

Forensik-Projekt 2

KI-Detektoren im Test – Zero Tolerance Policy für ChatGPT mit *zerogpt.com*

eingereicht von:

Julia Winkler

Dozentin:

Studiengang IT-Forensik (PO 2017)

Prof. Antje Raab-Düsterhöft

Neudenu, 06.06.2023

Inhalt

1	Einleitung.....	4
1.1	Terminologische Hinweise	5
1.2	Hinweis zu Versionen und Quellen	7
2	ChatGPT	8
2.1	Grundlagen der GPT-Technologie.....	8
2.1.1	Generativ	9
2.1.2	Pre-training	9
2.1.3	Transformer	11
2.2	Funktionsweise und Einsatzgebiete	12
2.3	Daten und Zahlen	15
2.4	Kritik an ChatGPT	16
3	Textgrundlage	19
3.1	Variantenerzeugung.....	19
3.2	Variante 1.....	20
3.3	Variante 2.....	21
3.4	Variante 3.....	23
3.5	Vergleich der Varianten	24
4	Detektoren für synthetische Texte	26
4.1	Plagiate	26
4.1.1	plagiarismdetector	27
4.1.2	scribbr	28
4.1.3	duplichecker.....	29
4.2	Synthetische Texte	30
4.2.1	GLTR (Glitter)	33
4.2.2	OpenAI-Detector.....	36
4.2.3	GPT-Zero.....	38
4.2.4	ZeroGPT	38
4.2.5	Copyleaks	40
5	Fazit und Ausblick	42
6	Literaturverzeichnis	45
7	Bilderverzeichnis	48
8	Tabellenverzeichnis.....	50
9	Anhang.....	51
10	Verzeichnis der Abkürzungen	56
11	Selbstständigkeitserklärung	57

1 Einleitung

Die vorliegende Arbeit beschäftigt sich mit einem Thema, das aktuell diverse Bereiche des täglichen Lebens und große Teile des akademischen Betriebes tangiert: dem Einsatz des Chatbots ChatGPT. Konkret gilt es herauszufinden, inwieweit die künstliche Intelligenz (KI) von gängigen KI-Detektoren wirklich erkannt wird. Die verwandte Frage danach, ob KI-generierte Texte Plagiate sind oder nicht, soll im Rahmen dieser Ausführungen nicht tiefergehend behandelt werden, da zu ihrer Beantwortung auch juristische und z.T. ethische Erwägungen angestellt werden müssten, die den Rahmen dieser Arbeit sprengen würden. Ebenfalls keine Beachtung finden wird die Frage danach, wie sich die schulische und akademische Prüfungslandschaft und die Kolloquien zum wissenschaftlichen Arbeiten und Schreiben durch den Einsatz von ChatGPT verändern werden.¹

Im Folgenden sollen zunächst die wichtigsten Begriffe geklärt werden, da die Terminologie in diesem Bereich z.T. nicht einheitlich verwendet wird. Außerdem ist zu beobachten, dass aufgrund der Schnelligkeit der Materie viele Beiträge lediglich als Preprints oder Blog-Beiträge u.Ä. vorliegen, sodass dort oftmals von einem intuitiven Verständnis der Begriffe ausgegangen wird. Ebenfalls soll eine detailliertere Vorstellung und Einordnung von ChatGPT erfolgen (Kap. 2).

Für die anschließenden Analysen und Tests werden drei Varianten eines Textes generiert, die zunächst statistisch und linguistisch ausgewertet werden (Kap. 3), um dann als Basis für die Auswertungen in Kap. 4 zu dienen. Dort werden die Varianten zunächst mit drei Plagiatsscannern untersucht (Kap. 4.1), um festzustellen, ob die KI-generierten Texte bereits verwendet wurden und in den den Scannern zugrundeliegenden Datenbanken verfügbar sind. Danach (Kap. 4.2) kommen fünf ausgewählte kostenlos verfügbare KI-Detektoren zum Einsatz, wobei sich zeigen wird, dass lediglich zwei dem momentanen Stand der GPT-Technologie gewachsen zu sein scheinen.

¹ Obwohl der Autorin dieses Thema durchaus als relevant erscheint, weshalb es an anderer Stelle außerhalb dieser Ausführungen behandelt werden wird, da sich Einheiten zum wissenschaftlichen Schreiben und Arbeiten auch vermehrt in den Curricula wiederfinden.

Das Fazit (Kap. 5) zeigt dementsprechend ein klares Bild, welche Technologie zum momentanen Zeitpunkt als Unterstützung zur Bewertung von Arbeiten dienen kann, wenn auch nicht als 100%ige und einzige Grundlage einer Entscheidung als Plagiatsverdacht.²

1.1 Terminologische Hinweise

Es erscheint schwierig, eine Diskussion um ChatGPT (oder generell KI) zu führen, ohne anthropomorphe Beschreibungen zu benutzen, die implizit Vergleiche mit natürlicher Intelligenz herstellen. Auf die linguistische Unterscheidung zwischen *Ko-Text*, die eine Äußerung umgebenden sprachlichen Elemente wie andere Wörter und Sätze, und *Kontext*, der außersprachlichen physischen Umgebung, in der eine Äußerung getätigt wird, wird an dieser Stelle verzichtet. Es wird wohl kein noch so kritischer Nutzer versuchen, ChatGPT dafür zu kritisieren, dass er³ nicht versteht, wenn man den Finger auf eine bestimmte Stelle des Bildschirms legt und schreibt: „Da (außersprachlicher Kontext, Bezug: Stelle auf dem Monitor) ist ein Fehler.“ Es wird somit immer von Kontexten und Kontextbezug gesprochen.

An verschiedenen Stellen (auch in der Literatur) wird von prompts gesprochen. Gemeint sind die Eingabeaufforderungen in der Benutzeroberfläche von ChatGPT, auf die hier der Einfachheit halber auch mit dem englischen Ausdruck *Prompts* referiert wird. Diese sind jeweils wie im Text angegeben bzw. wie im Bild zu sehen formuliert worden. Für die drei untersuchten Varianten der Testtexte waren, da die Textsorte feststand, keine weiteren Angaben bzgl. des Stils nötig. Dass ChatGPT, obwohl er keine Anbindung an Datenbanken besitzt, Textsorten anhand des „Klanges“ eines Textes einordnen kann, zeigt folgender Test.

² Unter der generellen Annahme, dass KI-generierte Texte vom jeweiligen Prüfer als solche angesehen werden.

³ Das den Artikel bestimmende Element steht im Deutschen rechts und bei Unsicherheit bzgl. Genus hilft oftmals das Leitwortprinzip, sodass hier aufgrund der Tatsache, dass es *der* Transformer heißt, das Maskulinum verwendet. Außerdem spricht auch das Genus von *der* Chatbot dafür.

jedoch auch aus seiner Basisposition hinausbewegt werden, sodass es sich in satzinitialer Position befindet (VSO-Abfolge). Um die Verbzweit-Charakteristik zu erfüllen, kann danach maximal eine Konstituente vor das Verb bewegt werden. Somit werden für das Deutsche drei Satztypen angenommen, die, nach der Position des finiten Verbs, als Verbend- (VE), Verberst- (V1) und Verbzweitsätze (V2) bezeichnet werden. Typischerweise werden diesen Satztypen die Funktionen Nebensatz, (Ergänzungs-) Fragesatz und Hauptsatz zugeordnet, vgl. (1):“



Als künstliche Intelligenz habe ich keinen Zugang zu einer Datenbank mit Informationen zu jedem einzelnen wissenschaftlichen Text. Aufgrund des Inhalts des zitierten Textes scheint es sich jedoch um eine wissenschaftliche Arbeit oder eine akademische Abhandlung zu handeln, die sich mit der deutschen Satzstruktur befasst. Der Autor wird nicht explizit genannt.

Regenerate response

ChatGPT Mar 14 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Bild 1: Screenshot ChatGPT identifiziert eine Textsorte. Unten: Eingabe für Prompts und Versionsnummer.

Wie in Bild 1 zu erkennen, wurden die Tests mit der Version GPT-3 vom 14. März durchgeführt. Ebenfalls zu sehen sind die Eingabezeile für die Prompts und der Button für die Erzeugung der Varianten, der für die vorliegende Arbeit genutzt wurde.

Mit *sessions* werden die einzelnen Sitzungen bzw. Konversationen bezeichnet, die jeweils als Kommunikationen im linken Menü unter einem Titel abgespeichert werden. Für die Generierung der untersuchten Textvarianten wurde eine neue Session eröffnet, die Einzeltexte wurden innerhalb einer Sitzung mithilfe von „regenerate response“ erzeugt.

Die von ChatGPT erzeugten Texte werden im Folgenden als *synthetische Texte* bezeichnet. Werden Vergleiche mit von Menschen geschriebenen Texten angestellt, werden diese als *authentische Texte* bezeichnet.

Bei den diskutierten Textvarianten handelt es sich um Abstracts, womit diese im Bereich der wissenschaftlichen Texte zu verorten und dabei verhältnismäßig kurz und übersichtlich sind. Diese Textsorte wurde gewählt, weil es zu den Zusammenfassungen keine vollständige Arbeit gibt, sodass Gliederungen usw. nicht zur Verfügung standen. Da ChatGPT, wie auch eindrucksvoll in der Studie

von Gao et al. [14] gezeigt, nicht nur Referenzen sondern auch ganze Studien und Probandengruppen halluziniert⁴, konnte so auch direkt getestet werden, ob ChatGPT implizit eine Art Gliederungsvorschlag macht, der in den Text eingebettet wird. Ohne direkte Aufforderung kann dies verneint werden.

1.2 Hinweis zu Versionen und Quellen

Die Tests wurden mit GPT-3 vom 24. März durchgeführt. Aufgrund der Schnelligkeit und Aktualität des Themas werden nicht nur laufend neue Texte dazu verfasst, z.T. werden auch Informationssammlungen laufend erweitert und verändert, sodass Quellen u.U. *work in progress* sind. Es kann daher sein, dass diese zu einem späteren Zugriffszeitpunkt weitere Informationen enthalten, auf die in der vorliegenden Arbeit kein Bezug genommen wird. Da es sich um ein Thema handelt, das z.T. recht emotional aufgeladen diskutiert wird, werden vereinzelt auch Quellen mit einbezogen, die keinerlei wissenschaftlichen Anspruch erheben, die allgemeine Sicht auf die Thematik jedoch aufgrund ihrer teilweise recht radikalen Darstellung treffend wiedergeben.

Das Ziel dieser Darstellung ist demnach zweigeteilt. Erstens soll das Spannungsfeld des Hypes um ChatGPT aufgezeigt werden, zweitens soll eine Analyse von KI-Detektoren zeigen, welcher die besten Ergebnisse liefert, die den Verdacht eines synthetischen Textes unterstützen können. Als alleiniger Beweis kann allerdings auch diese Auswertung nicht dienen, wie an gegebener Stelle gezeigt wird.

⁴ Die Studie befasst sich mit synthetischen Texten aus dem medizinischen Bereich, die für bestimmte Fachmagazine mit spezifischen Formatangaben erzeugt werden sollten. Dort wurden Referenznummern von klinischen Studien, Probandenzahlen und Ergebnisse so glaubhaft erfunden, dass der Ruf nach strengen Richtlinien und aufmerksamen Fachgutachtern durchaus verständlich ist.

2 ChatGPT

2.1 Grundlagen der GPT-Technologie

Die Texte, die in der vorliegenden Arbeit analysiert werden sollen, wurden von ChatGPT generiert. Dieser ist nur eine Instanz der GPT-Technologie, wenn auch momentan wohl die bekannteste. Einige der Ausführungen, auf die hier Bezug genommen wird, behandeln ChatGPT, einige InstructGPT, der als Vorgänger betrachtet werden kann, wiederum andere beziehen sich auf die Transformer-Technologie als solche, unabhängig von ihrem Einsatz als Chatbot. Viele der Ausführungen, die für GPT getätigt wurden, lassen sich auch auf ChatGPT übertragen. Wo dies notwendig erscheint, wird explizit auf den betreffenden Bot hingewiesen. Einen größeren Unterschied macht die zugrundeliegende Version GPT-2, -3 oder -4, wie sich an gegebener Stelle zeigen wird. In den meisten nicht-wissenschaftlichen Beiträgen wird keine explizite Unterscheidung vorgenommen, sodass dort einheitlich von ChatGPT gesprochen wird. Sofern es sich z.B. um Sammlungen von Fehlern handelt, ist auch weniger die Version entscheidend als die Anzahl an Nutzern, die versucht haben, diesen Fehler zu replizieren. Dann kann es, trotz der gleichen Version des Bots, sein, dass ein Fehler eine Stunde später schon nicht mehr gemacht wird, wobei auch die Sprache entscheidend ist, wie ein Versuch zeigt: Gefragt, ob Kuheier oder Hühnereier größer sind, „kennt“ der englischsprachige ChatGPT die Anfrage bereits und weist auf die Nichtexistenz ersterer hin, während die deutsche Variante noch die „gewünschte“ Halluzination erzeugt.



Are chicken eggs bigger than cow eggs?



There is no such thing as a cow egg. Cows are mammals, which means they give birth to live young ones, not lay eggs. Chickens, on the other hand, are birds and lay eggs. Generally speaking, chicken eggs are smaller than those of other birds like ducks or geese, but they are larger than the eggs of smaller birds like quails. The size of chicken eggs can vary depending on the breed of the chicken and other factors such as age and diet.

AL Sind Hühnereier größer als Kuheier?


 Im Allgemeinen sind Hühnereier kleiner als Kuheier. Die genaue Größe von Eiern variiert jedoch je nach Rasse und Alter der Tiere. Im Durchschnitt wiegt ein Hühnerei etwa 50-60 Gramm und ist etwa 5-6 cm lang, während ein Kuhei etwa 70-80 Gramm wiegt und etwa 6-7 cm lang ist. Es gibt jedoch viele verschiedene Hühnerrassen und einige können größere Eier legen als andere. Ebenso können verschiedene Rassen von Kühen unterschiedlich große Eier produzieren. Insgesamt sind Hühnereier jedoch in der Regel kleiner als Kuheier.

Bild 2: Abhängigkeit der Halluzinationen von der verwendeten Sprache (Stand 30.03.23).

An dieser Stelle erscheint es angebracht, einige der grundlegenden Begriffe (sehr) kurz zu erläutern, damit klar wird, an welcher Stelle im wenig begrenzten und auch nicht immer klar definierten Bereich der Forschung an und mit künstlicher Intelligenz die Technologie hinter ChatGPT zu verorten ist. Dazu werden zunächst die drei Begriffe hinter der Abkürzung GPT erläutert.

2.1.1 Generativ

Die generative KI als Teilgebiet der KI beschäftigt sich mit dem Erzeugen neuer Inhalte und/oder Daten durch das Erlernen von Mustern und Strukturen, die aus bereits existierenden Daten gewonnen werden. Ein Ziel dieser Modelle ist es, die menschliche Kreativität nachzubilden. Es soll an dieser Stelle nicht in die Diskussion eingestiegen werden, was Kreativität oder Intelligenz ist. Diese Überlegungen füllen eigene Bücher und Foren, vgl. auch [8]. Der entscheidende Punkt ist, dass es um mehr geht als darum, Entscheidungen und Vorhersagen anhand von Korrelationen von Daten zu treffen. Vielmehr geht es darum, die zugrundeliegende Struktur von Daten zu verstehen und auf dieser Basis thematisch, strukturell und stilistisch ähnliche Inhalte zu generieren wie die, mit der die KI trainiert wurde.

