Kai Spriestersbach

Angst vor Betrug oder dem nächsten Google-Update:

Lassen sich KI-Texte erkennen?

Der Einsatz von künstlicher Intelligenz (KI) hat in den letzten Monaten eine bemerkenswerte Entwicklung durchlaufen und beeinflusst verschiedene Branchen, darunter auch den Bereich des Schreibens. Spätestens seit zahlreiche Domains den Index mit KI-generierten Artikeln fluten, stellt sich für SEOs die Frage, ob es möglich ist, KI-generierte Texte zuverlässig zu erkennen. Auch wenn Google in deren Kommunikation vollkommen neue Töne in Sachen KI-generierte Inhalte anschlägt und den Fokus auf den Mehrwert und nicht die Art der Erstellung legt, sollten wir uns fragen, ob wir KI-Texte von menschlichen unterscheiden können. Schließlich wollen wir uns keinen GPT-generierten Text als aufwendig händisch verfasst andrehen lassen. Und wie sollen wir mit einem hybriden Text verfahren, in dem menschliche und KI-generierte Inhalte kombiniert wurden? Um fundierte Entscheidungen zu treffen und nicht auf falsche Versprechen von KI-Detektoren hereinzufallen, müssen wir uns mit den Methoden und

DER AUTOR

Kai Spriestersbach ist erfolgreicher Unternehmer, Tech-Journalist, SEO-Veteran und KI-Experte mit einem Masterabschluss in Webwissenschaften. Sein Sachbuch „Richtig Texten mit KI: ChatGPT, GPT-4, GPT-3 & Co." erschien im April 2023 im mvg Verlag.

Alle Fotos: app.leonardo.ai

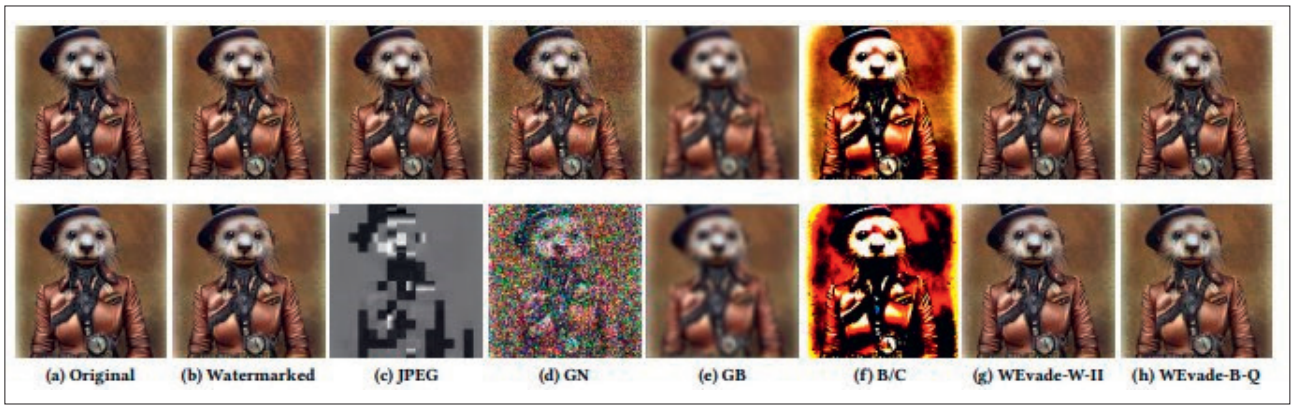

Abb. 1: Die Wasserzeichenmethode ist HiDDeN (GN: gaußsches Rauschen. GB: gaußsche Unschärfe. B/C: Helligkeit/Kontrast). Der Encoder/Decoder wird durch Standardtraining (erste Reihe) oder kontradiktorisches Training (zweite Reihe) trainiert *(Quelle: Zhengyuan Jiang et al., einfach.st/arc44).*

