Wer hat‘s geschrieben? Mensch oder Maschine?

WIKIPEDIA

Digitaler Themenstammtisch
mit Salino01 und Doris Weßels

Was sind die Gefahren der KI?

Künstliche Intelligenz (KI) kann auch Gefahren darstellen, insbesondere in Bezug auf die Integrität und Zuverlässigkeit der Informationen. Einige der möglichen Gefahren sind:

- **Fehlerhafte Informationen**: KI-basierte Systeme können aufgrund von Fehlern oder Fehlinterpretationen falsche Informationen generieren oder verbreiten.
- **Bias und Diskriminierung**: KI-Systeme können auch dazu neigen, bestehende Vorurteile und Diskriminierung zu replizieren, insbesondere wenn sie auf unstrukturierte Daten trainiert werden, die diese Vorurteile reflektieren.
- **Manipulation von Informationen**: KI-Systeme können auch dazu genutzt werden, um Informationen zu manipulieren oder zu verfälschen.

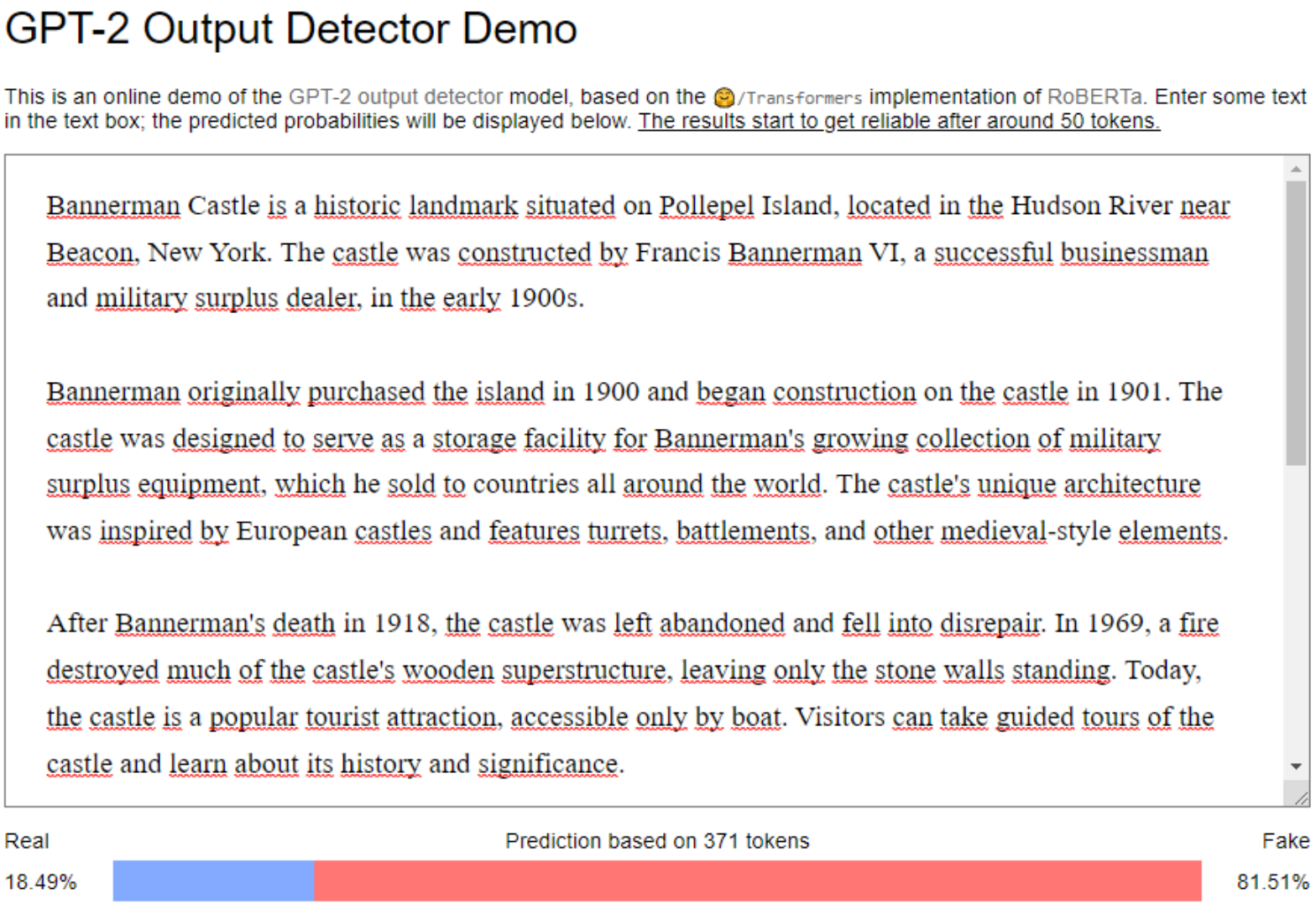
Wer hat Interesse KI-Texte zu identifizieren?