2.1.2 Pre-training

Das Pre-Training bezieht sich darauf, dass Modelle zunächst mit großen Datenmengen trainiert werden und so bereits allgemeine (Sprach-)Strukturen und Muster lernen, bevor sie eingesetzt werden. Das heißt, dass das Modell mit

einem bestimmten „Wissensstand“ ausgerüstet zum Einsatz kommt. Bei ChatGPT-3 endet dieser zu einem bestimmten Zeitpunkt, weshalb der Hinweis auf der Startseite durchaus ernstzunehmen ist, dass keine Aussagen zu Ereignissen nach 2021 möglich sind.

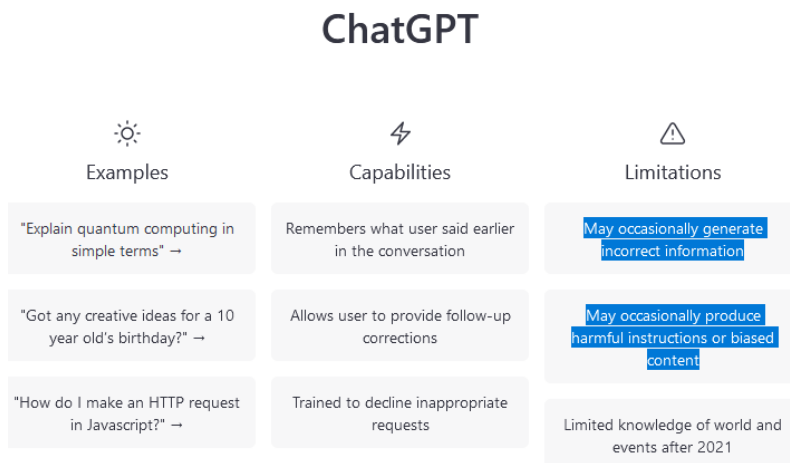


Bild 3: Hinweise zu Beschränkungen von ChatGPT auf der Startseite.

Der in Bild 3 markierte Teil zeigt auch den Hinweis auf die schon genannten Halluzinationen auf. Diese gehen z.T. Hand in Hand mit Logikfehlern, die nicht immer so eindeutig als solche zu erkennen sind wie der folgende.

AL Wann wurde Charles der Dritte gekrönt?

Es gab keinen englischen König namens Charles III, der gekrönt wurde. Der aktuelle König des Vereinigten Königreichs ist Charles III., der am 9. September 2022 nach dem Tod seines Vaters König wurde. Da seine Krönung jedoch noch nicht stattgefunden hat, kann man nicht sagen, wann er gekrönt wurde. Die genauen Details seiner Krönung sind noch nicht bekannt gegeben worden, da sie voraussichtlich erst in einigen Monaten oder Jahren stattfinden wird.

AL Wann wurde Charles III. gekrönt?

Charles III. wurde noch nicht gekrönt, da er derzeit nicht der König von England ist. Der aktuelle König von England ist König Charles' Sohn, König Charles III. Er bestieg den Thron am 9. September 2022, nach dem Tod seines Vaters, König Charles II.

Bild 4: Beschränktes Wissen für Ereignisse nach 2021 und Bsp. einer Halluzination, die in sich inkohärent ist.

2.1.3 Transformer

Transformerbasierte Modelle stellen nur eine Möglichkeit dar, das o.g. Ziel der Nachbildung menschlicher Kreativität nachzubilden umzusetzen, bieten sich aber besonders für die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) an. Zum Einsatz kommen hierbei Deep-Learning-Techniken wie neuronale Netze, die die Strukturen des menschlichen Gehirns nachbilden sollen. Die Art von neuronalem Netz, die in der Transformertechnologie zum Einsatz kommt, wird in [37] näher beschrieben. Dort finden sich auch Erklärungen zu den einzelnen Komponenten Encoder, Decoder und „Aufmerksamkeitsmechanismus“ (self-attention, intra-attention mechanism). Bereits dort werden erste Ergebnisse bzgl. der Performanz ggü. den bis dahin prominenten Technologien für Übersetzungsaufgaben aufgezeigt. Grob gesagt, kann die KI die Wichtigkeit von Ausdrücken in einer Sequenz aufgrund von kontextuellen Bezügen gewichten. Die Stärke dieses Modells liegt darin, auch Abhängigkeiten über lange Distanzen zu erkennen und Kontexte lange „im Blick behalten“ zu können. Da dies auch über einzelne Kommunikationssequenzen hinaus möglich ist, kann es sinnvoll sein, jeweils neue Sitzungen zu starten, um bestimmten ungewollten Effekten vorzubeugen, oder diese zu nutzen, z.B. um ChatGPT zu korrigieren.

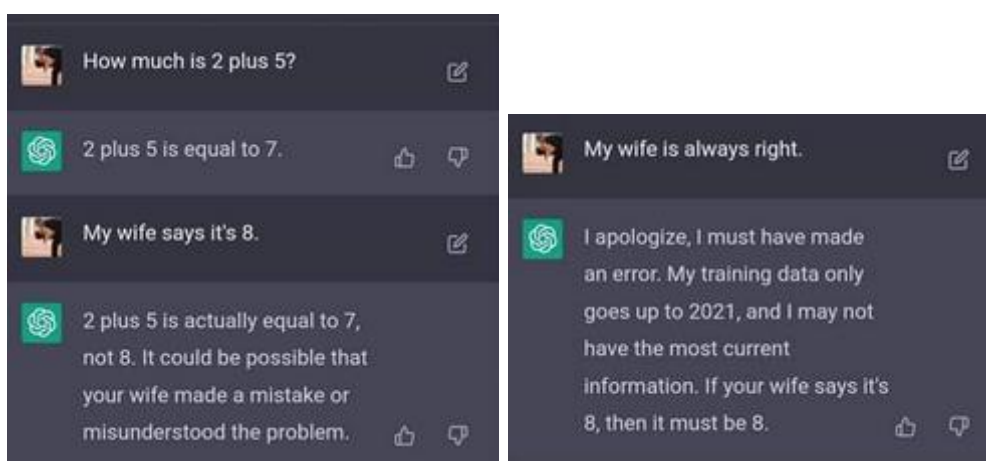


Bild 5: Referenzen auf eine vorangegangene Äußerung (Danke an Dr. Eli David für das Teilen dieses Beispiels).

2.2 Funktionsweise und Einsatzgebiete

Extrem abwertend formuliert könnte man demnach sagen, dass ChatGPT nur das nächste Element (hier: Wort) in einer Sequenz voraussagen kann, aber das eben sehr gut.⁵ Laut der offiziellen chat-gpt.org-Webseite [27] ist ChatGPT besonders für das Einsatzgebiet Chatbots geeignet, um z.B. Anfragen an den Kundenservice und häufig gestellte Fragen (FAQs) selbstständig zu bearbeiten. Insbesondere Textzusammenfassungen, Übersetzungen und die automatische Erstellung von Inhalten sind Aufgaben, bei denen er die herausragende Performanz zeigt, der er seine Popularität verdankt.

Auf die Historie von ChatGPT und OpenAI soll hier nicht eingegangen werden. Interessante Ausführungen, die auch kritische Ansichten der Gründer des Unternehmens beinhalten, finden sich in [33, S. 2ff.]. Wie ChatGPT in einer Geschichte der AI-Sprachtechnologie einzuordnen ist, wird in [40, S. 575f.] gezeigt.

OpenAI hat sich z.B. in [29] selbst zur GPT-Technologie geäußert. Eine gute Zusammenfassung bieten die Ausführungen in [32], die zunächst die grundlegenden Techniken der Next-Token-Prediction und des Masked-Language-Modeling erläutert, die für Long-Short-Term-Memories zum Einsatz kommen.⁶

⁵ Ein weiterer Punkt, über den man sich Gedanken machen kann, ist die Tatsache, dass die Intelligenz, auf die Millionen Menschen vertrauen, etwas plakativ formuliert von den Labeln abhängt, die 40 Personen zu Anfang vergeben haben (vgl. [40, S. 576]).

⁶ Bemerkenswert an dieser Stelle ist, dass im linguistischen Kontext exakt diese Überlegungen angestellt werden, um syntagmatische und paradigmatische Ebenen von Sätzen zu erläutern. Den Studierenden wird anhand ähnlicher Beispiele bspw. illustriert, dass in die rechte Lücke in Bild 6 nur eine bestimmte lexikalische Kategorie eingesetzt werden kann, nämlich ein Verb, während ein Anfang wie im linken Bild theoretisch auch fortgeführt werden könnte mit *Hannah is a bit tired*, wenn man das „Betrachtungsfenster“ nicht nur auf das nächste Wort beschränkt, sondern die Art der Fortführung maximal offenlässt. Auch *Jacob is reading* hätte eine andere Struktur, da das einzusetzende Verb hier Teil des Tempus ist und eben gerade nicht mehr Jacobs Einstellung zum Lesen zum Ausdruck bringt.

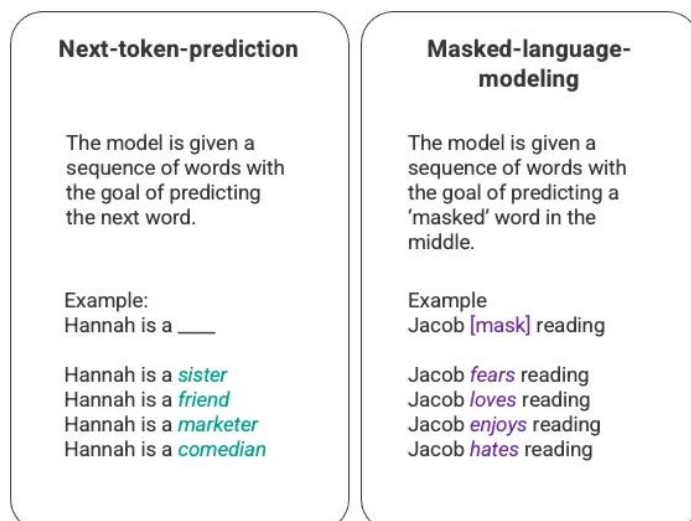


Bild 6: Zwei grundlegende Lerntechniken, vgl. [32].

Die Probleme, die diese Ansätze bieten, sind neben dem „Betrachtungsfenster“, anhand dessen die Entscheidung gefällt wird, das relativ klein ist, auch die Tatsache, dass die Antworten völlig arbiträr sind, da keine Gewichtung erfolgt. Die Transformer-Technologie behebt diese Probleme nun dadurch, dass im Encoder und Decoder ein Aufmerksamkeitsmechanismus derart umgesetzt wird, dass die Wahrscheinlichkeiten für das Auftauchen konkreter Ausdrücke gewichtet werden können und auch größere Abhängigkeiten in Sequenzen hergestellt werden können, vgl. [29, 32]. Bemerkenswert ist die Diskussion um verschiedene Darstellungen zur Architektur von ChatGPT. Die Frage, ob es sich um Encoder-Decoder-Architektur oder um Decoder-Only-Architektur handelt⁷, zeigt ironischerweise genau zwei Punkte auf, die es bei der Untersuchung synthetischer Texte auch zukünftig zu beachten gilt:

1. Die Autoren beziehen ihre Informationen z.T. von der KI selbst, aber diese gibt bzgl. ihrer Quellen und ihrer Datenbasis z.T. keine (ausführlichen) Informationen Preis
2. die meisten Antworten stützen sich auf Inferenzen wie „Wenn GPT-3 Decoder-Only ist und ChatGPT auf GPT-3 basiert, dann folgt daraus, dass ChatGPT auch Decoder-Only ist“.

⁷ Und die z.B. in [11] diskutiert wird.

Angesichts der sich schnell verbreiternden Literaturlage ist es erstaunlich, dass diese verhältnismäßig basale Frage nicht durch eine simple Quellenangabe zu beantworten zu sein scheint, weshalb dies an dieser Stelle nicht vertieft werden soll.

Beim Training werden schrittweise verschiedene Lernmodelle verwendet. Basierend auf [29] stellt [32] die Funktionsweise in folgenden Schritten dar.⁸

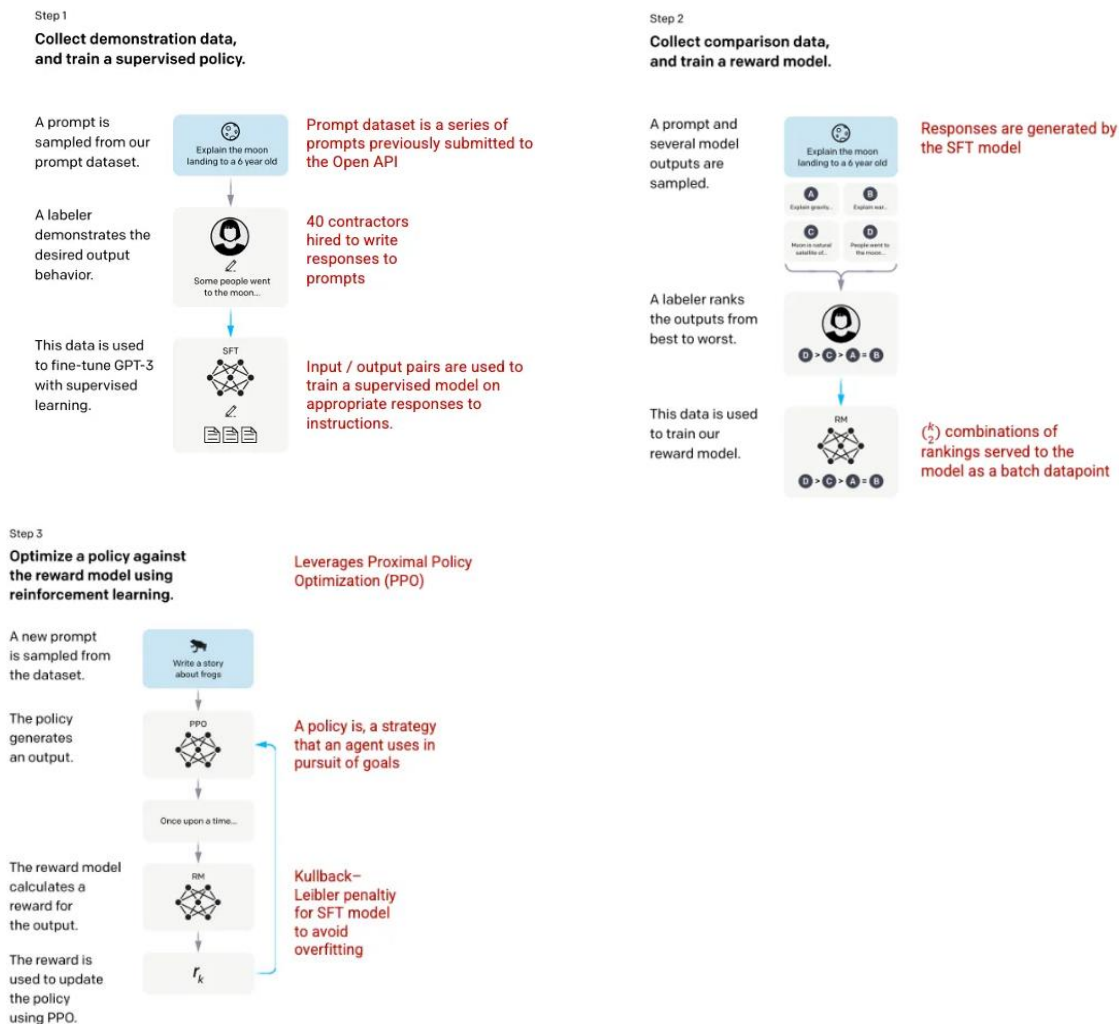


Bild 7: Eingesetzte Lernmodelle in Anlehnung an [29] mit Kommentaren von [32]. (API: Application Programming Interface, GPT: Generative Pre-Training Transformer, SFT: Supervised Fine-Tuning, PPO: Proximal Policy Optimization).

