizieren und zu reduzieren (siehe Abbildung 2).

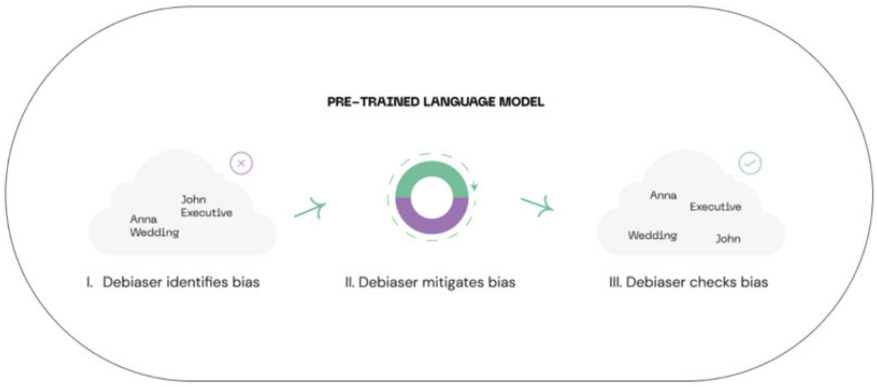


Abbildung 2: Ein Teil des Debiaser-Projekts untersucht, wie man Verzerrungen in vortrainierten Sprachmodellen verschiedener europäischer Sprachen messen und reduzieren kann.

Erste Einblicke in das Projekt und die laufenden Arbeiten am Debiaser wurden kürzlich auf dem European Workshop on Algorithmic Fairness EWAF'23 von Forscher*innen der Forschungsgruppe Angewandte Maschinelle Intelligenz der BFH in Zusammenarbeit mit Partner*innen des eLaw-Zentrums der Universität Leiden vorgestellt.

Publikation

Rigotti, C., Puttick A., Fosch-Villaronga, E., und Kurpicz-Briki, M. (2023) The BIAS project: Mitigating diversity biases of AI in the labour market. European Workshop on Algorithmic Fairness (EWAF'23), Winterthur, Schweiz, 7-9, Juni 2023. Verfügbar unter https://ceur-ws.org/Vol-3442/paper-47.pdf [https://ceur-ws.org/Vol-3442/paper-47.pdf]

Referenzen

[1] https://www.bbc.com/news/technology-54349538.amp [https://www.bbc.com/news/technology-54349538.amp]

[2] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G [https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G]

[3] https://www.bbc.com/news/technology-35902104 [https://www.bbc.com/news/technology-35902104]

[4] Rigotti, C., Puttick A., Fosch-Villaronga, E., und Kurpicz-Briki, M. (2023) The BIAS project: Mitigating diversity biases of AI in the labour market. European Workshop on Algorithmic Fairness (EWAF'23), Winterthur, Schweiz, 7-9, Juni 2023. Verfügbar unter https://ceur-ws.org/Vol-3442/paper-47.pdf [https://ceur-ws.org/Vol-3442/paper-47.pdf]

[5] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

[6] Kurpicz-Briki, M., & Leoni, T. (2021). A World Full of Stereotypes? Further Investigation on Origin and Gender Bias in Multi-Lingual Word Embeddings. Frontiers in Big Data, 4, 625290. Verfügbar unter https://www.frontiersin.org/articles/10.3389/fdata.2021.625290/full [https://www.frontiersin.org/articles/10.3389/fdata.2021.625290/full]

Link zu **eLaw – Zentrum für Recht und digitale Technologien** https://www.universiteitleiden.nl/en/law/institute-for-the-interdisciplinary-study-of-the-law/elaw [https://www.universiteitleiden.nl/en/law/institute-for-the-interdisciplinary-study-of-the-law/elaw]

Link zur **Forschungsgruppe AMI: bfh.ch/ami**

Link zum **Projekt BIAS biasproject.eu**

AUTHOR: MASCHA KURPICZ-BRIKI

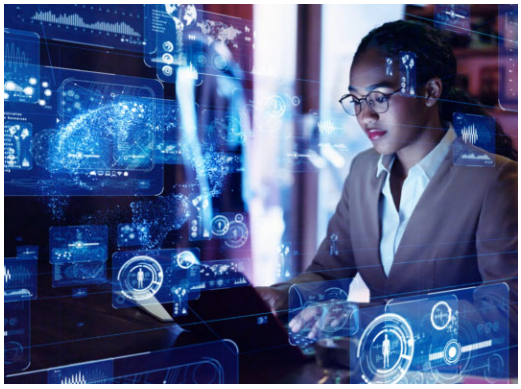
Dr. Mascha Kurpicz-Briki ist Professorin für Data Engineering am Institute for Data Applications and Security IDAS der Berner Fachhochschule, und stellvertretende Leiterin der Forschungsgruppe Applied Machine Intelligence. Sie beschäftigt sich in ihrer Forschung unter anderem mit dem Thema Fairness und der Digitalisierung von sozialen und gesellschaftlichen Herausforderungen.

Posts from Mascha Kurpicz-Briki

Create PDF

Ähnliche Beiträge

Gesellschaftliche Stereotypen in vortrainierten Sprachmodellen



Hi ChatGPT, hast du Vorurteile?

Wenn Mehmet und Peter nicht gleich sind - Vorurteile auf Grund der Namensherkunft in Wortvektoren

Wie Bias in KI-Systemen erkannt und verringert werden können

«Wir müssen Bias aus Sprachmodellen entfernen» - eine Podcastfolge über KI in

Verwaltung und Justiz

0

COMMENTS