Keine Frage: Große Sprachmodelle (LLMs) wie GPT-3, GPT-4 und Co. haben beeindruckende Fortschritte gemacht und können eine Vielzahl von Aufgaben sehr gut bewältigen, beispielsweise das Vervollständigen von Dokumenten und das Beantworten von Fragen. Es gibt jedoch vermehrt Bedenken hinsichtlich des unregulierten Einsatzes dieser Modelle, da sie für unerwünschte Zwecke wie Plagiarismus, Erzeugung von Fake News und Spamming missbraucht werden können. Daher wäre es wichtig, künstlich generierten Text zuverlässig erkennen zu können. Denn die jüngsten Fortschritte in der KI-Technologie haben dazu geführt, dass immer mehr Texte im Internet von generativer KI geschrieben werden. Die Frage, ob man solche Texte automatisch erkennen kann, steht aber nicht nur für Betreibenden von Suchmaschinen im Raum, auch als Auftraggebende für Texte möchte man sicherstellen, dass die Texter:innen nicht einfach nur ein Knöpfchen gedrückt haben. Bei schriftlichen Leistungsnachweisen an Schulen und Hochschulen stehen Lehrpersonen und Dozent:innen vor gigantischen Herausforderungen und nicht zuletzt für Verlage wäre eine zuverlässige Erkennung von KI-generierten Texten ein Segen, schließlich hat dies auch Auswirkungen auf die Urheberrechte der veröffentlichten Werke.

Aus der Ecke der KI-Ethik werden bereits Forderungen laut, Organisationen, die Grundlagenmodelle für die öffentliche Nutzung entwickeln, sollten verpflichtet werden, einen zuverlässigen Erkennungsmechanismus für die von dem Modell generierten Inhalte bereitzustellen, bevor es öffentlich freigegeben wird. Dieser Mechanismus sollte zudem kostenlos öffentlich zugänglich sein und den Nutzenden erlauben, zu prüfen, ob ein bestimmter Inhalt (ganz oder teilweise) von dem Modell erstellt wurde. In ihrem Open-Access-Paper (*einfach.st/spring53*) argumentieren die Autor:innen zwar, dass diese Anforderung technisch umsetzbar sei und somit dazu beitragen würde, bestimmte Risiken neuer AI-Modelle zu reduzieren, doch wie sich im Laufe dieses Artikels zeigt, ist diese Betrachtung relativ naiv.

Unterschiedliche Ansätze für die Erkennung von KI-generierten Texten

Ein unter Lehrenden in den USA beliebtes Tool auf dem Markt, GPTZero, misst im Wesentlichen die Zufälligkeit eines bestimmten Textes und vergleicht diese mit dem Zufallsmaß bekannter menschlicher Texte. Laut der FAQ-Seite von GPTZero ist dieses Tool in der Lage, 99 % der Zeit von Menschen verfassten Text und 85 % der Zeit von KI generierten Text zu erkennen.

In der Praxis zeigt sich jedoch schnell, dass diese Zahlen lediglich in Best-Case-Szenarien erreicht werden, bei denen relativ lange und unveränderte Texte geprüft werden, die auf Englisch und im Stil von schulischen Essays geschrieben wurden. In den meisten anderen Einsatzszenarien werden diese Werte nach meinen Beobachtungen nicht einmal ansatzweise erreicht.

Auch der anfänglich angebotene Detektor für KI-Texte von OpenAI hat sich als erschreckend ungenau herausgestellt. Hierbei handelte es sich um ein System, das per maschinellem Lernen mit menschlichen und KI-generierten Texten trainiert wurde und dadurch lernen sollte, diese zu unterscheiden.

OpenAI gab jedoch zu, dass das Tool nur 26 % der von KI geschriebenen Texte richtig identifiziert (True Positives). Außerdem werden 9 % der von Menschen geschriebenen Texte fälschlicherweise als KI-Texte eingestuft (falsch positive Ergebnisse). Die Treffsicherheit des sogenannten „AI Classifiers" war am Ende so schlecht, dass OpenAI diesen am 20. Juli 2023 offline gestellt hat (*platform.openai.com/ai-text-classifier*).

Man arbeite nun daran, das Feedback einzubeziehen, und erforsche „effektivere Verfahren zur Herkunftsbestimmung von Texten". Offenbar konzentriert man sich nun lieber dar-

auf, zumindest generierte Bilder mit einem Wasserzeichen zu versehen und diese verlässlich wiederzuerkennen (Abbildung 1). Ich sehe dies ebenfalls skeptisch, denn eine aktuelle Studie von Forschenden der Duke University (*einfach.st/arc43*) zeigt, dass die derzeitigen auf Wasserzeichen basierenden Erkennungsmethoden für KI-generierte Inhalte ebenfalls nicht ausreichend sind, und betont die dringende Notwendigkeit neuer Methoden.