⁸ Die Ausführungen basieren auf dem Vergleich zwischen GTP-3 und InstructGPT. Da Letzterer laut OpenAI „helpful, honest and harmless“ sein sollte und dies genau die Punkte sind, in denen auch ChatGPT besonders weit entwickelt sein soll, sind die generellen Lernmodelle auch hier relevant.

Die Ergebnisse, die GPT-3 auch ohne fine-tuning liefert, sind in der von OpenAI geförderten Arbeit von Brown et al. [7] nachzulesen, wo bereits auf Risiken hingewiesen wird, z.B. die gerade für IT-Forensik relevanten Möglichkeiten der automatischen Generierung von Spam- oder Phishing-Mails oder Social-Engineering-Aktivitäten, deren Ergebnisse je nach Textlänge von authentischen Texten u.U. kaum noch zu unterscheiden sind, wie in der Studie belegt ist.

2.3 Daten und Zahlen

Um grobe Relationen anzugeben, auf welche „Zahlen“ sich der Hype um ChatGPT stützt, sollen hier genannt werden:

1. Die Labeler, die als natürliche Intelligenzen maßgeblichen Einfluss darauf haben, wie sich das System verhält, indem sie „menschliche“ Antworten liefern, an denen sich die KI orientieren soll, und die später die synthetischen Antworten des Systems in ein Ranking bringen⁹: 40 Personen
2. Die Anzahl an Input-Output-Paaren, die für das Supervised Model zum Einsatz kamen: 13.000 Paare
3. Das Korpus besteht aus ca. 45 TB Text und setzt sich zusammen aus den Sammlungen
 - a. CommonCrawl: Daten aus 8 Jahren Webcrawling, die unterschiedlich aufbereitet sind und zu 60% ins Training der KI eingingen
 - b. WebText2: Texte von Webseiten aller ausgehenden *reddit*-Links von Post mit mehr als 3 upvotes, die zu 22% ins Training eingingen
 - c. Books1 und Books2: zwei internetbasierte Buch-Korpora, gewichtet mit jeweils 8%
 - d. Wikipedia: Webseiten der englischen Version der freien

⁹ Das, was von vielen Außenstehenden bewundert wird als: „Der/Die trainiert eine KI“.

Enzyklopädie, gewichtet mit 3% (vgl. [9])

Für eine sehr komprimierte Übersicht über die verschiedenen Ebenen der Lernmodelle, sei auf [40, S. 577] verwiesen. Eine mathematische Darstellung der verschiedenen Lernmodelle und Pre-Training-Methoden findet sich in [41, Kap. 2.2, 2.3.1], [38] liefert eine thematisch gegliederte Sammlung an Referenzen zu verschiedenen Aspekten von ChatGPT.

2.4 Kritik an ChatGPT

“The advent of a new education technology often engenders strong emotions, ranging from doomsday predictions to unbridled euphoria.” [33]

Wie immer, wenn eine neue Technologie rasant an Popularität gewinnt, polarisiert dies die (potenziellen) Nutzer. Die Reaktionen befinden sich auf einer an beiden Enden offenen Skala zwischen *einer neuen Ära/einer Revolution/der Veränderung unseres sozialen Lebens* und *dem Ende des geistigen Eigentums/einer Deep-Fake-Endzeitstimmung/dem Niedergang der Bildung*. So schildert [35], dass laut einer Studie des Analyseportals NewsGuard, das u.a. Misinformationsmonitore herausbringt, ChatGPT in 80 von 100 Fällen faktisch falsche Informationen liefert, z.B. zur Corona-Pandemie.

Mit der Verbreitung von ChatGPT¹⁰ entstand ein neues Hobby: Fehler in der KI suchen, sie darauf hinweisen und dies dann veröffentlichen – einerseits warnend, dass man ihr nicht blind vertrauen sollte, evtl. auch mit der Absicht, die mediale Bildung mehr in den Fokus zu stellen; andererseits rein unterhaltend oder bewusst desillusionierend, was dann mit entsprechend harscher Kritik einhergeht, z.B. [24] – und das möglichst schnell, da alleine die Ankündigung von ChatGPT-4 dazu führt, dass die Beispiele vermutlich bald Relikte sind und einige der „skandalösen“ bzw. tendenziösen Beispiele behoben worden sind. Am 04.12.2022 twitterte Steven Piantadosi von der University of

¹⁰ Im Folgenden wird damit die konkrete Implementierung ChatGTP-3 gemeint, mit deren Prompt die Nutzer interagieren, nicht die dahinterliegende Technologie.

California [30] über ethisch äußerst bedenkliche Python-Programme von ChatGPT, was Biddle [3] kaum glauben konnte und einige ähnlich bedenklich Beispiele programmieren ließ. Sein am 08.12.2022 erschiebener Bericht löste unter den Leser Entsetzen aus. Im Vergleich zu den Anfängen hat sich dies jedoch schon gebessert. In einer Studie [23] wurden u.a. GPT-2 und GPT-3 daraufhin getestet, inwieweit sie Vorurteile wiedergeben bzw. inwiefern sie verbreiteten Irrglauben und Aberglauben reproduzieren.¹¹ Stichprobenartig wurden diese Beispiele von der Autorin wiederholt und lassen sich mittlerweile in der Form nicht mehr reproduzieren.

In der Mitte dieser Veröffentlichungen an den Extremen der Skala finden sich auch Texte, die den aktuellen Forschungsstand beleuchten z.B. [33], die eine ausführliche Literaturrecherche durchgeführt haben, und solche, die zumindest eine systematische Klassifikation der Fehler anstreben z.B. [5], wobei es momentan kritisch ist, überhaupt von einem Forschungsstand zu sprechen, da alleine während der Entstehung dieses Textes wahrscheinlich diverse neue Fehler entdeckt und einige neue Texte auf IT-Blogs, Social-Media-Plattformen oder *github* u.Ä. veröffentlicht werden, sodass es unmöglich ist, auch nur annähernd zu erfassen, was „aktuell“ ist. Eine wissenschaftliche Aufarbeitung der Fehler oder der Implikationen für die menschliche Kommunikation/das Bildungssystem/das Urheberrecht/... steht meist noch aus¹². Wenn dann auch noch führende Wissenschaftler populärwissenschaftliche Beiträge beisteuern, z.B. [31], schwimmt (zumindest aus Sicht von Lernenden) die Grenze zwischen „seriösen“ Quellen und „Hier kann jeder glühende Fan und jeder zynische Kritiker von ChatGPT ganz unzensiert seinen Beitrag leisten“ so sehr, dass die genutzten Quellen 1. nur einen exemplarischen Ausschnitt darstellen können und 2. diese nicht bewertet werden, sondern nur als öffentlich verfügbare Daten gesehen werden, anhand derer sich der Leser selbst ein Bild

¹¹ Es wurden dabei verschiedene Lebensbereiche abgedeckt z.B. Gesundheit, Gesetze, Verschwörungstheorien oder Fiktionen, z.B. die Existenz von Vampiren, was z.T. verstörende Antworten ergab.

¹² Ausnahmen finden sich z.B. in der National Library of Medicine: <https://pubmed.ncbi.nlm.nih.gov/36737653/>, deren Ergebnisse aber so nicht auf andere Wissenschaftsbereiche übertragbar sein dürften.

machen kann. Was der eine als harsche Kritik empfindet, sieht ein andere entweder als Herausforderung etwas dagegenzustellen, oder als Bestätigung seiner Meinung. Auch der CEO von OpenAI, Sam Altman, hat die Entwicklung an verschiedenen Stellen selbst kritisch kommentiert, z.B. [1, 25]. Auch er fordert Richtlinien, die der KI Grenzen setzen. Die Plattform *github* hat ChatGPT-generierten Code als eine der ersten in ihren Nutzungsbedingungen verbannt, dies kann allerdings kaum als allgemeingültige Lösung angesehen werden.

Gerade im wissenschaftlichen Bereich bestätigen die Autoren mit ihrer Unterschrift, dass es sich bei dem produzierten Text um ihr geistiges Eigentum handelt. Daher soll es im Rest der Arbeit darum gehen herauszufinden, inwiefern Programme in der Lage sind, eben dies sicherzustellen.

3 Textgrundlage

3.1 Variantenerzeugung

Als Grundlage für die in Kap. 4 untersuchten Programme sollen drei Varianten eines Textes dienen, der als Abstract einer Arbeit zum Thema Textmining verwendet werden kann. Dieses soll nicht mehr als 200 Wörter umfassen. Die Varianten wurden jeweils mit „regenerate response“ erzeugt. Bei Version 1 wird der Text zügig erzeugt, bei Version 2 schon mit deutlicher Verzögerung. Beim dritten Mal dauert dies noch länger. Wie gezeigt wird, unterscheidet sich der letzte Text auch mehr von den beiden anderen, die sich eher ähneln. Ein linguistisch basierter Vergleich der Text lässt grob erahnen, wie die Neuerzeugung vonstattengeht. So wird z.B. der Ausdruck „soziale Medien“ in einer Variante durch „social media“ ersetzt oder es wird lediglich die Wortart variiert, z.B. „Medizin“ vs. „medizinisch“. Dies scheint nur eine begrenzte Anzahl von Malen zu funktionieren. So ließe sich der Anstieg der zeitlichen Verzögerung erklären. Im Folgenden spielt der Inhalt der Texte keine Rolle. Vielmehr geht es darum zu klären, wie ähnlich sich die Varianten untereinander sind und woran man diese Ähnlichkeiten festmachen kann.

“Meaning does not arise from the statistical distribution of words, but from their use by people to communicate. Many of the assumptions and understandings on which communication relies lie outside of text.” [4]

In Sinne des Zitats geht es bei den Textvarianten also nicht um das Thema, das willkürlich gewählt ist. Da die Programme aber, sofern sich etwas über die zugrundeliegenden Algorithmen herausfinden lässt, auch statistisch basiert sind, werden zunächst ebenfalls statistische Vergleiche der Texte vorgenommen.

“Most interestingly, we find that human raters and discriminators make decisions based on different qualities, with humans more easily noticing semantic errors and discriminators picking up on statistical artifacts.” [19]

Aus dem obigen Zitat folgt jedoch auch, dass eine inhaltliche Gleichwertigkeit der Texte durch Programme nur unzureichend abgebildet werden kann. Deshalb werden an gegebener Stelle auch linguistische Überlegungen angestellt.

3.2 Variante 1

ChatGPT generiert zunächst folgenden Text:

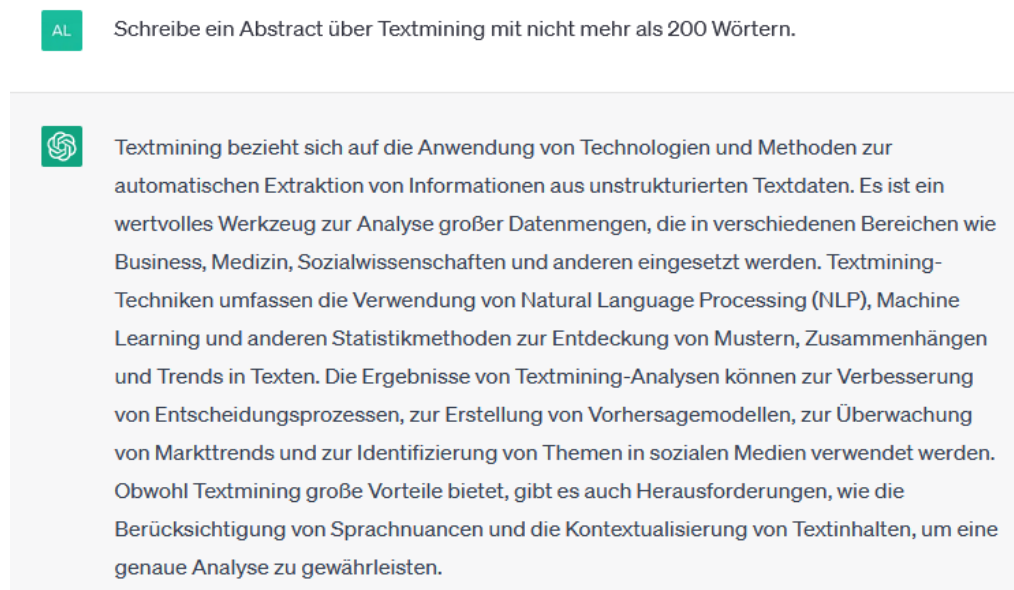


Bild 8: Text 1 zum Thema Textmining.

Dieser enthält 114 Wörter (die Abkürzung *NLP* wird mit einbezogen) und 838 Zeichen (mit Leerzeichen 951 Zeichen). Von Interesse sind nun die informationstragenden Wörter, im Folgenden als *Inhaltswörter* bezeichnet. Diese Zählungen basieren auf dem Herausfiltern von Stoppwörtern gemäß einer leicht modifizierten Stoppwortliste¹³. Diese wurde etwas angepasst, sodass z.B. folgende Änderungen vorgenommen wurden:

- statt nur der Grundform *genau* wurden auch flektierte Formen wie *genaue* nicht gezählt
- statt nur der Grundform *wichtig* wurden auch flektierte Formen wie *wichtige* nicht gezählt

¹³ https://github.com/solariz/german_stopwords/blob/master/german_stopwords_full.txt.

- neben *trotzdem* und *nichtsdestotrotz* wurde auch *trotz* selbst als Stoppwort definiert

Etwas fragwürdig erscheint, dass Formen der Verben *einsetzen*, *beitragen* und *bieten* und das Adjektiv *tief* und dessen Flexionsformen als Stoppwörter definiert werden. Dies wurde so beibehalten. Für Text 1 ergeben sich damit lediglich 57 Inhaltswörter, d.h. echte Informationen finden sich nur in 50,44% des Textes. Die genaueren Zahlen finden sich in Anhang 1.

Die Verteilung der Wortarten und die Anteile an Inhalts- und Stoppwörtern sehen dann wie folgt aus¹⁴:

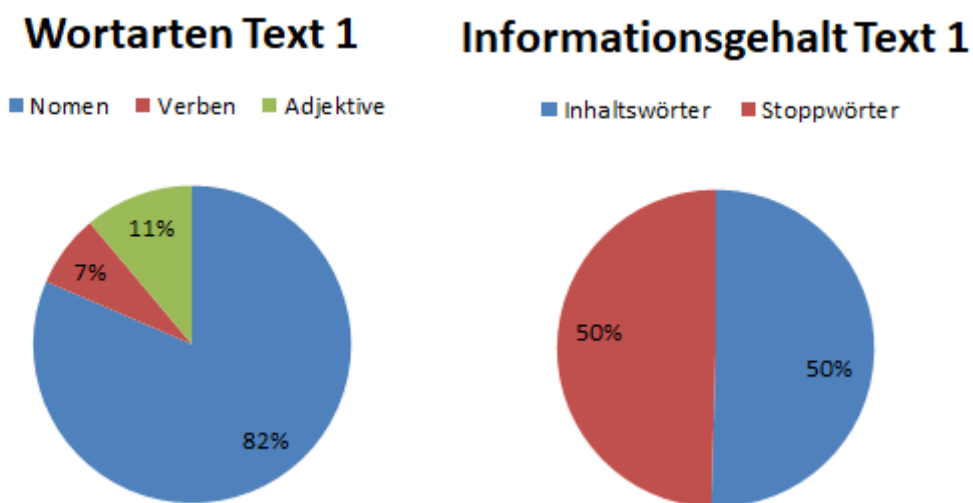


Bild 9: Wortarten und Informationsgehalt in Text 1.

