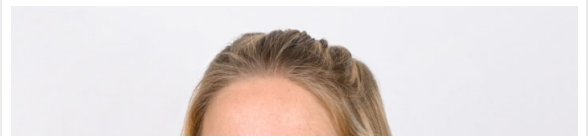
SocietyByte Wissenschaftsmagazin der Berner Fachhochschule

«Wir müssen Bias aus Sprachmodellen entfernen» – eine Podcastfolge über KI in Verwaltung und Justiz

Von Anne-Careen Stoltze (BFH Wirtschaft) | 0 Kommentare

Chat GPT ist gerade in aller Munde. Das trainierte Programm kann eigentlich über jedes Thema plaudern und Gespräche führen, die menschlich anmuten. Aber ChatGPT unterlaufen auch Fehler. In der neuesten Episode des Podcasts «Let's Talk Business» reden KI-Expertin Mascha Kurpicz-Briki [https://www.bfh.ch/de/ueber-die-bfh/personen/diqa4uibb7gl/] und Datenforscher Matthias Stürmer darüber, wie man Bias vermeiden kann, warum der Begriff Augmented Intelligence besser passt als KI und wie sie in Justiz und Verwaltung eingesetzt werden kann. Hier geht es zur Episode und zu einer gekürzten schriftlichen Version.

Heute sprechen wir über ein Thema, das seit Anfang des Jahres für Furore sorgt. ChatGPT. Das trainierte Sprachmodell kann über jedes Thema plaudern und Gespräche führen, die ziemlich menschlich anmuten. Aber ChatGPT unterlaufen auch sachliche Fehler und ein weiteres Problem sind Bias, also Vorurteile. Du hast in einem Forschungsprojekt untersucht, das Künstliche Intelligenz diskriminiert. Wie zeigt sich das?



Prof. Dr. Mascha Kurpicz-Briki ist stellvertretende Leiterin der Applied Machine Intelligence Research Group am Departement Technik und Informatik der BFH.

Mascha Kurpicz-Briki: Bias ist ein grosses Problem. Es gibt bereits Verzerrungen in den Trainingsdaten, auf denen solche Systeme trainiert wurden. Dies sind Daten, die unsere Gesellschaft produziert hat. Das heisst, die Daten enthalten alle die Diskriminierungen und Stereotypen, die wir in der Gesellschaft haben. Wenn wir dann die KI darauf trainieren, dann sind automatisch diese Stereotypen vorhanden. Das ist allgemein ein Problem und auch im Bereich der automatischen Sprach- oder Textverarbeitung. Das ist das, was wir Natural Language Processing nennen und wo eben auch ChatGPT drunter fällt. Am Anfang war es einfacher, einen Bias nachzuweisen. Inzwischen reagiert das System häufig sehr ausweichend, indem es sagt: «Ich möchte mich nicht dazu äussern.» Oder: «Ich bin nur ein Sprachmodell und das Thema ist mir zu heikel.»

Müsste man ChatGPT also mit diskriminierungsfreien Daten trainieren? Und geht das überhaupt?

Mascha Kurpicz-Briki: Das Thema wird in der Forschung und von grossen Firmen angegangen. Aber es ist gar nicht so einfach. Das Einfachste wäre, wenn wir als Gesellschaft einfach aufhören würden, Stereotypen zu wiederholen oder in den Daten zu codieren. Aber schlussendlich ist das nicht von einem Tag auf den anderen möglich. Das heisst, wir arbeiten mit Daten, die historisch entstanden sind, sie sind ein Produkt ihrer Zeit. Es ist Teil unserer Forschung, diese Verzerrungen aus den Trainingsdaten und Sprachmodellen rauszubringen.

Matthias, ihr entwickelt an eurem Institut derzeit ein Sprachmodell für juristische Texte. Was kann ich mir darunter vorstellen?