- Google (Suchmaschinen sollen reale Texte als Ergebnis liefern, Qualität statt Quantität)
- Wikipedia (Vermeidung von falschen Inhalten)
- Hochschule und Lehre (Erkennen von Plagiaten / unerlaubten Hilfsmitteln bei wissenschaftlichen Arbeiten)
- Erkennen von Fake-News
- …

Kann man KI-Texte automatisch erkennen?

- Es wurden verschiedene auf KI basierende Programme entwickelt, um herauszufinden, ob ein Text vom Computer erzeugt wurde.
- Häufig sind diese „noch“ auf englischsprachige Texte limitiert und für den Test benötigt man meist relativ lange Texte von mindestens 1000 Zeichen.
- Beispiel: GPT-2 Output Detector Demo (https://openai-openai-detector.hf.space/)

Test: GPT-2 Output Detector Demo



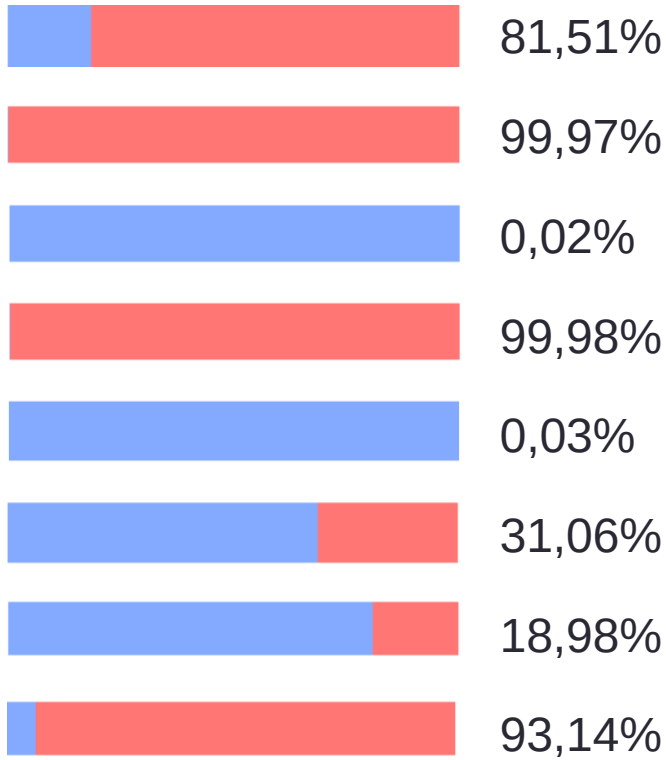
Zu prüfender Text wird einfach in den Editor kopiert und man erhält eine Bewertung mit Wahrscheinlichkeiten für Fake (KI-generiert).

GPT-2 Output Detector wurde ursprünglich für GPT-2 entwickelt, soll nach verschiedenen Berichten auch verlässliche Prüfungen bei GPT-3 erlauben.