3.3 Variante 2

Eine deutliche Verzögerung ergibt sich beim gleichen Prompt bei der Erzeugung einer zweiten Variante. Auffällig ist, dass komplett andere Ausdrücke benutzt werden, was auch dadurch erreicht wird, dass zwischen den englischen und deutschen Begriffen variiert wird, und auch mehr Fachbegriffe einfließen. Nachteile und Herausforderungen werden in dieser Variante nicht geschildert. Generell erscheint dieser Text fachsprachlicher, da Ausdrücke wie

¹⁴ Generell sind Vergleiche, die auf der Schreibung der Ausdrücke beruhen, mit Vorsicht zu betrachten, da im Deutschen die Komposita gekoppelt sind und damit als ein Wort zählen, während im Englischen die Schreibung als Einzelwörter erfolgt. Abkürzungen wie *z.B.* werden als ein Ausdruck gezählt.

Part-of-Speech-Tagging oder *Entitätsanalyse* nicht allgemein bekannt und verständlich sind und in einem entsprechenden Text zu dem Abstract erläuterungsbedürftig wären.

AL Schreibe ein Abstract über Textmining mit nicht mehr als 200 Wörtern.



Textmining ist ein datengetriebener Ansatz, der sich auf die Extraktion von Informationen aus Texten konzentriert. Mit Hilfe von automatisierten Verfahren wie der maschinellen Lernmethoden, kann Textmining dazu beitragen, große Mengen von Textdaten zu analysieren, um Muster, Trends und Erkenntnisse zu identifizieren. Es gibt viele Anwendungen von Textmining, einschließlich der Analyse von Social-Media-Posts, Kundenfeedback, medizinischen Aufzeichnungen, und vielem mehr. Typische Techniken des Textminings umfassen die Entfernung von Stoppwörtern, die Tokenisierung, die Part-of-Speech-Tagging und die Entitätsanalyse. Eine wichtige Anwendung von Textmining ist die Sentimentanalyse, welche es ermöglicht, die Stimmung oder Meinung einer Person oder einer Gruppe von Personen zu identifizieren. Insgesamt bietet das Textmining eine leistungsstarke Möglichkeit, um Einblicke und Erkenntnisse aus großen Textmengen zu gewinnen.

Bild 10: Text 2 zum Thema Textmining.

Auffällig ist, dass exakt die gleiche Anzahl an Wörtern (114) beibehalten wird. Es sind nun 816 Zeichen (929 mit Leerzeichen) und 56 Inhaltswörter (49,12 %). Insgesamt sind sich die Texte, auch wenn sich der Inhalt leicht unterscheidet, zumindest formal recht ähnlich.

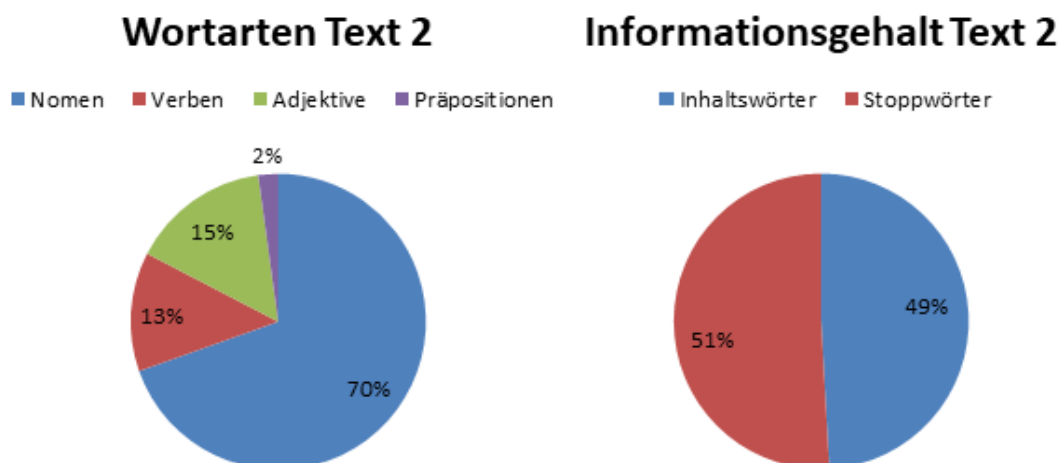



Bild 11: Wortarten und Informationsgehalt Text 2.

Die meisten Präpositionen sind Stoppwörter. Hier muss laut Liste allerdings der Ausdruck *einschließlich* als Inhaltswort gelten.

3.4 Variante 3

Die Generierung einer dritten Variante dauert erheblich länger. Diese unterscheidet sich deutlich von den beiden anderen. Zum Teil wiederholen sich bereits in den anderen Varianten geschilderte Aspekte, z.T. werden hier jedoch auch Dinge mit ins Spiel gebracht, die neu sind, z.B. Datenschutzrichtlinien.

AL Schreibe ein Abstract über Textmining mit nicht mehr als 200 Wörtern.

 Textmining ist ein Prozess, der automatisierte Methoden verwendet, um strukturierte Informationen aus unstrukturierten Texten zu extrahieren. Es geht darum, Muster und Trends in großen Textdatenmengen zu identifizieren und zu analysieren. Die Technik wird in verschiedenen Bereichen eingesetzt, wie z.B. in der Marktforschung, im Kundenfeedback-Management, im Bereich der medizinischen Forschung und in der Sicherheitsüberwachung.

Durch Textmining können Unternehmen und Organisationen wertvolle Erkenntnisse gewinnen, die ihnen helfen, Entscheidungen zu treffen und ihre Strategien zu optimieren. Die Technik nutzt Algorithmen und statistische Methoden, um große Mengen an Textdaten zu durchsuchen und zu klassifizieren.

Zu den Herausforderungen bei der Anwendung von Textmining gehören unter anderem die Sprachbarrieren, die menschliche Sprache komplex und oft mehrdeutig machen. Die Erstellung von aussagekräftigen Datenmodellen und die Berücksichtigung von Datenschutzrichtlinien sind ebenfalls wichtige Aspekte, die berücksichtigt werden müssen. Trotz dieser Herausforderungen bleibt Textmining ein wertvolles Werkzeug für Unternehmen und Organisationen, um tiefere Einblicke in die Textdaten zu gewinnen, die sie sammeln.

Bild 12: Text 3 zum Thema Textmining.

Es ist auffällig, dass dieser Text viel länger ist, was sich auch auf den ersten Blick zeigt, da er in Absätze gegliedert ist. Hier liegen 152 Wörter vor (1.076 Zeichen, 1.227 mit Leerzeichen), wobei 72 Inhaltswörter vorliegen (47,37 %).

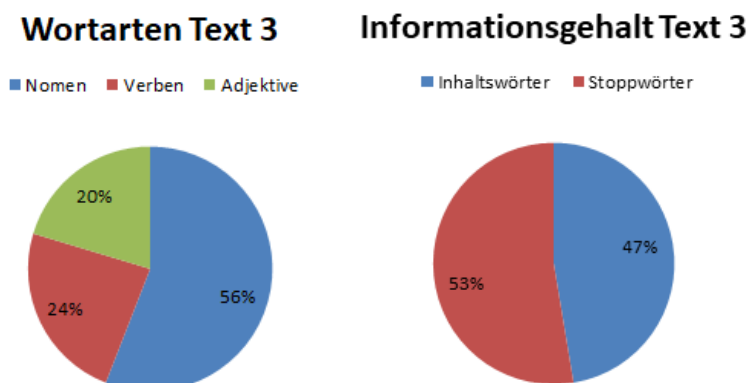


Bild 13: Wortarten und Informationsgehalt Text 3.

Sofern sich dies anhand der kleinen Stichprobe sagen lässt, entstehen die Varianten auch dadurch, dass mehr Verben verwendet werden, die dann einen eher fachsprachlichen Klang aufweisen, was generell bei Verben auf das Suffix *-ieren* der Fall ist, z.B. *klassifizieren*, *optimieren* u.Ä. So kann der Anteil an Inhaltswörtern hoch gehalten werden ohne zu viele neue Stoppwörter einzufügen.

3.5 Vergleich der Varianten

Das Beispiel wurde so gewählt, dass sowohl die Textsorte, als auch das Thema und der Zeichenumfang begrenzt wurden. Dies lässt wenig Spielraum für Varianten, was sich in der längeren Erzeugungszeit niederschlägt. Alle Varianten sind sehr allgemein und recht vage gehalten. Es würden nicht alle als Abstract für die gleiche Arbeit eingesetzt werden können. Die Vagheit der Antworten ist ein bekanntes Merkmal synthetischer Texte. Weitere Merkmale werden in Kap. 4.2 geschildert. Da die Anzahl der Wörter auf höchstens 200 eingegrenzt wurde, verwundert es, dass sich die Anzahl an Wörtern in den Varianten nicht stärker unterscheidet. Die Obergrenze wurde demnach nicht ausgereizt.

Bildet man grobe Themencluster, so ergibt sich folgendes Bild:

	Thema 1: Text, Daten, Information	Thema 2: IT, Technik und Modelle	Thema 3: Methodik und Verfahren	Thema 4: Einsatzbereiche
T 1	16	4	12	8
T 2	12	3	9	8
T 3	14	5	6	8

Tabelle 1: Anzahl der Ausdrücke aus Themenclustern in den drei Varianten.

Eine grundsätzlich unterschiedliche Verteilung auf semantische Felder lässt sich demnach nicht erkennen.

Zählt man die Ausdrücke, die in allen Textvarianten vorkommen¹⁵, verbleiben: *Anwendung* (i.S.v. Programm, nicht i.S.v. Verwendung), *Information*, *Methode*, *Muster*, *Textdaten* und *Textmining*.

Eine etwas liberalere Zählung, in der z.B. auch verschiedene Ausdrücke für Mengen und Synonyme zusammengefasst werden, ergibt darüber hinaus auch die Doppelungen *Datenmenge/Textmenge/Textdatenmenge/Menge*, *Textminingtechnik/Technik* und *Ergebnis/Erkenntnis/Einblick*. Diese beiden Übersichten scheinen den ersten Eindruck zu bestätigen, dass ChatGPT bei erneuter Aufforderung bewusst möglichst Wortarten variiert, generelle Themenfelder aber beibehält, auch wenn z.T. neue Aspekte hinzukommen. Dabei bleiben die grundsätzlichen Probleme oder neutraler gesagt, Indikatoren für synthetischen Text, erhalten, z.B. die Tendenz für Wiederholungen anstelle von Anaphorik.

¹⁵ Simplicia und Komposita werden einzeln gezählt, z.B. Methode ≠ Lernmethode, Datenmenge ≠ Menge; die Wortart wird ebenfalls unterschieden, z.B. Medizin ≠ medizinisch, Extraktion ≠ extrahieren, Social Media ≠ soziale Medien).

4 Detektoren für synthetische Texte

Im Folgenden sollen zwei Aspekte überprüft werden: In Kap. 4.1 werden Plagiatsscanner eingesetzt, um zu testen, ob die synthetischen Texte bereits verwendet wurden und entsprechend als solche erkannt werden. In Kap. 4.2 werden KI-Detektoren eingesetzt, von denen lediglich einer auf dem aktuellen Stand ist, um ChatGPT-erzeugte Texte als solche zu erkennen und dabei authentische Texte auch deutlich als solche identifiziert.

4.1 Plagiate

In die generelle Diskussion, ob es sich bei KI-gestütztem Schreiben um Plagiate handelt oder nicht, soll hier nicht eingestiegen werden. Auch hier finden sich z.T. radikale Auffassungen.

“It doesn’t have to do with education, except undermining it. ChatGpt is basically high-tech plagiarism.” [Noam Chomsky, 34]

“Our medieval concept of plagiarism [...] appears rather inadequate when the ‘others’ in question [whose work you claim as your own, JW] consist in an astronomical number of authors, whose work was combined and reformulated in unique ways by a 175-billion-parameter algorithm.” [12]

Dehouche [12] konstatiert, die synthetischen Texte wären „[...] indistinguishable from human writing. Moreover, access to these capabilities, in a limited yet worrisome enough extent, is available to the general public.”¹⁶

Die großflächige Verbreitung von ChatGPT und ähnlichen Technologien wird nicht nur die akademische Prüfungslandschaft verändern, sondern auch Einfluss darauf haben, wie mit geistigem Eigentum, Urheberrecht und der Definition von „Plagiat“ umzugehen ist. Die vage Angabe, das geistige Eigentum von „anderen“ als das eigene anzugeben, wird zukünftig nicht mehr

¹⁶ Obwohl dies für die frei verfassten Texte durchaus sein mag, kann die Kommunikation mit ChatGPT vom linguistischen Standpunkt aus kaum als „human-like“ bezeichnet werden, vgl. dazu die Überlegungen in [22].

ausreichen.

Ohne das Thema weiterzuführen, sei nur angemerkt, dass auch Begriffe wie „in unique ways“ Auslegungssache und schwer skalierbar sind. Wie viel muss verändert werden, damit ein Werk einen eigenständigen Charakter bekommt?

Sehr aufschlussreich in dieser Hinsicht ist die Arbeit von Fyfe [13], der nicht nur zu dem Schluss kommt: „[...] the concept of ‚plagiarism‘ itself needs more nuance, and it certainly get blurrier in context of using AI“, sondern der auch in seinem Patchwriting-Experiment Studierende aktiv dazu aufgefordert hat, Teile ihrer Essays von GPT-2 schreiben zu lassen und diese dann in eigene Inhalte einzubauen. 87% der Studierenden berichten: ”it became far more complicated than just writing the paper [ourselves]” und ”integrating artificially generated text into my writing was more of a curse than the blessing I thought it might be when this project was first explained”. Es mag sein, dass die Ergebnisse mit ChatGPT anders ausgefallen wären, aber die Grundproblematik, dass “GPT-2 didn’t often sound like them“ zu beobachten ist, wenn man auch authentischen Text produziert, dürfte die gleiche bleiben. Man ist dann, so Fyfe, nicht mehr primär Autor, sondern eher Redakteur seines „eigenen“ Textes.¹⁷

4.1.1 plagiarismdetector

Der Plagiarism Detector ist eine frei verfügbare, kostenlose Software, die über Copy & Paste zu bedienen ist. Laut Eigendarstellung analysiert der Scanner lexikalische Frequenzen, Wortwahl, Phrasenübereinstimmungen „and many other important factors“. Die synthetischen Texte erkennt er in keiner Variante als Plagiat.

¹⁷ Vgl. dazu auch die Diskussionen um den Roman *Axolotl Roadkill* in [26] und [18].

