Fabian Haak – Untersuchung von geschlechtsspezifischen Verzerrungen in automatisierten Systemen zur Bewertung von Online-Suchen[1]

Fabian Haak

As part of his PhD, Fabian Haak researches bias in queries, query suggestions, and search results in web search engines.

At TH Köln, his work focuses in particular on information retrieval, natural language processing, and bias in online information systems.

Beschreibung des Fellowvorhabens

- In der Information Retrieval (IR) Forschung bleibt die Quantifizierung, wie problematisch oder voreingenommen ein Suchergebnis ist, eine große Herausforderung. Bias existiert und muss erforscht werden, doch die Quantifizierung der Schwere (Severity) ist hochgradig subjektiv und menschliche Annotation ist teuer.
- Um dieses Skalierungsproblem zu lösen, nutzen Forschende zunehmend Large Language Models (LLMs) als Annotatoren, um Bias automatisch zu bewerten. Dabei stellten wir fest, dass die Modelle selbst einen Gender Bias in ihrer Wahrnehmung von „Bias“ (schädlich/ toxisch/ diffamierend) aufzuweisen scheinen.
- Wenn LLMs als Evaluatoren in der Informationswissenschaft eingesetzt werden sollen, müssen wir diese internen Inkonsistenzen untersuchen und verstehen. Dies sowie die Untersuchung, welche weiteren geschlechtsspezifischen Kontextfaktoren Entscheidungen bei der Bias-Annotation beeinflussen, stellen eine signifikante Forschungslücke dar, die in diesem Vorhaben adressiert werden soll.

Forschungsfragen:

1. Bewerten LLMs die Schwere eines verzerrten Suchvorschlags unterschiedlich, abhängig vom Geschlecht des Subjekts im Vorschlag (z. B. „[Weibliche Person] Beine“ vs. „[Männliche Person] Beine“)?

1. Beeinflusst die dem LLM zugewiesene „Persona“ das Urteil? Wenn wir das Modell anweisen, als ein bestimmtes Geschlecht zu agieren (z. B. „Du bist ein männlicher Annotator“), ändert sich dann die Sensibilität für Gender Bias? Spiegeln diese Änderungen reale menschliche Diskrepanzen wieder oder induziert das Modell stereotype Halluzinationen?

1. Wenn das LLM die Aufgabe hat, das Schadenspotenzial für einen spezifischen User einzuschätzen, verändert sich dann das Urteil des Modells abhängig vom Geschlecht der Person?

Description of the fellow project

- In Information Retrieval (IR) research, quantifying how problematic or biased a search result is remains a major challenge. Bias exists and must be studied, yet quantifying its severity is highly subjective, and human annotation is expensive.
- To address this scaling problem, researchers increasingly use Large Language Models (LLMs) as annotators to automatically assess bias. In doing so, we found that the models themselves appear to exhibit a gender bias in their perception of “bias” (harmfulness/ toxic).
- If LLMs are to be used as evaluators in information science, we must examine and understand these internal inconsistencies. This, as well as investigating which additional gender-specific contextual factors influence decisions in bias annotation, represents a significant research gap that this project aims to address.

Research Questions:

1. Do LLMs evaluate the severity of a biased search suggestion differently depending on the gender of the subject in the suggestion (e.g., “[Female Person] legs” vs. “[Male Person] legs”)?

1. Does the “persona” assigned to the LLM influence the judgment? If we instruct the model to act as a specific gender (e.g., “You are a male annotator”), does its sensitivity to gender bias change? Do these changes reflect real human discrepancies, or does the model induce stereotypical hallucinations?

1. If the LLM is tasked with assessing the potential harm for a specific user, does the model’s judgment vary depending on the gender of the person?

[1]Study of Gender Bias in Automated Systems for Evaluating Online Searches

Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

Technology Arts Sciences TH Köln