Speziell in Bezug auf KI-generierte Texte zeigt sich immer deutlicher, dass eine verlässliche Erkennung offenbar nicht möglich zu sein scheint.

KI-Text-Erkennung: eine Sisyphusarbeit

Mehrere aktuelle Studien legen jedoch nahe, dass diese Erkennungsmethoden in realen Szenarien an ihre Grenzen stoßen. Insbesondere zeigen die Ergebnisse, dass einfache Textumformulierungen solche Erkennungstechniken wirkungslos machen können. Noch besorgniserregender ist die Erkenntnis, dass, je fortschrittlicher und menschenähnlicher die Modelle werden, desto herausfordernder deren Erkennung wird. Das lässt die Frage aufkommen, wie Unternehmen und Webmaster in der digitalen Landschaft künstlich generierte Inhalte sicher identifizieren können. Die Debatte um den verantwortungsbewussten Umgang mit KI-generierten Texten dürfte somit weiter an Fahrt gewinnen.

> „Kann von KI erzeugter Text zuverlässig erkannt werden? Nein, sagen Forschende."

So fanden Forschende der Universität von Maryland beispielsweise jüngst heraus, dass selbst die besten Detektoren keine Sicherheit im Umgang bieten können. Die Wissenschaftler:innen bezweifeln in ihrem Pre-Print (*einfach.st/arc45*) sogar, dass es jemals möglich sein wird, von KI generierten Text zuverlässig zu erkennen. Professor Soheil Feizi und vier Informatik-Doktoranden haben in ihrer Studie die Frage gestellt: „Kann von KI erzeugter Text zuverlässig erkannt werden?" – und ihre Antwort lautet leider Nein. Selbst einfache Umformulierungen oder geringfügige Änderungen an KI-generierten Texten konnten die Detektoren bereits täuschen. Die Auswertungen der Arbeitsgruppe haben gezeigt, dass selbst die derzeit modernsten Erkennungsmethoden in der Praxis kaum in der Lage sind, LLM-generierten Text zuverlässig zu identifizieren. Die Studie zeigt klar, dass selbst die besten Erkennungsmethoden ihre Grenzen haben, da sie durch einfaches Paraphrasieren leicht umgangen werden können. Bei sehr fortgeschrittenen Modellen ist selbst der beste Detektor kaum besser als ein zufälliger Klassifikator.

> „Würfeln wäre genauso zuverlässig bei der Ernennung, ob ein Text durch eine KI erzeugt wurde."

Das Paper zeigt auch, dass selbst durch Wasserzeichen geschützte Sprachmodelle anfällig für Angriffe sind, bei denen Menschen verborgene Signaturen des Modells erkennen und in Text einfügen können, die von Menschen verfasst wurden. Ein weiteres Problem, das speziell für nicht englische Texte gilt: Die meisten GPT-Detektoren wurden mit englischen Texten trainiert und für den US-Markt entwickelt und sind damit per se nicht für Texte anderer Sprachen geeignet. Wie eine aktuelle Stanford-Studie (Pre-Print: *einfach.st/arc46*) zeigt, sind diese Tools daher gegenüber nicht englischen Muttersprachler:innen voreingenommen. Die Stanford-Forschenden untersuchten sieben weitverbreitete GPT-Detektoren: Originality.ai, Quil.org, Sapling, OpenAI (inzwischen eingestellt), Crossplag, GPTZero und ZeroGPT. Dabei wurden Texte von Nichtmuttersprachler:innen häufig fälschlicherweise als KI-generiert klassifiziert.