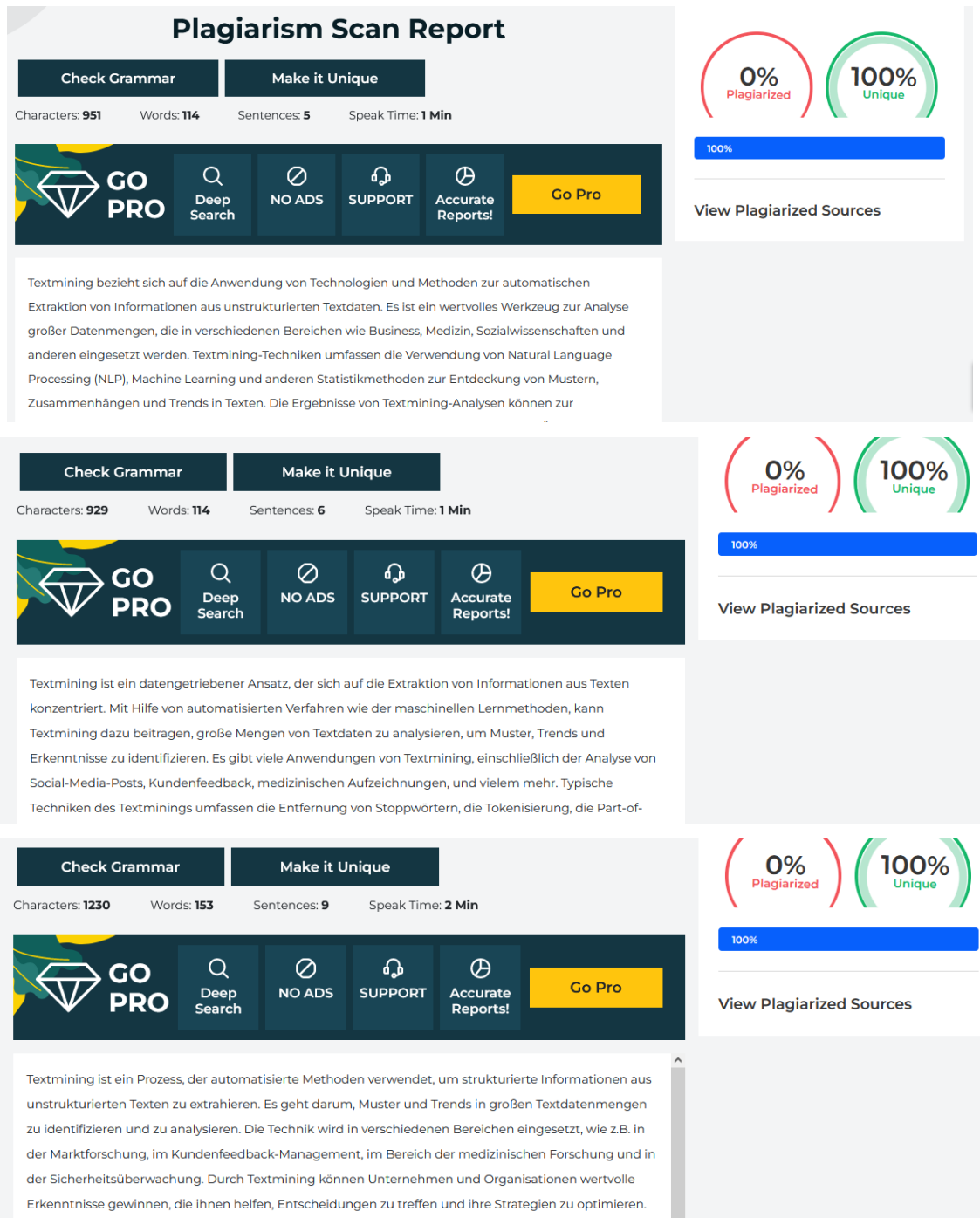


Bild 14: Ergebnisse des Plagiatsscanners von *plagiarismdetector.net*.

4.1.2 scribbr

Scribbr ist ebenfalls frei verfügbar, erfordert jedoch die Eingabe einer E-Mail-Adresse, an die die Scanergebnisse geschickt werden können. Diese können jedoch problemlos auch ohne Bestätigung der Adresse abgerufen werden, möchte man Scribbr nur einmalig nutzen. Scribbr arbeitet mit Turnitin, das momentan immer noch als zuverlässigster Scanner für schulische und

akademische Texte gilt, da es auch Zugriff auf akademische Datenbanken hat. Es spricht also Einiges dafür, dass die synthetischen Texte in dieser Form noch nicht verwendet wurden. Turnitin findet im Gegensatz zu anderen Scannern auch Ähnlichkeiten in stärker paraphrasierten Passagen und gibt die Quellen an, in denen die Ähnlichkeiten gefunden wurden. Die Software verfügt auch über einen KI-Detektor. Inwieweit dieser in der kostenlosen Version bei scribbr zum Einsatz kommt, ließ sich nicht herausfinden.



Text 1.docx

Niedriges Risiko eines Plagiats

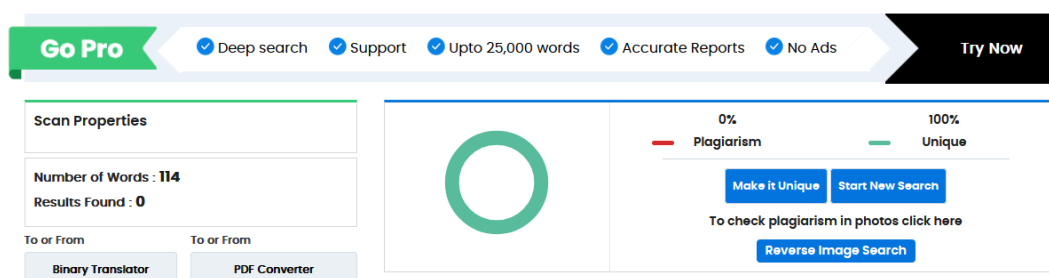
Wir haben keine erheblichen Ähnlichkeiten entdecken können. Dennoch kann dein Text Plagiate aus Quellen enthalten, die nicht in unserer Datenbank sind, wie z. B. aus eigenen alten Arbeiten, aus der Arbeit eines Bekannten oder einer Bekannten, oder aus Artikeln hinter einer Paywall.

Bild 15: Ergebnisse des Plagiatsscans mit scribbr.

Der Scanreport selbst weist auf die Grenzen seiner Aussagekräftigkeit hin. Etwas merkwürdig erscheint der Hinweis, dass der Text aus früheren eigenen Arbeiten sein könnte. Seine eigenen Texte für verschiedene Veröffentlichungen umzuschreiben, dürfte wohl kaum als Plagiat angesehen werden, egal welcher Plagiatsbegriff zugrunde gelegt wird.

4.1.3 duplichecker

Der Scanner von duplichecker.com versteht sich mehr oder weniger als eines von einer Reihe von SEO-Tools und sollte vermutlich eher dafür verwendet werden, die Inhalte von Websites auf versehentliche Plagiate zu prüfen. Auch dieses Programm findet an den Texten nichts zu beanstanden.



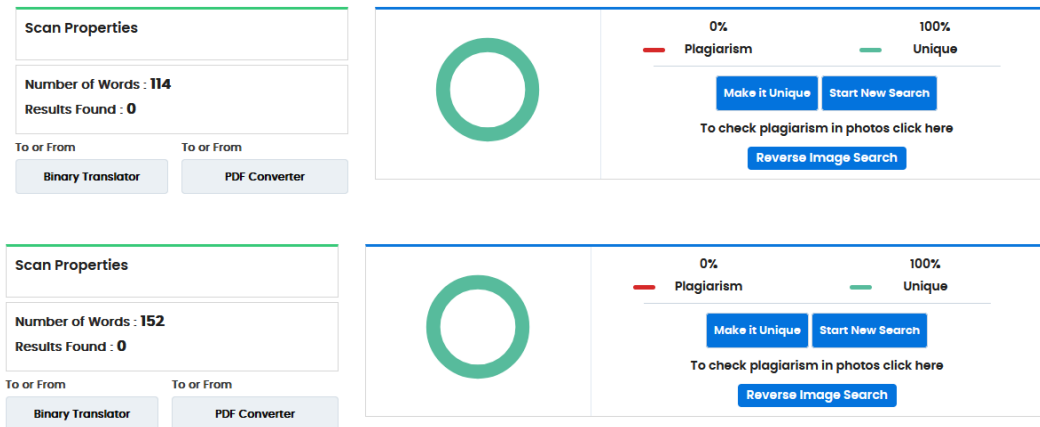


Bild 16: Ergebnisse des Scans mit *duplichecker.com*.

4.2 Synthetische Texte

ChatGPT selbst gibt folgende Antwort, wenn man ihn nach der Verwendung seiner eigenen Texte fragt.

AL Do you think it is plagiarism to hand in your text as a term paper?

If you were to hand in the text I provided as a term paper without proper citation or attribution, it would be considered plagiarism. Plagiarism is the act of presenting someone else's work, ideas, or words as your own without giving them credit. It is a serious academic offense that can result in disciplinary action, including failing the assignment, failing the course, or even expulsion from the institution.

To avoid plagiarism, it is essential to properly cite and attribute any sources that you use in your work, including text, images, and ideas. This means giving credit to the original author or creator of the work and providing a citation that allows others to locate the source. If you use my text in your term paper, be sure to provide proper citation and attribution to me as the source.

Bild 17: ChatGPTs Antwort auf die Frage, ob man seinen Text als Hausarbeit einreichen darf.

Da ChatGPT nicht über eine Anbindung an Datenbanken verfügt, kennt er auch nicht zwangsläufig den Autor der Texte. Bei bekannten Beispielen funktioniert dies jedoch.

AL: Wer ist der Autor des Textes "colorless green ideas sleep furiously"?


 Der Autor des Textes "Colorless green ideas sleep furiously" ist Noam Chomsky. Dieser Satz ist ein bekanntes Beispiel für eine grammatisch korrekte, aber sinnlose Aussage, die er 1957 als Beispiel für die Grenzen von Sprachmodellen und Grammatiktheorie verwendete.

Bild 18: ChatGPT erkennt den Autor bei populären Beispielen und Texten.

Startet man nicht jedes Mal eine neue Konversation, so wäre zu erwarten, dass ChatGPT seine eigenen Texte erkennt, die er wenige Turns¹⁸ zuvor selbst produziert hat. Dies klappt allerdings nicht (immer).

In this term paper, we will explore the RGB color cube and its applications in digital color representation. We will examine the basic principles of the RGB color model, including how it is used to create and manipulate colors. Additionally, we will explore the different techniques used to visualize the RGB color cube, including 2D and 3D representations, and we will examine how it is used to illustrate color gamuts and color spaces."

AL: Do you know who wrote this text?


 No, I do not have access to information on who wrote the text you provided. As an AI language model, my purpose is to provide helpful responses to your queries based on my programmed knowledge base.

Bild 19: ChatGPT erkennt seinen eigenen Text nicht als solchen. Die Ausführungen zum RGB-Farbwürfel hat er selbst wenige Turns zuvor produziert.

Da die KI-Detektoren ebenso wie die Plagiatsscanner meist nicht offenlegen, woran Plagiate bzw. synthetische Texte erkannt werden, kann man sich bei der Erzeugung von Testtexten nur auf das stützen, was anhand von Studien als Merkmale identifiziert werden konnte.

Guo et al. [16] haben verschiedene Merkmale gefunden, die authentische Texte von synthetischen Texten unterscheiden. Bei der Auswahl der Textsorte wurde darauf geachtet, dass diese Aspekte nach Möglichkeit nur eine geringe Rolle spielen.

¹⁸ Im linguistischen Sinne bezeichnet ein *Turn* einen einzelnen Gesprächsbeitrag in einer Konversation.

Die wichtigsten Merkmale von ChatGPT-Texten sind demnach:

- die Texte sind klar strukturiert und logisch
- die Texte sind tendenziell eher lang und detailliert
- die Texte zeigen weniger voreingenommene und schädliche Informationen
- ChatGPT antwortet nicht auf Fragen, deren Antwort außerhalb seiner Wissensbasis liegt
- die Texte können potenziell erfundene Informationen enthalten

Die Hauptunterschiede zu den menschlichen Antworten sind:

- ChatGPT ist auf die Frage fokussiert und beantwortet diese, ohne vom Thema abzukommen, was bei authentischen Antworten passieren kann
- ChatGPT antwortet objektiv, Menschen benutzen häufiger subjektive Ausdrücke
- ChatGPT antwortet formell, wenn nicht anders aufgefordert, Menschen tendieren zu einem höheren umgangssprachlichen Anteil
- ChatGPT drückt keine Emotionen aus, was Menschen i.d.R. tun, auch über Interpunktion und/oder grammatische Merkmale [16]

Da für die vorliegende Untersuchung eine bestimmte Textsorte (und damit auch ein bestimmter „Tonfall“) vorgegeben wurde, sind einige der Merkmale nicht einschlägig.

Gao et al. [14] ließen ChatGPT 50 Abstracts für spezifische medizinische Magazine verfassen, wobei das Format der jeweiligen Zeitschrift der KI vorgegeben wurde. Dies funktionierte nur bei 16% der synthetischen Texte. Dies deckt sich mit der Beobachtung von [13], obwohl hier kein authentischer Inhalt mit eingebunden werden musste.

Zur Überprüfung, ob es sich um synthetische oder authentische Texte handelt, wurden die Texte 1. einem AI-Output-Detektor, 2. einem Plagiatsprüfer und 3. Fachgutachtern zur Prüfung vorgelegt. Die Ergebnisse sind im Folgenden kurz zusammengefasst:

- Der AI-Output-Detektor erkannte 99,98% der synthetischen Texte als solche, lediglich 0,02% der authentischen Texte wurden falsch klassifiziert.
- Der Plagiatsprüfer stellte fast 100% der synthetischen Texte als authentisch fest, die „Plagiate“ wurden durchweg nicht erkannt.
- Die Gutachter erkannten 68% der synthetischen Texte als solche und hatten bei den authentischen Texten ebenfalls eine hohe Trefferquote von 86%, diese als authentisch zu erkennen.

Die absoluten Zahlen sind in Bild 20 wiedergegeben. Es wurden insgesamt 50 Abstracts erzeugt.

		Truth	
		Original	Generated
Reviewer guess	Original	43	16
	Generated	7	34

Bild 20: Authentische und synthetische Abstracts nach [14].

Zusammenfassend wurde festgestellt, dass es bestimmte Muster gibt, die synthetische Texte als solche erkennbar machen: „[...] there are pattern that allow it to be detected by AI output detection.“ [14] Diese Punkte betrafen u.a. die „ausgedachten“ Zahlen für die klinischen Studien und die relativ oberflächliche und vage Darstellung. Die erfundenen Probandenzahlen hingegen erschienen plausibel und auch der Fokus der Studien anhand der Details aus echten Abstracts gab zunächst keinen Anlass für Zweifel.

Die im Folgenden verwendeten KI-Detektoren wurden anhand von Vorschlägen aus der Literatur ausgewählt. Lediglich die letzten beiden sind zu aktuell, um in den Studien erwähnt zu werden. Diese liefern jedoch die besten Ergebnisse.