Praktischer Test GPT-2 Output Detector Demo

Beispieltext: Text über ‚Bannerman Castle‘

- KI-Text in englisch erzeugt
- KI-Text in englisch erzeugt (enzyklopädisch)
  - Englischer KI-Text von KI in Deutsch übersetzt
  - Rückübersetzung durch deepl
- KI-Text in Deutsch erzeugt (enzyklopädisch)
  - Übersetzung in Englisch durch deepl
  - + „Translated with …“
  - Englisch doppelt eingefügt



Probleme bei der Erkennung

- Qualität der KI-Texte und der Detektoren entwickelt sich in hoher Geschwindigkeit weiter (vgl. Viren / Antivirenprogramme).
- Texte werden ggf. nach der Erstellung noch durch einen Mensch überarbeitet.
- Der Schreibstil kann angepasst werden (z.B. verwende einen Schreibstil, in dem…), so dass Detektoren getäuscht werden.

Kann man KI-Texte automatisch erkennen?

- Die Erkennungsrate von KI-Text ist nicht besonders zuverlässig und von der Art (Schreibstil) abhängig.
- Ergebnis hängt von der Länge des Textes ab (z.B. Wiederholungen).
- Bereits kleine Ergänzungen oder Änderungen haben einen großen Einfluss auf das Ergebnis.
- Es gibt auch zahlreiche Falschmeldungen (menschliche Texte werden als potentiell KI gewertet).

Kann man KI-Texte von Hand erkennen?

- Plagiatssuchen sind meist erfolglos, da die Sätze von der KI selbst ‚gewürfelt' werden (könnte sich ändern, wenn KI-Texte immer häufiger auf normalen Internetseiten erscheinen). Plagiate sind eher Hinweise auf von Menschen kopierte Texte!

- Es gibt einige typische Merkmale wie die Textgestaltung, die für eine manuelle Identifizierung herangezogen werden können.

Keine grammatikalischen Fehler

- Während Texte auf Basis von GPT-1 und GPT-2 noch deutliche sprachliche Mängel aufwiesen, kommt es bei GPT-3 äußerst selten vor, dass ein KI-Textgenerator grammatische Fehler produziert oder der Text Inkonsistenzen in der Grammatik aufweist. Sofern der Autor dem KI-Tool keinen fehlerhaften Text als Vorlage gibt, wird es sich auch im Deutschen an die Richtlinien für korrekte Grammatik und Syntax halten.

Wortwiederholungen und Redundanzen

- Textgeneratoren folgen Mustern, so dass KI-generierte Texte dazu neigen, dieselben Wörter zu wiederholen, insbesondere wenn diese in der Eingabeaufforderung enthalten sind.
- Menschen versuchen in der Regel, abwechslungsreicher zu schreiben und insbesondere häufig wiederkehrende Wörter zu vermeiden. Statt immer wieder „Die Burg...“ zu wiederholen, könnten z.B. passende Synonyme wie die Befestigungsanlage, das Gebäude, das Anwesen, der Turm, ... verwendet werden. Ersatzausdrücke wie „diese …“ oder „sie…“ sind bei KI selten.
- KI-Texte können dazu neigen, dieselben Ausdrücke oder Ideen zu wiederholen, ohne neue Informationen oder Perspektiven einzuführen.

Einheitliche Formatierung

- KI Textgeneratoren neigen dazu, kürzere Sätze zu erstellen und weniger Neben- und vor allem Schachtelsätzen zu verwenden.
- Die Wortwahl ist meist einfacher und Fachausdrücke oder zusammengesetzte Worte werden seltener verwendet.
- KI Text ist oft relativ einheitlich formatiert (gleiche Satz-/ Absatzlänge). Von Menschen geschriebene Texte bieten eine größere Satzzeichenvielfalt, z. B. Gedankenstriche, Semikola, Klammern etc. oder stärker variierende Absatzlängen. Füllwörter wie „auch“ oder „dann“ sind in KI-Texten seltener.