In einem weiteren Pre-Print untersuchte eine Arbeitsgruppe um die emeritierte Professorin der HTW Berlin, Debora Weber-Wulff, zwölf kostenlose KI-Prüfwerkzeuge sowie zwei kostenpflichtige KI-Erkennungstools (*einfach.st/arc47*). Die besagte Arbeitsgruppe „Technology & Academic Integrity" des European Network for Academic Integrity widmet sich speziell der Bewertung von Programmen zur KI-Text-Erkennung und hat bislang 14 öffentlich verfügbare Tools auf ihre Fähigkeit geprüft, KI-generierte Texte zu identifizieren. Diese umfassten Check For AI, Compilatio, Content at Scale, Crossplag, DetectGPT, Go Winston, GPT Zero, GPT-2 Output Detector Demo, OpenAI Text Classifier, PlagiarismCheck, TurnItIn, Writeful, GPT Detector, Writer und Zero GPT. Obwohl Tools wie Copyleaks oder undetectable.ai in dieser Untersuchung nicht berücksichtigt wurden, zogen die Forschenden den Schluss, dass alle vorhandenen Erkennungswerkzeuge weder genau noch zuverlässig sind.

> „Alle diese Systeme neigen aus ihrer statistischen Natur heraus dazu, Texte fälschlicherweise als menschlich verfasst zu klassifizieren, anstatt KI-generierte Inhalte zu erkennen."

Eine weitere Studie (*einfach.st/jour52*), die im Journal of Applied Learning & Teaching erschienen ist, kam zu dem gleichen Schluss, dass keines der getesteten Tools vollständig in der Lage ist, KI-generierte Inhalte in verschiedenen Kontexten genau zu erkennen, was negative Auswirkungen auf das Problem des KI-generierten Plagiats in wissenschaftlichen Arbeiten haben könnte. In dieser Studie wurde die Genauigkeit von fünf KI-Inhaltserkennungstools getestet, um AI-generierte Inhalte in den Antworten von ChatGPT, YouChat und Chatsonic zu erkennen. Die Antworten dieser Chatbots wurden mit englischen Prompts im Feld der angewandten Anglistik erzeugt. Die generierten Antworten wurden zusätzlich mittels Google Translate in verschiedene Sprachen übersetzt und in verschiedene Tools eingegeben, um die KI-generierten Inhalte zu erkennen. Bei den auf Englisch generierten Antworten von ChatGPT, YouChat und Chatsonic schnitt Copyleaks AI Content Detector am besten ab, gefolgt von OpenAIs Text Classifier. Bei den in fünf Sprachen übersetzten Antworten von ChatGPT identifizierte GPTZero alle als von Menschen erstellt. Bei den auf Deutsch, Französisch und Spanisch übersetzten Antworten erkannte Copyleaks jedoch einige Texte korrekt als von KI generiert. Insgesamt zeigt diese Studie jedoch, dass keines der derzeitig erhältlichen Tools zuverlässig funktioniert.

Eine weitere interessante Arbeit einer Gruppe der University of Maryland *(einfach.st/arc48)* befasst sich ebenfalls mit dem Problem, Texte, die von großen Sprachmodellen (LLMs) erzeugt wurden, von menschlich verfassten Texten zu unterscheiden. Mit Bezug auf die Informationstheorie argumentieren die Forschenden darin, dass, wenn maschinell erzeugter Text menschenähnliche Qualität annimmt, die benötigte Stichprobengröße zur Erkennung immer weiter steigt, was die Praktikabilität enorm einschränkt.

> „Gerade in Umgebungen, in denen es wichtig ist, zwischen menschlichen und KI-generierten Texten zu unterscheiden, zum Beispiel in der Wissenschaft oder im Journalismus, könnte eine falsche Identifikation schwerwiegende Konsequenzen für die Autor:innen haben."

Doch das stoppt die Anbieter von KI-Detektoren keinesfalls. Unternehmen wie Winston AI, Content at Scale und Turnitin machen sich im schulischen Umfeld einen Namen mit ihrer angeblichen Fähigkeit, KI-Beteiligung in Arbeiten von Schüler:innen zu erkennen. Diese Unternehmen bieten Abo-Dienste an, bei denen Lehrkräfte die Arbeiten ihrer Schüler:innen über ein Web-Dashboard überprüfen können und eine Wahrscheinlichkeitsbewertung erhalten, die angibt, wie „menschlich" oder „KI" der Text ist. Dass diese Tools nicht halten, was sie versprechen, wissen wahrscheinlich die wenigsten Nutzenden, die aus der Verzweiflung nach einer Lösung auf die vollmundigen Werbeversprechen der Firmen hereinfallen. Noch immer werben zahlreiche Unternehmen wie Copyleaks und undetectable.ai damit, KI-Texte zu erkennen. Im zweiten Fall bietet man sogar noch eine Lösung an, mit der Texte so umgeschrieben werden können, dass diese nicht mehr als KI-generiert erkannt werden können. Ein lukratives Geschäftsmodell, das auf beiden Seiten mitverdient. Doch mittels einfacher Tests zeigt sich sehr schnell, dass diese Tools keineswegs zuverlässig sind. Copyleaks lieferte in einem Kurztest derart viele „False Positives" (Texte, die als KI-generiert eingestuft werden, es in Wirklichkeit aber gar nicht sind), dass man niemandem empfehlen kann, sich darauf zu verlassen.