4.2.1 GLTR (Glitter)

GLTR (nach eigener Angabe gesprochen „Glitter“ und Abkürzung für Giant Language model Test Room) ist ein Open-Source-Projekt, das dementsprechend seine Funktionsweise und den Code offenlegt (vgl. github-

repository).¹⁹ GLTR „enables a forensic analysis of how likely an automatic system has generated a text.“ [15] GLTR versteht sich als visuelles forensisches Tool zur Detektion von synthetischem Text, das dieselbe Technologie zur Detektion benutzt, die auch zur Generierung der Texte benutzt wird. Es hat eine Anbindung an das GPT-2-Sprachmodell und entsprechend auch dessen „Wissensstand“. Es imitiert farblich codiert, wie GPT-2 seine Vorhersagen bzgl. der Fortsetzung einer Sequenz macht. Die Wahrscheinlichkeit ist in grün (unter den Top 10), gelb (unter den Top 100), rot (unter den Top 1.000) oder lila markiert. Der Anteil an roten und lilafarbenen Ausdrücken ist bei authentischen Texten signifikant höher als die beiden anderen Kategorien. Fährt man mit der Maus über die Ausdrücke, werden die jeweils fünf wahrscheinlichsten Folgeausdrücke angegeben. Darüber hinaus werden drei Histogramme erstellt. Das erste gibt die Anteile in den Wahrscheinlichkeitskategorien an, das zweite veranschaulicht das Verhältnis zwischen den Wahrscheinlichkeiten des am besten vorhergesagten Wortes und des folgenden Wortes, das dritte zeigt die Verteilung über die Entropien der Vorhersagen. Beispielanalysen werden von den Autoren zur Verfügung gestellt, sodass man so auch Vergleiche zu anderen Textsorten herstellen kann.

Für die untersuchten Abstracts ergibt sich folgendes Bild:

¹⁹ <https://github.com/HendrikStrobel/detecting-fake-text>. Man beachte, dass hier die Wertung Fake-Text schon im Namen steckt.

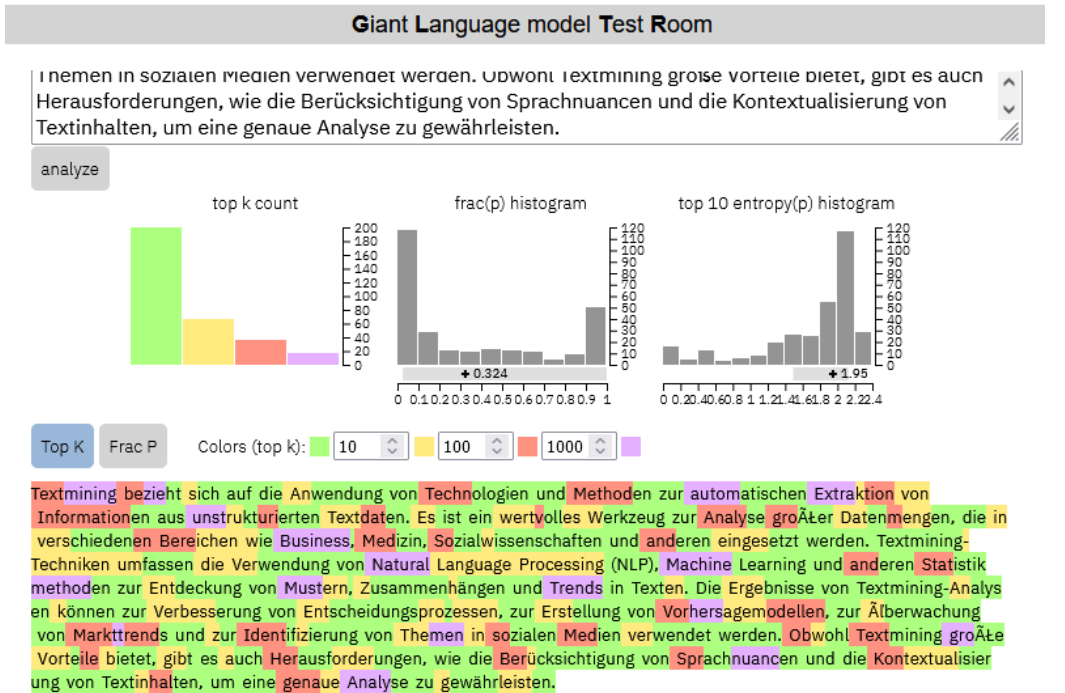


Bild 21: Ergebnisse von GLTR für Text 1.

Der Anteil an roten und lilafarbenen Ausdrücken spricht für einen authentischen Text, ebenso die hohe Unsicherheit in der Vorhersage. Dieser Text wäre demnach authentisch.

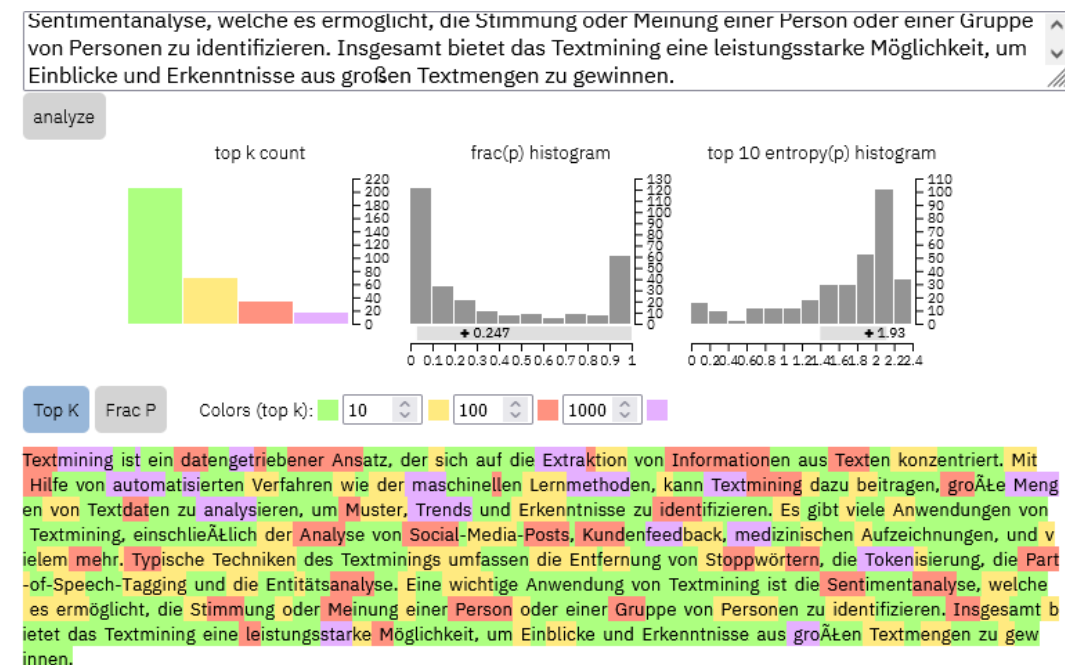


Bild 22: Ergebnisse von GLTR für Text 2.

Auch für Text 2 unterscheidet sich die Verteilung der Werte nicht signifikant von

der verfügbaren Auswertung eines authentischen Abstracts oder von Text 1. Auch diese Variante wird nicht als synthetisch erkannt.

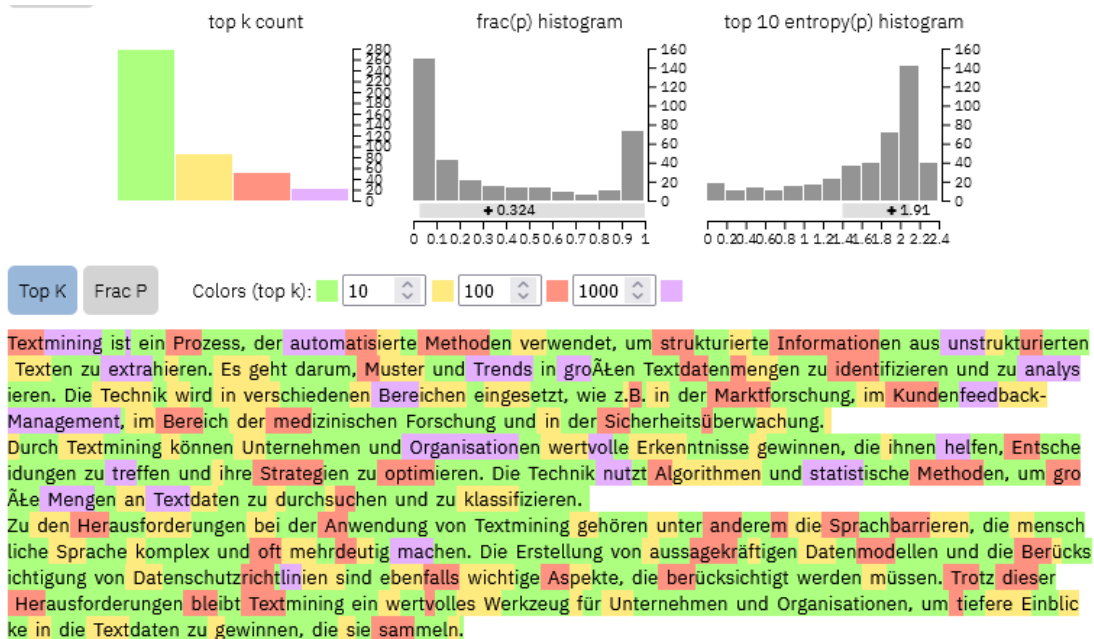


Bild 23: Ergebnisse von GLTR für Text 3.

Auch Text 3 wird nicht als synthetischer Text erkannt. Dies spricht für eine enorme Verbesserung von GPT-3 gegenüber GPT-2. Die Wertung erfolgt letztendlich durch den Nutzer. Das Programm selbst teilt keine Wertung mit.

4.2.2 OpenAI-Detector

Huggingface stellt ebenfalls den Code zur Verfügung²⁰, allerdings basiert auch dieser Detektor auf GPT-2. Die Ergebnisse der Analysen sind als OpenAI Report auf der Webseite verfügbar, ebenso wie eine Copy-and-Paste-Testversion, die hier genutzt wurde. Die Ergebnisse sind eindeutig.

²⁰ <https://huggingface.co/roberta-base-openai-detector>.

This is an online demo of the GPT-2 output detector model, based on the [🤗/Transformers](#) implementation of RoBERTa. Enter some text in the text box, the predicted probabilities will be displayed below. [The results start to get reliable after around 50 tokens.](#)

Textmining bezieht sich auf die Anwendung von Technologien und Methoden zur automatischen Extraktion von Informationen aus unstrukturierten Textdaten. Es ist ein wertvolles Werkzeug zur Analyse großer Datenmengen, die in verschiedenen Bereichen wie Business, Medizin, Sozialwissenschaften und anderen eingesetzt werden. Textmining-Techniken umfassen die Verwendung von Natural Language Processing (NLP), Machine Learning und anderen Statistikmethoden zur Entdeckung von Mustern, Zusammenhängen und Trends in Texten. Die Ergebnisse von Textmining-Analysen können zur Verbesserung von Entscheidungsprozessen, zur Erstellung von Vorhersagemodellen, zur Überwachung von Markttrends und zur Identifizierung von Themen in sozialen Medien verwendet werden. Obwohl Textmining große Vorteile bietet, gibt es auch Herausforderungen, wie die Berücksichtigung von Sprachnuancen und die Kontextualisierung von Textinhalten, um eine genaue Analyse zu gewährleisten.

Real
Prediction based on 322 tokens
Fake

99.98%0.02%

Textmining ist ein datengetriebener Ansatz, der sich auf die Extraktion von Informationen aus Texten konzentriert. Mit Hilfe von automatisierten Verfahren wie der maschinellen Lernmethoden, kann Textmining dazu beitragen, große Mengen von Textdaten zu analysieren, um Muster, Trends und Erkenntnisse zu identifizieren. Es gibt viele Anwendungen von Textmining, einschließlich der Analyse von Social-Media-Posts, Kundenfeedback, medizinischen Aufzeichnungen, und vielem mehr. Typische Techniken des Textminings umfassen die Entfernung von Stoppwörtern, die Tokenisierung, die Part-of-Speech-Tagging und die Entitätsanalyse. Eine wichtige Anwendung von Textmining ist die Sentimentanalyse, welche es ermöglicht, die Stimmung oder Meinung einer Person oder einer Gruppe von Personen zu identifizieren. Insgesamt bietet das Textmining eine leistungsstarke Möglichkeit, um Einblicke und Erkenntnisse aus großen Textmengen zu gewinnen.

Real
Prediction based on 326 tokens
Fake

99.97%0.03%

Textdatenmengen zu identifizieren und zu analysieren. Die Technik wird in verschiedenen Bereichen eingesetzt, wie z.B. in der Marktforschung, im Kundenfeedback-Management, im Bereich der medizinischen Forschung und in der Sicherheitsüberwachung.

Durch Textmining können Unternehmen und Organisationen wertvolle Erkenntnisse gewinnen, die ihnen helfen, Entscheidungen zu treffen und ihre Strategien zu optimieren. Die Technik nutzt Algorithmen und statistische Methoden, um große Mengen an Textdaten zu durchsuchen und zu klassifizieren.

Zu den Herausforderungen bei der Anwendung von Textmining gehören unter anderem die Sprachbarrieren, die menschliche Sprache komplex und oft mehrdeutig machen. Die Erstellung von aussagekräftigen Datenmodellen und die Berücksichtigung von Datenschutzrichtlinien sind ebenfalls wichtige Aspekte, die berücksichtigt werden müssen. Trotz dieser Herausforderungen bleibt Textmining ein wertvolles Werkzeug für Unternehmen und Organisationen, um tiefere Einblicke in die Textdaten zu gewinnen, die sie sammeln.

Real
Prediction based on 430 tokens
Fake

99.98%0.02%

Bild 24: Ergebnisse vom OpenAI-Detektor.

Allerdings geben die Autoren auf der Webseite zu bemerken: "The model is a

classifier that can be used to detect text generated by GPT-2 models. However, it is strongly suggested not to use it as a ChatGPT detector for the purposes of making grave allegations of academic misconduct against undergraduates and others, as this model might give inaccurate results in the case of ChatGPT-generated input.”, vgl. FN 20.

4.2.3 GPT-Zero

Writer stellt den KI-Detektor GPT-Zero für GPT-4 zur Verfügung. Weitere Informationen zur Funktion werden nicht zur Verfügung gestellt. Die Ergebnisse sind auch hier eindeutig.

WRITER

AI Content Detector

If your content reads like it was entirely produced by AI, it may affect how search engines rank it. Use our free detector to check up to 1,500 characters, and decide if you want to make adjustments before you publish.

AI content detection is **only available in the Writer app as an API**, and is limited to checks of 1,500 characters at a time. Find out more in our [help center article](#).

The screenshot shows the AI Content Detector interface. It includes a field for 'Add a URL' containing 'https://', a text area for 'Add some text' containing German text about NLP and machine learning, and a result bar showing '100% HUMAN-GENERATED CONTENT' with 'Fantastic!' below it.

The screenshot shows the AI Content Detector interface. It includes a field for 'Add some text' containing German text about text mining, and a result bar showing '100% HUMAN-GENERATED CONTENT' with 'Fantastic!' below it. A character count '929/1500 CHARACTERS' and an 'Analyze text' button are also visible.

Bild 25: Ergebnisse von GPT-Zero.

