idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Technische Universität Berlin
Stefanie Terp

09/11/2023
http://idw-online.de/en/news820371

Transfer of Science or Research
Information technology
transregional, national

Gutachter*innen auf Bestellung?

Forscher*innen an der TU Berlin tricksen KI-gestützte Textanalyse bei der Auswahl von wissenschaftlichen Gutachter*innen aus

Wissenschaftler*innen des Berlin Institute for the Foundations of Learning and Data (BIFOLD) an der TU Berlin haben signifikante Schwachstellen in einigen KI-gestützten Textanalyse-Systemen gefunden: Bereits mit kleinen Änderungen von Wörtern, Sätzen und Referenzen, die für Menschen keinerlei Unterschied machen, konnten sie die Textanalyse beeinflussen. Als Anwendungsbeispiel diente ihnen der Peer-Review-Prozess im Zusammenhang mit wissenschaftlichen Publikationen: In ihrer jüngsten Veröffentlichung demonstrierten sie, wie man durch kleine Adaptionen der eingereichten Publikation „missliebige“ oder „kritische“ Gutachter*innen umgehen kann.

Algorithmen des maschinellen Lernens sind nicht erst seit ChatGPT zu einem wichtigen Werkzeug bei der Analyse von Texten geworden. Insbesondere die automatische Einordnung von Texten in bestimmte Oberthemen mit Hilfe von sogenannten Topic-Models ist weitverbreitet, um schnell viele Texte zu sortieren. Nicht zuletzt nutzen auch wissenschaftliche Fachzeitschriften oder Kongresse diese Lernmodelle, um eingereichte wissenschaftliche Artikel automatisch auf qualifizierte Gutachter*innen zu verteilen. Diese entscheiden letztlich darüber, ob der Artikel veröffentlicht wird.

Angriff auf Topic-Models zur Textanalyse

Um die Sicherheit dieser KI-gestützten Textanalyse zu untersuchen, hat das Team um den BIFOLD-Wissenschaftler Prof. Dr. Konrad Rieck einen Angriff auf Topic-Models untersucht. Dazu haben sie eine neue Methode entwickelt, mit der sie minimale Manipulationen an Texten durchführen, deren Effekt auf die Textanalyse messen und die Texte schrittweise immer wieder neu anpassen. Schlussendlich ist es den Forscher*innen gelungen, das von den Topic-Models erkannte Oberthema eines Textes gezielt zu verändern – ohne wesentliche inhaltliche Veränderungen vorzunehmen. Während Menschen kaum einen Unterschied in den Texten bemerken, führt das gezielte Weglassen und Hinzufügen einzelner Begriffe die Algorithmen der Topic-Models schnell in die Irre.

Betrügen beim Peer-Review-Prozess

Was mit solchen Manipulationen erreicht werden kann, hat das Team an Beispielen aus der eigenen Zunft demonstriert: Dazu simulierten sie den sogenannten Peer-Review-Prozess von zwei großen wissenschaftlichen Konferenzen, die ein entsprechendes Topic-Model für die automatisierte Zuweisung von Oberthemen und dazu passenden Gutachter*innen nutzen. Im Rahmen des Peer-Reviews wird die Qualität einer wissenschaftlichen Arbeit anonymisiert von Wissenschaftler*innen aus dem gleichen Forschungsfeld bewertet. Aufgrund dieser Begutachtung entscheidet sich, ob die Arbeit in einem wissenschaftlichen Journal oder für die Vorstellung auf einem Kongress akzeptiert wird.

Nun kann es aber sein, dass Forscher*innen gerade nicht wollen, dass Kolleg*innen aus dem gleichen Forschungsfeld ihre Arbeit begutachten. Zum Beispiel weil sie wissen, dass ihre Arbeit fehlerhaft ist und dem kritischen Blick einer Expert*in auf dem Gebiet nicht standhält. Oder weil das Forschungsfeld so klein ist, dass es sich bei den Gutachter*innen nur um direkte Konkurrent*innen im Wettlauf um wissenschaftliche Reputation oder Forschungsfördermittel handeln kann. Könnten Wissenschaftler*innen die KI-gestützten Textanalyse austricksen und die Auswahl manipulieren?

Minimale Änderungen von Wörtern, Sätzen und Referenzen reichen aus

In ihren Experimenten manipulierte das Team von Konrad Rieck wissenschaftliche Arbeiten durch minimale Änderungen von Wörtern, Sätzen und Referenzen und sorgten so dafür, dass die Topic-Models die Arbeiten gezielt falschen Themenbereichen zuordneten. Im Ergebnis bekamen diese manipulierten Arbeiten in den simulierten Prozessen automatisch andere Gutachter*innen zugeordnet als die Original-Arbeiten. So wird es potenziell möglich, „unliebsame“ Gutachtende von vorneherein auszuschließen und stattdessen eher wohlgesonnene Gutachter*innen einzubeziehen. „Es ist natürlich nicht im Sinne der Wissenschaft, wenn manipulierte Arbeiten ihre eigenen Gutachter*innen auswählen“, meint Konrad Rieck.

Algorithmen nehmen Dinge anders wahr als wir

Die Arbeit der Wissenschaftler*innen wurde auf dem 32. USENIX Security Symposium in den USA vorgestellt. Parallel dazu informierten die Forscher*innen die Betreiber*innen von automatisierten Systemen zur Begutachtung von Konferenzbeiträgen über ihre Ergebnisse. Das grundsätzliche Problem von automatisch verfälschten Texten lässt sich aktuell allerdings noch nicht lösen. „So wie Algorithmen Bilder anders wahrnehmen als Menschen, so nehmen sie auch Texte anders wahr als wir. Beides kann von potenziellen Angreifer*innen ausgenutzt werden. Hier stehen wir mit unserer Forschung sicherlich erst am Anfang und haben noch viel Arbeit vor uns", erklärt Konrad Rieck.

Publikation:
Thorsten Eisenhofer, Erwin Quiring, Jonas Möller, Doreen Riepel, Thorsten Holz, Konrad Rieck: “No more Reviewer #2: Subverting Automatic Paper-Reviewer Assignment using Adversarial Learning”, Proceeding of the 32nd USENIX Security Symposium, 2023.

Weitere Informationen erteilt Ihnen gern:
Prof. Dr. Konrad Rieck
TU Berlin/BIFOLD
Chair of Machine Learning and Security
E-Mail: rieck@tu-berlin.de