Sachliche Fehler

- Programme wie ChatGPT basieren auf NLP (Natural Language Processing) und wurden konstruiert, um möglichst natürlich wirkende Texte und plausibel erscheinende Texte zu erstellen. Dem Wahrheitsgehalt einer Aussage wird weniger Priorität zugemessen. Inhaltliche Fehler sind um so häufiger, je weniger Informationen zu einem Thema vorliegen.

- Häufig sind Zahlen falsch oder werden gerundet, was z.B. bei Jahreszahlen zum Problem wird. Teilweise sind Daten veraltet, da die Datenbasis, mit der die KI trainiert wurde, nicht aktuell ist.

Quellenprobleme

- Quellenangaben von KI-Bots sind derzeit oft falsch oder nicht vorhanden.
- Existieren die Quellen, so ist der Inhalt des (KI-)Textes an Hand der angegebenen Quelle zu verifizieren.
- Die Quellen sind hinsichtlich ihrer fachlichen Qualität zu bewerten.

Zusammenfassung Textgestaltung

- Es ist häufig schwierig Texte von Mensch und Maschine zu unterscheiden. Teilweise sind wünschenswerte Texteigenschaften, wie fehlende Rechtschreib- und Grammatikfehler, einfachere Wortwahl, einfachere Satzgestaltung gerade Eigenschaften von KI-Texten. Menschliche Texte wirken hingegen meist noch lebendiger / abwechslungsreicher.

Zusammenfassung Textinhalt

- KI-Generatoren sind darauf programmiert, plausibel erscheinende Texte zu schreiben, so dass rein logische Analysen des Textes häufig erfolglos sind.
- KI-Texte stellen normalerweise keine Plagiate dar.
- Inhaltlich können KI-Texte trotzdem grobe Fehler enthalten, die aber nur durch eigenes Wissen oder Quellenstudium gefunden werden. Andererseits gilt diese Aussage auch für schlecht recherchierte menschliche Texte.

- *»Das Problem ist, dass wenn man anfängt, nach typischen menschlichen Eigenschaften in Texten zu suchen, dass es wahrscheinlich die KI schneller als Menschen schafft, die menschlichen Eigenschaften nachzuahmen.«*

Wie kann man KI-Texte erkennen?

Es gibt einige Ansätze, um KI-Texte zu erkennen:

- Unnatürlicher Schreibstil: KI-Systeme haben oft Schwierigkeiten, einen natürlichen Schreibstil zu imitieren, und können daher Texte produzieren, die unnatürlich oder stilistisch auffällig sind.
- Wiederholte Muster: KI-Systeme können auch dazu neigen, dieselben Muster und Phrasen wiederholt zu verwenden, insbesondere wenn sie auf bestimmte Datensätze trainiert wurden.
- Fehlende Kontextualisierung: KI-Systeme haben oft Schwierigkeiten, Kontextualisierungen und Nuancen in Texten zu erfassen, was zu unzureichenden oder fehlerhaften Inhalten führen kann.
- Tools zur Textanalyse: Es gibt auch einige Tools, die speziell entwickelt wurden, um KI-Texte zu erkennen. Diese Tools nutzen oft maschinelles Lernen und künstliche Intelligenz, um Muster in Texten zu erkennen, die auf KI-Generierung hinweisen.

Quellen

- *ChatGPT: Kostenloses Tool soll KI-generierte Texte erkennen helfen*. In: *Der Spiegel*. 1. Februar 2023, ISSN 2195-1349 (spiegel.de).
- marketing: *KI Texte erkennen - Diese 12 Tools habe ich getestet.* 5. Februar 2023.
- Roman: *KI Texte erkennen 2023: 6 Besten KI Text Detektoren, Tools, Prüfer.* 13. Januar 2023.
- Annika Mittelbach: *ChatGPT & Co.: So erkennen Sie von einer KI geschriebene Texte.* Chip.de, 26. Januar 2023

Wer hat‘s geschrieben: Mensch oder Maschine?