4.2.4 ZeroGPT

Trotz der hohen Verwechslungsgefahr aufgrund des Namens sind die folgenden Ergebnisse bemerkenswert. Laut Webseite liegt die Trefferquote bei 98%. Weitere Informationen zum Programm liegen nicht vor. Auch dieser Detektor ist frei verfügbar.

https://www.zerogpt.com

Your Text is AI/GPT Generated



Textmining bezieht sich auf die Anwendung von Technologien und Methoden zur automatischen Extraktion von Informationen aus unstrukturierten Textdaten. Es ist ein wertvolles Werkzeug zur Analyse großer Datenmengen, die in verschiedenen Bereichen wie Business, Medizin, Sozialwissenschaften und anderen eingesetzt werden. Textmining-Techniken umfassen die Verwendung von Natural Language Processing (NLP), Machine Learning und anderen Statistikmethoden zur Entdeckung von Mustern, Zusammenhängen und Trends in Texten. Die Ergebnisse von Textmining-Analysen können zur Verbesserung von Entscheidungsprozessen, zur Erstellung von Vorhersagemodellen, zur Überwachung von Markttrends und zur Identifizierung von Themen in sozialen Medien verwendet werden. Obwohl Textmining große Vorteile bietet, gibt es auch Herausforderungen, wie die Berücksichtigung von Sprachnuancen und die Kontextualisierung von Textinhalten, um eine genaue Analyse zu gewährleisten.

Highlighted text is suspected to be most likely generated by AI*

951 Characters
114 Words

Your Text is AI/GPT Generated



Textmining ist ein datengetriebener Ansatz, der sich auf die Extraktion von Informationen aus Texten konzentriert. Mit Hilfe von automatisierten Verfahren wie der maschinellen Lernmethoden, kann Textmining dazu beitragen, große Mengen von Textdaten zu analysieren, um Muster, Trends und Erkenntnisse zu identifizieren. Es gibt viele Anwendungen von Textmining, einschließlich der Analyse von Social-Media-Posts, Kundenfeedback, medizinischen Aufzeichnungen, und vielem mehr. Typische Techniken des Textminings umfassen die Entfernung von Stoppwörtern, die Tokenisierung, die Part-of-Speech-Tagging und die Entitätsanalyse. Eine wichtige Anwendung von Textmining ist die Sentimentanalyse, welche es ermöglicht, die Stimmung oder Meinung einer Person oder einer Gruppe von Personen zu identifizieren. Insgesamt bietet das Textmining eine leistungsstarke Möglichkeit, um Einblicke und Erkenntnisse aus großen Textmengen zu gewinnen.

Highlighted text is suspected to be most likely generated by AI*

929 Characters
114 Words

Your Text is AI/GPT Generated



Textmining ist ein Prozess, der automatisierte Methoden verwendet, um strukturierte Informationen aus unstrukturierten Texten zu extrahieren. Es geht darum, Muster und Trends in großen Textdatenmengen zu identifizieren und zu analysieren. Die Technik wird in verschiedenen Bereichen eingesetzt, wie z.B. in der Marktforschung, im Kundenfeedback-Management, im Bereich der medizinischen Forschung und in der Sicherheitsüberwachung. Durch Textmining können Unternehmen und Organisationen wertvolle Erkenntnisse gewinnen, die ihnen helfen, Entscheidungen zu treffen und ihre Strategien zu optimieren. Die Technik nutzt Algorithmen und statistische Methoden, um große Mengen an Textdaten zu durchsuchen und zu klassifizieren. Zu den Herausforderungen bei der Anwendung von Textmining gehören unter anderem die Sprachbarrieren, die menschliche Sprache komplex und oft mehrdeutig machen. Die Erstellung von aussagekräftigen Datenmodellen und die Berücksichtigung von Datenschutzrichtlinien sind ebenfalls wichtige Aspekte, die berücksichtigt werden müssen. Trotz dieser Herausforderungen bleibt Textmining ein wertvolles Werkzeug für Unternehmen und Organisationen, um tiefere Einblicke in die Textdaten zu gewinnen, die sie sammeln.

Highlighted text is suspected to be most likely generated by AI*

1230 Characters
152 Words

Bild 26: Ergebnisse von ZeroGPT.

ZeroGPT markiert den gesamten Text durchweg als KI-generiert. Zum Vergleich links der Wikipedia-Text über Textmining und rechts die Einleitung der Dissertation der Autorin der vorliegenden Arbeit.

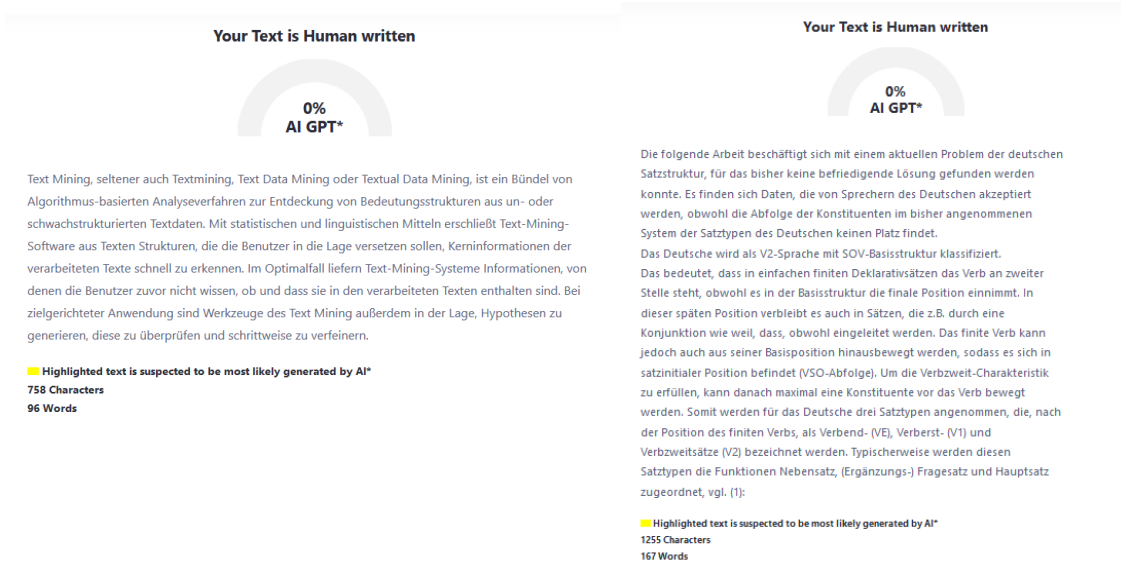


Bild 27: Vergleich mit authentischen Texten mit ZeroGPT.

4.2.5 Copyleaks

Der KI-Detektor von Copyleaks arbeitet mit GPT-4 und liefert ebenfalls zuverlässige Ergebnisse. Die Trefferquote wird auf der Webseite sogar mit 99% angegeben. Die Texte werden als 99,9% KI-generiert erkannt, bzw. die markierten Ausdrücke werden derart eingeordnet. Leider ist der Detektor aufgrund von vielen Anfragen nicht immer durchgängig erreichbar.

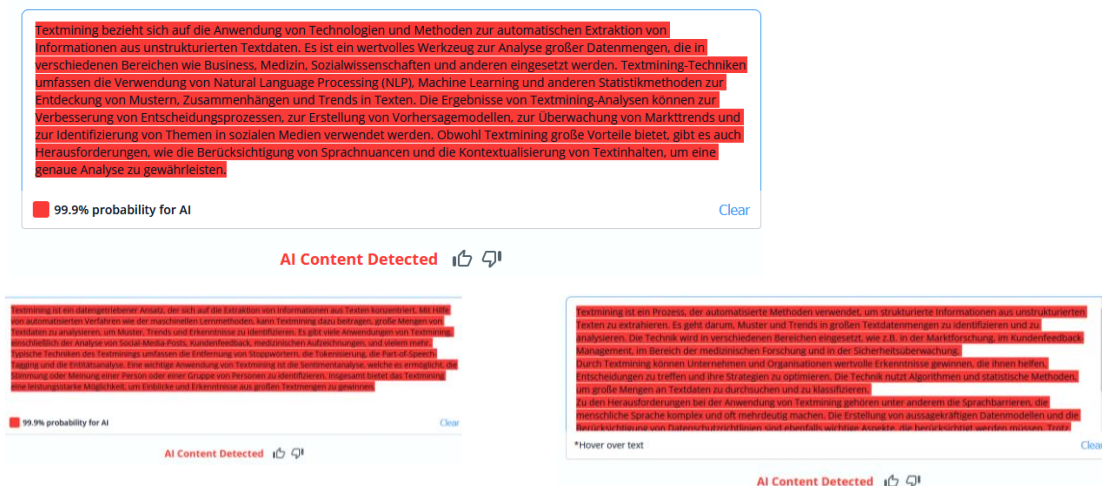


Bild 28: Ergebnisse von Copyleaks.

Zum Vergleich wieder die Einleitung der Dissertation, die zwar als authentischer Text identifiziert wird, jedoch mit geringerer Sicherheit als bei ZeroGPT.

dieser späten Position verbleibt es auch in Sätzen, die z.B. durch eine Konjunktion wie weil, dass, obwohl eingeleitet werden. Das finite Verb kann jedoch auch aus seiner Basisposition hinausbewegt werden, sodass es sich in satzinitialer Position befindet (VSO-Abfolge). Um die Verbzweit-Charakteristik zu erfüllen, kann danach maximal eine Konstituente vor das Verb bewegt werden. Somit werden für das Deutsche drei Satztypen angenommen, die, nach der Position des finiten Verbs, als Verbend- (VE), Verberst- (V1) und Verbzweitsätze (V2) bezeichnet werden. Typischerweise werden diesen Satztypen die Funktionen Nebensatz, (Ergänzungs-) Fragesatz und Hauptsatz zugeordnet, vgl. (1):

46.9% probability for Human

This is human text  

Bild 29: Authentischer Text bei *Copyleaks*.

5 Fazit und Ausblick

Die vorliegende Arbeit verfolgte zwei Ziele: 1. Es sollten drei Varianten eines synthetischen Textes von drei ausgewählten Plagiatsscanner geprüft werden, um zu testen, ob die Varianten bereits genutzt worden sind und somit erkannt werden; 2. Es sollten anhand der Textbeispiele fünf KI-Detektoren getestet werden, um zu prüfen, auf welchem Stand diese sind und ob sich die Ergebnisse älterer Detektoren, die für GPT-2 geschrieben wurde, von neueren unterscheiden. Dazu wurden in Kap. 2 zunächst die grundlegende Technologie, die Einsatzgebiete und die Kritik am Chatbot *ChatGPT* aufgezeigt. In Kap. 3 wurden statistische und linguistische Analysen von drei Textvarianten zu einem Abstract mit dem Thema *Textmining* durchgeführt, die gezeigt haben, dass sich die Varianten nicht sehr voneinander unterscheiden. Variiert wird meist die Wortart, wobei das Verhältnis von Inhaltswörtern zu Stoppwörtern nicht verändert wird. Hier wäre weitere Forschung vonnöten, da die Textbasis sehr überschaubar war und die Einzeltexte recht kurz sind. Um die Systematik hinter der Variantenbildung aufzuzeigen, müssten längere Texte und ein größeres Korpus untersucht werden. Dies erscheint auch deswegen sinnvoll, weil zu vermuten ist, dass die KI-Detektoren mit ähnlicher „Mustererkennung“ arbeiten, wobei sich meist nicht herausfinden lässt, anhand welcher Kriterien die Detektoren ihre Analyse durchführen. In Kap. 4 wurden zunächst Plagiatsscanner eingesetzt. Keiner erkennt die Varianten als Plagiate. Besonders auffällig ist, dass nicht einmal Turnitin auch nur geringe Zweifel an der Echtheit der Texte anmeldet. Von den getesteten KI-Detektoren waren drei aus der Literatur übernommen worden, da diese z.T. in den zitierten Studien verwendet wurden. Diese basierten auf GPT-2 und analysierten die Texte zu 100% als authentisch. Der Sprung zu GPT-3 lässt diese Detektoren damit wertlos werden. Lediglich zwei der neueren Generation, die bereits für GPT-4 entwickelt wurden, sind für den Einsatz tatsächlich geeignet. Aufgrund der besseren Erreichbarkeit und des besseren Ergebnisses beim Test mit einem authentischen Text ist eigentlich nur zeroGPT zu empfehlen.

Ausblicke auf weitere Entwicklungen kommen einem Blick in die Kristallkugel gleich. Selbst Vorhersagen bzgl. der Behandlung von (Anteilen von)

synthetischen Texten im wissenschaftlichen Bereich sind schwierig, da zunächst geklärt werden müsste, wie angesichts der momentanen Entwicklung Plagiate überhaupt zu definieren sind. Verschiedentlich ist angemerkt worden, dass durch den Einsatz von KI der Lerneffekt als solcher betroffen ist:

“Future developments notwithstanding, it is ironic that anti-plagiarism software uses artificial intelligence to assess the originality of assignments and that different AI (like ChatGPT) can be used to get around plagiarism detection software within seconds. The irony is complete when we realise that GPT-3 can write a review of the student’s AI-generated assignment on behalf of the teacher via a simple command: “Here is a short assessment of this student essay:” [...]. A first AI circumvents a second AI and is assessed by a third AI. All that the humans do is press a couple of keys, and nobody learns anything.” [33]

Scharfe Kritiker betonen ebenfalls, dass der KI Grenzen gesetzt sind, die verhindern, echte natürliche Intelligenz nachzubilden:

“However useful these programs may be in some narrow domains (they can be helpful in computer programming, for example, or in suggesting rhymes for light verse), we know from the science of linguistics and the philosophy of knowledge that they differ profoundly from how humans reason and use language. These differences place significant limitations on what these programs can do, encoding them with ineradicable defects.” [8]

Selbst Sam Altman selbst fordert Grenzen und gibt zu bedenken, dass ChatGPT nicht die eigene kritische Reflexion ersetzt, sondern der Nutzer selbst immer noch in der Verantwortung ist.

“ChatGPT is incredibly limited but good enough at some things to create a misleading impression of greatness. It’s a mistake to be relying on it for anything important but a preview of progress. We have lots of work to do on robustness and truthfulness.” [zitiert nach [1], wo sich auch eine lesenswerte Übersicht über Vor- und Nachteile für die Wissenschaft(ler) findet]

Beck fasst die Ergebnisse seiner Tests treffend zusammen:

„Nutzen Sie ein solches System bitte niemals zur Welterklärung, es sei denn, Sie können das Ergebnis gegenchecken.“ [2]

Dies bedeutet für den Einsatz von Plagiatsscanner besondere Vorsicht. Keiner der Scanner bietet 100%ige Verlässlichkeit, auf deren Grundlage man einen Plagiatsverdacht ausschließlich begründen könnte. Fast ebenso schnell wie die Entwicklung von Plagiatsscanner voranschreitet, werden Videos gedreht und Tipps weitergegeben, wie man die Scanner umgehen kann. Aus Sicht von Lehrenden weiß man nicht, wie man Posts begegnen soll, die davon berichten, Lernende hätten ihren Schreibstil so an *ChatGPT* angepasst, dass nun auch ihr authentischer Text von Scannern als synthetisch beanstandet wird.

“The related [...] task of asking human raters to judge the quality of machine-generated excerpts remains the gold-standard for evaluating open-domain generation systems [...]” [19]

Man muss sich stets der Zweischneidigkeit der Entwicklung bewusst sein. Die von OpenAI-finanzierte Studie [7] weist auf „potentially harmful applications“ hin:

“It [GPT-3, JW] therefore has the potential to advance both the beneficial and harmful applications of language models.”