INFO

Obwohl es verlockend ist, sich auf KI-Tools zu verlassen, um KI-generierten Text zu erkennen, hat sich gezeigt, dass diese nicht praxistauglich sind. KI-Text-Detektoren wie GPTZero, ZeroGPT und der Text Classifier von OpenAI erkennen KI-generierte Texte nicht zuverlässig genug, da sie häufig falsch positive Ergebnisse liefern.

Die praktischen Auswirkungen dieser Ergebnisse sind beträchtlich. Eine falsche Identifizierung durch Detektoren kann schwerwiegende Folgen haben, die zu falschen Anschuldigungen und Reputationsschäden führen können, insbesondere in akademischen und journalistischen Kontexten. Diese Unsicherheit führt dazu, dass die Erkennungsmechanismen kaum zuverlässiger sind als ein reiner Zufallsgenerator.

Die vergebliche Suche nach einer Lösung

Selbst Marktführer OpenAI erklärte seine Bemühungen um seinen „AI Classifier" nach nur sechs Monaten für gescheitert und schaltete diesen Dienst am 20. Juli 2023 ab. Der KI-Klassifikator sei aufgrund seiner geringen Genauigkeit nicht mehr verfügbar (*openai.com/blog/new-ai-classifier-for-indicating-ai-written-text*). Man suche nun nach „effektiveren Verfahren" zur Herkunftsbestimmung von Texten. Gleichzeitig habe man sich verpflichtet, Mechanismen zu entwickeln und einzusetzen, die es den Nutzenden ermöglichen, zu erkennen, ob Audio- oder visuelle Inhalte von KI stammen. Die führenden KI-Unternehmen wie OpenAI, Alphabet (Google), Anthropic und Meta haben sich gegenüber dem Weißen Haus sogar verpflichtet, Maßnahmen wie die Kennzeichnung von KI-Inhalten mit Wasserzeichen umzusetzen, um die Technologie sicherer zu machen, wie Präsident Joe Biden jüngst bekannt gab.

> „Unser AI Classifier ist gescheitert!", OpenAI.

Diese Wasserzeichen sollen auf „technische Weise" in den Inhalt eingebettet werden und den Nutzenden ermöglichen, zumindest gefälschte

Bilder oder Audiodateien zu erkennen. Es ist derzeit jedoch noch vollkommen unklar, wie das Wasserzeichen bei der Weitergabe der Informationen sichtbar sein wird. Außerdem wurde bereits gezeigt, dass nicht ausgeschlossen werden kann, dass Menschen die Wasserzeichen entschlüsseln und in andere Bilder oder Tondokumente einfügen könnten, die nicht von KI erzeugt wurden.

Was bedeutet das für SEO-Texte und was sagt Googles Search Team dazu?

Die Rolle von künstlicher Intelligenz und maschinell erstellten Inhalten im Bereich der Suchmaschinenoptimierung hat sich weiterentwickelt, insbesondere das jüngste Update im September 2023 hat die Haltung von Google gegenüber von KI erstellten oder erweiterten Inhalten verdeutlicht:

In den aktualisierten Richtlinien der Suchmaschine wird die Rolle der KI bei der Erstellung von Inhalten subtil anerkannt, indem der Ausdruck „von Menschen geschrieben“ durch „für Menschen geschrieben“ ersetzt wurde. Der Hauptfokus liegt nun einzig und allein auf der Qualität der Inhalte, unabhängig davon, wie sie erstellt wurden.