Matthias Stürmer: Wir fokussieren uns auf rechtliche Themen aus dem Verwaltungskontext. Wir prüfen, ob Gerichtsurteile, die anonymisiert sind, de-anonymisiert werden können, um Personen zu re-identifizieren. Das wäre eine Gefahr für die Privatsphäre. Dennoch werden Gerichtsurteile veröffentlicht, damit Transparenz über das Gerichtswesen besteht. Wir prüfen zusammen mit Juristen, was sind die Möglichkeiten oder eben zum Glück auch die Grenzen der KI, dass das nicht so schnell möglich ist. In dem Zusammenhang haben wir untersucht, wie juristische Texte aufgebaut sind. Gerichtsurteile haben klare Strukturen und gewisse Absätze haben bestimmte Funktionen. Dabei haben wir gemerkt, dass es ein Sprachmodell, was auf juristische Terminologie fokussiert ist sowie Deutsch, Französisch und Italienisch versteht, noch gar nicht gibt. Diese Grundlagen erarbeiten wir nun.

Wofür ist das Anonymisieren der Gerichtsurteile wichtig?

Prof. Dr. Matthias Stürmer leitet das Institut Public Sector Transformation am Departement Wirtschaft der BFH.

Matthias Stürmer: Die Anonymisierung wird generell in Gerichten praktiziert, damit die Privatsphäre der Klägerparteien gewahrt bleibt. Wenn Firmen- und Personennamen anonymisiert sind, kann der Inhalt des Gerichtsfalls veröffentlicht werden. Unser Forschungsprojekt heisst deshalb auch Open Justice versus Privacy, denn es ist ein Spannungsfeld: Man will einerseits die Privatsphäre schützen und andererseits Transparenz, damit Urteile insbesondere Bundesgerichtsentscheide nachvollziehbar sind. Bei Jurist*innen gibt es die Angst, dass man mit Big Data, mit künstlicher Intelligenz, plötzlich einfach Daten de-anonymisieren kann. Aber das konnten wir schon widerlegen. Selbst ChatGPT kann Gerichtsurteile nicht de-anonymisieren und auch mit vertretbarem Aufwand ist es nicht zu machen.

Neben dem Datenschutz – welche Risiken gibt es noch?

Matthias Stürmer: Bias und Diskriminierung sind sicher ein Riesenthema. Womit wir uns aber noch beschäftigen, ist die Idee der digitalen Souveränität. Heute werden sehr viele KI-Modelle von Firmen aus den USA und China produziert. Wir möchten aufzeigen, dass die Schweiz mit eigenen Ressourcen solche Modelle herstellen kann. Denn all die bestehenden Chat-Programme sind eine Blackbox: Wir wissen nicht, wie, mit welchen Daten trainiert wurde, mit welchen Algorithmen und mit welchen Schutzmechanismen die ausgestattet sind. Wir wollen das transparent machen und zeigen. Damit wir weniger abhängig sind und Aspekte wie Fake News eingedämmt werden. Das ist letztlich für die Demokratie und unsere Gesellschaft wichtig.

Wofür kann euer Modell später angewendet werden?

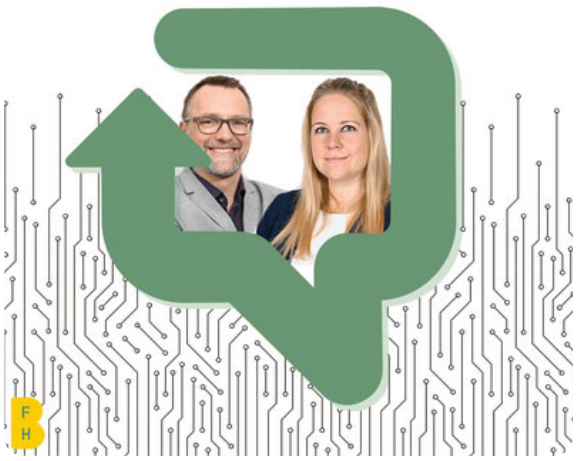
Matthias Stürmer: Das ist wie ein Fundament, auf dem verschiedene Anwendung gebaut werden. Ich vergleiche es oft mit einem Motor, quasi einem Düsenjet, den man an verschiedenen Orten einsetzen kann. In unserem Fall generieren wir fürs Bundesgericht ein Sprachmodell, das sehr gut anonymisiert. Bisher wurde das manuell, also sprich mit Suchen und Ersetzen-Befehlen gemacht. Was schon sehr gut funktionierte, aber oftmals sind in Gerichtsurteilen auch Personenidentifizierende Merkmale drin und das soll unser Sprachmodell noch besser herausfiltern. Zudem haben wir experimentiert: Was ist der Unterschied zwischen generischen und spezifischen Sprachmodellen? Letzteres erkennt beispielsweise Lückentexte viel besser und kann die Lücken präziser füllen als ein generisches Sprachmodell. Und jetzt kann man sich beliebige Ideen ausdenken.

