idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Bauhaus-Universität Weimar
Claudia Weinreich
04/24/2025
http://idw-online.de/en/news851105

Research projects
Information technology
transregional, national

Digitale Wasserzeichen helfen dabei, KI-generierte Texte zu identifizieren

Ob Chatbots, Sprachassistenten oder automatische Texterstellung – die Künstliche Intelligenz schreibt an vielen Stellen mit. Wie aber lässt sich nachweisen, welchen Ursprung ein Text hat – ob er von einem Menschen stammt oder durch ein Computerprogramm generiert wurde? Können unsichtbare Wasserzeichen helfen, Autorschaftsinformation unterzubringen? Antworten auf diese Fragen sucht das neue Forschungsprojekt »Lantmark«, das Anfang April 2025 an der Bauhaus-Universität Weimar gestartet ist und vom BMBF mit ca. 1 Mio. Euro gefördert wird.

Gemeinsam mit dem Fraunhofer-Institut für Digitale Medientechnologie IDMT in Ilmenau und der Artefact Germany aus Hamburg erforschen Wissenschaftler*innen der Fakultät Medien digitale Wasserzeichen für Texte und gesprochene Sprache. Ziel ist es, automatisch generierte Inhalte künftig als solche erkennbar zu machen und so Transparenz und Vertrauenswürdigkeit in digitalen Kommunikationsräumen zu stärken. Nutzende sollen überprüfen können, welches Programm einen Text erstellt hat oder ob ein Text von einem Menschen verfasst oder gesprochen wurde.

Im Fokus steht das sogenannte Text-Watermarking, ein Verfahren, mit dem sich Markierungen als Teil eines Textes versteckt unterbringen lassen, um beispielsweise die Herkunft des Textes nachzuvollziehen oder Textänderungen zu erkennen. Ein weiteres Ziel des Projektes ist es, große Sprachmodelle, sogenannte Large Language Models (LLMs) so zu modifizieren, dass sie »gebrandet« sind. Sie tragen also eine Art digitale Signatur, mit deren Hilfe sich gebrandete Sprachmodelle von nicht-gebrandeten unterscheiden lassen. Mit den so zertifizierten Sprachmodellen können Unternehmen oder Personen nicht-autorisierte Meldungen über sich aufdecken, weil diese eben nicht die korrekten Wasserzeichen enthalten.

Bisher einzigartig ist die Kombination aus Text- und Audioanalyse in Lantmark: Die Antragsteller schlagen ein neuartiges Verfahren zur bimodalen Authentifizierung vor, bei der geschriebene und gesprochene Inhalte gemeinsam überprüft werden. Hiermit könnten beispielsweise vertonte Aussagen festgestellt werden, die einer Person mittels Sprachsynthese »in den Mund gelegt wurden«.

Mit dem Projekt verfolgen die Forschenden u. a. das Ziel, konkret anwendbare Verfahren zu entwickeln, die sich in der Praxis bewähren – etwa bei der Bekämpfung von Manipulation in digitalen Räumen. »Große Sprachmodelle wie ChatGPT werden in vielen professionellen und privaten Anwendungen eingesetzt. Digitale Wasserzeichen und ›zertifizierte‹ Sprachmodelle können künftig eine wichtige Rolle spielen, um Nachweise und Garantien für den Ursprung von Information zu geben«, erklärt Prof. Dr. Benno Stein, Leiter des Projekts und Inhaber der Professur für Intelligente Informationssysteme an der Bauhaus-Universität Weimar.

Das Verbundvorhaben »Lantmark – Potential und Grenzen von Large Language Models für Anwendungen im Watermarking von Text und gesprochener Sprache« wird vom Bundesministerium für Bildung und Forschung im Rahmen der Förderlinie »Sichere Zukunftstechnologien in einer hypervernetzten Welt: Künstliche Intelligenz« mit rund 1,07 Millionen Euro gefördert.

Weitere Informationen zum Forschungsprojekt gibt es unter:
https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/lantmark

contact for scientific information:
Für Rückfragen steht Ihnen Tina Meinhardt, Referentin für Forschungskommunikation & Öffentlichkeitsarbeit an der Fakultät Medien, telefonisch unter +49 (0) 36 43 / 58 37 65 oder per E-Mail an tina.meinhardt@uni-weimar.de zur Verfügung.

URL for press release: https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/lantmark