In Googles eigenem „Leitfaden der Google Suche zu KI-generierten Inhalten“ (*einfach.st/godev743*) wird die Qualität der Inhalte als höchstes Gut herausgestellt. Die Algorithmen von Google belohnen Inhalte von hoher Qualität und Authentizität, unabhängig von ihrer Erstellungsart. Ein besonderes Augenmerk wird dabei auf die Eigenschaften E E A T gelegt, die für Erfahrung, Expertise, Autorität und Vertrauenswürdigkeit stehen. Von Anfang an lag der Fokus auf der Qualität der Inhalte, ungeachtet ihrer Entstehungsweise. Das Ziel sei die Bereitstellung zuverlässiger und hochwertiger Suchergebnisse und nicht die Erkennung von KI-generierten Texten.

INFO

„Für mich persönlich spielt es keine Rolle, ob ein Text von einer KI oder einem Menschen geschrieben wurde. Entweder es ist ein guter Text oder es ist kein guter Text. So sieht es auch aus Sicht der Suchmaschine aus. Entweder es ist Spam oder es ist kein Spam. Von Menschen geschriebener Spam ist genauso schlecht für die Qualität der Suchergebnisse wie von KI geschriebener Spam. Und ein richtig guter Artikel, der von der KI geschrieben wurde, ist im Endeffekt genauso gut, wie wenn ihn ein Mensch geschrieben hätte.“

Das ist für mich wenig überraschend, denn es gab im Anti-Spam-Team von Google bereits vor mehr als einem Jahrzehnt Bedenken angesichts einer Flut an Inhalten, die maschinell generiert wurden, doch eine radikale Verbannung nicht menschlich erstellter Inhalte wurde damals schon als unverhältnismäßig angesehen. Stattdessen konzentrierte man sich darauf, die Systeme so zu optimieren, dass sie herausragende Inhalte honorieren. Dieser Ansatz setzt sich bis heute fort, etwa durch spezielle Ranking-Systeme, die verlässliche Informationen bevorzugen, und durch das System für hilfreiche Inhalte. Letzteres wurde eingeführt, um Nutzenden Inhalte zu präsentieren, die primär ihrem Zweck dienen und nicht nur auf die Platzierung in den Suchergebnissen abzielen.

Automatisierte Inhalte und die Rolle der KI

In Bezug auf automatisch generierte Inhalte sind klare Richtlinien bei Google schon seit Langem etabliert: Automatisierung, einschließlich KI, zur Erzeugung von Inhalten mit dem Ziel der Manipulation von Suchergebnissen verstößt gegen die Spamrichtlinien. Googles Anti-Spam-Maßnahmen sind unabhängig von der Erzeugungsweise von Inhalten wirksam.

Trotzdem ist zu beachten, dass nicht jede Form der Automatisierung, einschließlich KI, zwangsläufig als Spam betrachtet werden muss! Automatisierung wird im Journalismus bereits seit Jahren eingesetzt, um nützliche Inhalte wie Sportergebnisse oder Wettervorhersagen zu erstellen. Wichtig sei es, KI verantwortungsvoll für die Content-Erstellung einzusetzen. Dabei sollen Content-Creator, die in der Google-Suche erfolgreich sein möchten, qualitativ hochwertige, nutzerzentrierte Originalinhalte mit den E E A T-Eigenschaften erstellen. Durch diese Vorgehensweise – unabhängig davon, ob KI-generierte Inhalte verwendet werden oder nicht – werden Googles Anforderungen erfüllt.

Fazit

Aktuelle Detektorsysteme sind nicht ausreichend zuverlässig und liefern oft falsch positive Ergebnisse. Es ist offensichtlich, dass eine verlässliche und einfache Lösung für das Erkennen von KI-generierten Texten derzeit nicht in Sicht ist. Die ethische und verantwortungsvolle Nutzung von solchen Texten sollte dennoch höchste Priorität haben. Ob ein Text von einer KI oder einem Menschen geschrieben wurde, wird für den Erfolg jedoch keine Rolle spielen. Qualität und Nutzen für die Lesenden stehen im Vordergrund und zum Glück sehen die Suchmaschinen dies ebenfalls so.

Die Entwicklung von KI-Texterkennungstools steht jedoch noch am Anfang und vielleicht wird es eines Tages jemandem gelingen, eine verlässliche Lösung zu entwickeln. In der Zwischenzeit ist es wichtig, die Grenzen dieser Technologie zu verstehen und kritisch zu hinterfragen, bevor man sich auf sie verlässt.