Welche zum Beispiel?

Matthias Stürmer: Eine Idee ist es die oft schwer verständlichen juristischen Texte einfacher zu formulieren. Beim Natural Language Processing gibt es einen Prozessschritt, der heisst Textvereinfachung. Mit diesem kann man eine einfachere juristische Sprache produzieren. Wir haben es zwar noch nicht ausprobiert, aber das ist ein bekanntes Forschungsgebiet in der NLP. Damit werden schwierige Texte inklusiver, so dass alle Menschen sie gut verstehen können.

Mascha Kurpicz-Briki: Das ist ein sehr schönes Beispiel, an dem man sieht, wie Sprachmodelle zur Inklusion eingesetzt werden können. Weil einerseits haben wir zwar die Diskriminierungsgefahr, aber andererseits gibt es viele spannende Use Cases, wo Sprachmodelle uns sehr gut nützen.

Dies ist eine gekürzte Version des Gesprächs, die ganze Länge hören Sie hier:



Let's Talk Business #Episode 14: KI-Nutzung in Justiz un...

Links zum Thema

Institut Public Sector Transformation [https://www.bfh.ch/de/forschung/forschungsbereiche/public-sector-transformation/]

Projekt OpenJustice vs. Privacy [https://www.bfh.ch/de/forschung/forschungsbereiche/digital-sustainability-lab/forschungsprojekt-open-justice-privacy/]

Arbeitsgruppe Applied Machine Intelligence [https://www.bfh.ch/de/forschung/forschungsbereiche/applied-machine-intelligence/]

Projekt Erkennung & Abschwächung von Vorurteilen in auf dem Arbeitsmarkt eingesetzter KI [https://www.bfh.ch/de/forschung/forschungsprojekte/2022-025-172-803/]

TRANSFORM 2023: Künstliche Intelligenz im öffentlichen Sektor

Das Thema der TRANSFORM 2023 ist «künstliche Intelligenz im öffentlichen Sektor». Machine Learning, Chatbots, Natural Language Processing und weitere Methoden, die auf künstlicher Intelligenz (KI) basieren, bieten viele Chancen, aber auch gewisse Risiken für die Behörden. Wo stehen wir heute mit der Anwendung von KI? Was für Erfahrungen macht die Verwaltung mit KI? Wo gibt es Potenziale für den Einsatz von KI in administrativen Arbeitsprozessen? Was sind mögliche Risiken und Chancen? Diese Fragen werden zusammen mit Referent*innen aus der Wissenschaft, der Verwaltung und weiteren Organisationen diskutiert. Keynotes halten Paulina Grnarova (DeepJudge) und Bertrand Loison (BFS) gefolgt von einem Reality Check von Matthias Mazenauer (Kanton ZH) und Kommentar von Marc Steiner (IPST). Darauf folgen eine Vielzahl unterschiedlicher Beiträge mit Praxiserfahrungen, rechtlichen Überlegungen und kritischer Hinterfragung.

Weitere Informationen und die Anmeldung finden Sie hier. [https://www.bfh.ch/de/aktuell/fachveranstaltungen/transform-2023/]

Dieser Podcast wird produziert mit freundlicher Unterstützung von: Audioflair Bern [https://audioflair.ch/] und Podcastschmiede Winterthur [https://www.podcastschmiede.ch/] .

AUTOR/AUTORIN: ANNE-CAREEN STOLTZE

Anne-Careen Stoltze ist Redaktorin des Wissenschaftsmagazins SocietyByte und Host des Podcasts "Let's Talk Business". Sie arbeitet in der Kommunikation der BFH Wirtschaft, sie ist Journalistin und Geologin.

Posts von Anne-Careen Stoltze | Website

PDF erstellen

Ähnliche Beiträge

Mit welchen Techniken sich die Arbeitsweise von KI entschlüsseln lässt

0

KOMMENTARE