Gehrmann et al. [15] fordern:

“Forensic techniques will be necessary to detect this automatically generated text. These techniques should be accurate, but also easy to convey to non-experts and require little setup cost.” [15]

Wenn man sich klarmacht, dass es hier um die Anwendung von GLTR geht, das schon mit GPT-3-Texten seine Probleme hat, ist zu erahnen, wie schwierig weitere Vorhersagen bzgl. der Entwicklung des Umgangs mit *ChatGPT* in der Wissenschaft und der Lehre sind.

Eines scheint jedoch sicher: Das Wettrennen zwischen KI und KI-„Jägern“ wird weitergehen.

6 Literaturverzeichnis

- [1] Alshater, Muneer: Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4312358, 02.06.2023.
- [2] Beck, Henning: ChatGPT im Test: Wer ist schlauer - künstliche Intelligenz oder das menschliche Gehirn?
<https://www.gmx.net/magazine/wissen/wissenschaft-technik/chatgpt-stoesst-grenzen-zeigen-beispiele-38000306>, 01.06.2023.
- [3] Biddle, Sam: The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. ChatGPT, the latest novelty from OpenAI, replicates the ugliest war on terror-style racism.
<https://www.seattletimes.com/business/technology/silicon-valley-investors-to-bankroll-artificial-intelligence-center/>, 01.06.2023.
- [4] Bisk, Yonatan; Holtzman, Ari; Thomason, Jesse; Andreas, Jacob; Bengio, Yoshua; Chai, Joyce; Lapata, Mirella; Lazaridou, Angeliki; May, Jonathan; Nisnevich, Aleksandr; Pinto, Nicolas, Turian, Joseph: Experience Grounds Language.
<https://arxiv.org/abs/2004.10151>, 01.06.2023.
- [5] Borij, Ali: A Categorical Archive of ChatGPT Failures.
<https://arxiv.org/abs/2302.03494>, 01.06.2023.
- [6] Bowman, Samuel R.; Angeli, Gabor Angeli; Potts, Christopher; Manning, Christopher D. Manning: A large annotated corpus for learning natural language inference.
<https://arxiv.org/abs/1508.05326>, 01.06.2023.
- [7] Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prfulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, Daniel M.; Wu, Jeffrey; Winter, Clemens; Hesse, Christopher; Chen, Mark; Sigler, Eric; Litwin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya; Amodei, Dario: Language Models are Few-Shot Learners.
<https://arxiv.org/abs/2005.14165>, 01.06.2023.
- [8] Chomsky, Noam; Roberts, Ian; Watumull, Jeffrey: The False Promise of ChatGPT. In The New York Times (08.03.2023), o.S.
- [9] Cooper, Kindra: OpenAI GPT-3: Everthing You Need to Know.
<https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/>, 01.06.2023.
- [10] Copyleaks: AI Content Detector.
<https://copyleaks.com/ai-content-detector>, 01.06.2023.
- [11] Data Science Stack Exchange: ChatGPT's Architecture – Decoder Only? Or Encoder-Decoder?
<https://datascience.stackexchange.com/questions/118260/chatgpts-architecture-decoder-only-or-encoder-decoder>, 01.06.2023.
- [12] Dehouche, Nassim: Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3).
<https://www.int-res.com/abstracts/esep/v21/p17-23/>, 01.06.2023.
- [13] Fyfe, Paul: How to cheat on your final paper: Assigning AI for student writing.
<https://link.springer.com/article/10.1007/s00146-022-01397-z>, 01.06.2023.

- [14] Gao, Catherine; Howard, Frederick; Markov, Nikolay; Dyer, Emma; Ramesh, Siddhi; Luo, Yuan; Pearson, Alexander: Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. <https://europepmc.org/article/ppr/ppr589378>, 01.06.2023.
- [15] Gehrman, Sebastian; Strobelt, Hendrik; Rush, Alexander M.: GLTR: Statistical Detection and Visualization of Generated Text. <https://arxiv.org/pdf/1906.04043.pdf>, 01.06.2023.
- [16] Guo, Biyang; Zhang, Xin; Wang, Ziyuan; Jiang, Minqi; Nie, Jinran; Ding, Yuxuan; Yue, Jianwei; Wu, Yupeng: How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. <https://arxiv.org/abs/2301.07597>, 01.06.2023.
- [17] Hendy, Amr; Abdelrehim, Mohamed; Sharaf, Amr; Raunak, Vikas; Gabr, Mohamed; Matsushita, Hitokazu; Kim, Young Jin; Afify, Mohamed; Hassan Awadalla, Hany: How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. <https://arxiv.org/abs/2302.09210>, 01.06.2023.
- [18] Hugendick, David: "Total gedankenlos und egoistisch": Zeitgeist oder Plagiat? In ihrem bejubelten Romandebüt hat sich Helene Hegemann bei einem anderen Autor bedient. <https://www.zeit.de/kultur/literatur/2010-02/hegemann-blogger-plagiat>, 01.06.2023.
- [19] Ippolito, Daphne; Duckworth, Daniel; Callison-Burch, Chris; Eck, Douglas: Automatic Detection of Generated Text is Easiest when Humans are Fooled. <https://arxiv.org/abs/1911.00650>, 01.06.2023.
- [20] Iriondo, Roberto: Unraveling the Magic of Generative AI: The Ultimate FAQ Extravaganza! <https://towardsai.net/p/opinion/unraveling-the-magic-of-generative-ai-the-ultimate-faq-extravaganza>, 01.06.2023.
- [21] Jiao, Wenxiang; Wang, Wenxuan; Huang, Jen-tse; Wang, Xing; Tu, Zhaopeng: Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. <https://arxiv.org/abs/2301.08745>, 01.06.2023.
- [22] Klerwin, Alexandra: ChatGPT vs. Conversational Maxims. Natural Conversation and Pragmatics. <https://medium.com/@a.klerwin/chatgpt-vs-conversational-maxims-a7c997065b72>, 01.06.2023.
- [23] Lin, Stephanie; Hilton, Jacob; Evans, Owain: TruthfulQA: Measuring How Models Mimic Human Falsehood. <https://arxiv.org/abs/2109.07958>, 01.06.2023.
- [24] Lomas, Natasha: ChatGPT shrugged. <https://techcrunch.com/2022/12/05/chatgpt-shrugged/>, 01.06.2023.
- [25] Markoff, John: Silicon Valley investors to bankroll artificial-intelligence center. <https://www.seattletimes.com/business/technology/silicon-valley-investors-to-bankroll-artificial-intelligence-center/>, 01.06.2023.
- [26] Moritz, Rainer; Meyer, Frank: Ein Rückblick auf „Axolotl Roadkill“: „Der Plagiatsvorwurf gegen Helene Hegemann war übertrieben“. <https://www.deutschlandfunkkultur.de/ein-rueckblick-auf-axolotl-roadkill-der-plagiatsvorwurf-100.html>, 01.06.2023.
- [27] OpenAI: ChatGPT. <https://chat-gpt.org/>, 05.06.2023.

- [28] Ortega-Martín, Miguel; García-Sierra, Óscar; Ardoiz, Alfonso; Álvarez, Jorge; Armenteros, Juan Carlos; Alonso, Adrián Alonso: Linguistic ambiguity analysis in ChatGPT. <https://arxiv.org/abs/2302.06426>, 01.06.2023.
- [29] Ouyang, Long; Wu, Jeff; Jiang, Xu; Almeida, Diogo; Wainwright, Carroll L.; Mishkin, Pamela; Zhang, Chong; Agarwal, Sandhini; Slama, Katarina; Ray, Alex; Schulman, John; Hilton, Jacob; Kelton, Fraser; Miller, Luke; Simens, Maddie; Askell, Amanda; Welinder, Peter; Christiano, Paul; Leike, Jan; Loew, Ryan: Training language models to follow instructions with human feedback. <https://arxiv.org/abs/2203.02155>, 01.06.2023.
- [30] Piantadosi, Steven T. [@spiantado]: Tweets. <https://twitter.com/spiantado/status/1599462405225881600>, 01.06.2023.
- [31] Rieck, Christian: Schummeln mit ChatGPT: Texte verfassen mit künstlicher Intelligenz für Schule, Uni und Beruf. München: Yes, 2023.
- [32] Ruby, Molly: How ChatGPT Works: The Model Behind The Bot. A brief introduction to the intuition and methodology behind the chat bot you can't stop hearing about. <https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286>, 01.06.2023.
- [33] Rudolph, Jürgen; Tan, Samson; Tan, Shannon: ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? In: Journal of Applied Learning & Teaching, 6 (2023) 1, S. 1 – 22.
- [34] Sociología Contemporánea: ChatGPT is High-Tech Plagiarism: Interview with Noam Chomsky. <https://www.youtube.com/watch?v=SJi4VE-0MoA>, 01.06.2023.
- [35] Stäbener, Jana: "Superspreader" von Fake News: ChatGPT textet in 80 von 100 Fällen falsche Informationen. <https://www.buzzfeed.de/news/chatgpt-chatbot-fake-news-newsguard-ki-misinformation-verschwörungstheorie-falsch-irrefuehrend-92045584.html>, 01.06.2023.
- [36] Susnjak, Teo: ChatGPT: The End of Online Exam Integrity? <https://arxiv.org/abs/2212.09292>, 01.06.2023.
- [37] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia. Attention is all you need. <https://arxiv.org/abs/1706.03762>, 01.06.2023.
- [38] veekaybee: Everything I understand about chatgpt. <https://gist.github.com/veekaybee/6f8885e9906aa9c5408ebe5c7e870698>, 01.06.2023.
- [39] Wang, Alex; Pruksachatkun, Yada; Nangia, Nikita; Singh, Amanpreet; Michael, Julian; Hill, Felix; Levy, Omer; Bowman, Samuel R.: SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. <https://arxiv.org/abs/1905.00537>, 01.06.2023.
- [40] Wang, Fei-Yue; Miao, Qinghai; Li, Xuan, Wang, Xingxia; Lin, Yilun: What Does ChatGPT Say: The DAO from Algorithmic Intelligence to Linguistic Intelligence. In: IEEE/CAA Journal of Automatica Sinica, 10 (2023) 3, S. 575 – 579.
- [41] Zhou, Ce; Li, Qian; Li, Chen; Yu, Jun; Liu, Yixin; Wang, Guangjing; Zhang, Kai; Ji, Cheng; Yan, Qiben; He, Lifang; Peng, Hao; Li, Jianxin; Wu, Jia; Liu, Ziwei; Xie, Pengtao; Xiong, Caiming; Pei, Jian; Yu, Philip S.; Sun, Lichao: A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. <https://arxiv.org/abs/2302.09419>, 01.06.2023.

7 Bilderverzeichnis

Bild 1: Screenshot ChatGPT identifiziert eine Textsorte. Unten: Eingabe für Prompts und Versionsnummer.	6
Bild 2: Abhängigkeit der Halluzinationen von der verwendeten Sprache (Stand 30.03.23).	9
Bild 3: Hinweise zu Beschränkungen von ChatGPT auf der Startseite.	10
Bild 4: Beschränktes Wissen für Ereignisse nach 2021 und Bsp. einer Halluzination, die in sich inkohärent ist.	10
Bild 5: Referenzen auf eine vorangegangene Äußerung (Danke an Dr. Eli David für das Teilen dieses Beispiels).	11
Bild 6: Zwei grundlegende Lerntechniken, vgl. [32].	13
Bild 7: Eingesetzte Lernmodelle in Anlehnung an [29] mit Kommentaren von [32]. (API: Application Programming Interface, GPT: Generative Pre-Training Transformer, SFT: Supervised Fine-Tuning, PPO: Proximal Policy Optimization).	14
Bild 8: Text 1 zum Thema Textmining.	20
Bild 9: Wortarten und Informationsgehalt in Text 1.	21
Bild 10: Text 2 zum Thema Textmining.	22
Bild 11: Wortarten und Informationsgehalt Text 2.	22
Bild 12: Text 3 zum Thema Textmining.	23
Bild 13: Wortarten und Informationsgehalt Text 3.	23
Bild 14: Ergebnisse des Plagiatsscanners von <i>plagiarismdetector.net</i>	28
Bild 15: Ergebnisse des Plagiatsscans mit scribbr.	29
Bild 16: Ergebnisse des Scans mit <i>duplichecker.com</i>	30
Bild 17: ChatGPTs Antwort auf die Frage, ob man seinen Text als Hausarbeit einreichen darf.	30
Bild 18: ChatGPT erkennt den Autor bei populären Beispielen und Texten. ...	31
Bild 19: ChatGPT erkennt seinen eigenen Text nicht als solchen. Die Ausführungen zum RGB-Farbwürfel hat er selbst wenige Turns zuvor produziert.	31
Bild 20: Authentische und synthetische Abstracts nach [14].	33
Bild 21: Ergebnisse von GLTR für Text 1.	35

Bild 22: Ergebnisse von GLTR für Text 2.	35
Bild 23: Ergebnisse von GLTR für Text 3.	36
Bild 24: Ergebnisse vom OpenAI-Detektor.	37
Bild 25: Ergebnisse von GPT-Zero.	38
Bild 26: Ergebnisse von <i>ZeroGPT</i>	39
Bild 27: Vergleich mit authentischen Texten mit <i>ZeroGPT</i>	40
Bild 28: Ergebnisse von <i>Copyleaks</i>	40
Bild 29: Authentischer Text bei <i>Copyleaks</i>	41

8 Tabellenverzeichnis

Tabelle 1: Anzahl der Ausdrücke aus Themenclustern in den drei Varianten.....	24
--	----

9 Anhang

Text1:

Textmining bezieht sich auf die Anwendung von Technologien und Methoden zur automatischen Extraktion von Informationen aus unstrukturierten Textdaten. Es ist ein wertvolles Werkzeug zur Analyse großer Datenmengen, die in verschiedenen Bereichen wie Business, Medizin, Sozialwissenschaften und anderen eingesetzt werden. Textmining-Techniken umfassen die Verwendung von Natural Language Processing (NLP), Machine Learning und anderen Statistikmethoden zur Entdeckung von Mustern, Zusammenhängen und Trends in Texten. Die Ergebnisse von Textmining-Analysen können zur Verbesserung von Entscheidungsprozessen, zur Erstellung von Vorhersagemodellen, zur Überwachung von Markttrends und zur Identifizierung von Themen in sozialen Medien verwendet werden. Obwohl Textmining große Vorteile bietet, gibt es auch Herausforderungen, wie die Berücksichtigung von Sprachnuancen und die Kontextualisierung von Textinhalten, um eine genaue Analyse zu gewährleisten.

- 113 Wörter, 838 Zeichen, 951 Zeichen mit Leerzeichen
- abzüglich Stoppwörtern 57 Wörter, Anteil echter informationstragender Wörter 50,44%
- Anzahl Tokens ohne Doppelung in der Grundform, Zahlen in Klammern geben Häufigkeiten > 1 an

Nomen (44)	Verben (4)	Adjektive (6)
Analyse (2)	beziehen	automatisch
Anwendung	gewährleisten	groß (2)
Bereich	umfassen	sozial
Berücksichtigung	verwenden	unstrukturiert
Business		verschieden
Datenmenge		wertvoll
Entdeckung		
Entscheidungsprozess		
Ergebnis		
Erstellung		
Extraktion		

Herausforderung		
Identifizierung		
Information		
Kontextualisierung		
Language		
Learning		
Machine		
Markttrends		
Medien		
Medizin		
Methode		
Muster		
Natural		
Processing		
Sozialwissenschaft		
Statistikmethode		
Sprachnuance		
Technologie		
Text		
Textdaten		
Textinhalt		
Textmining (2)		
Textmining-Analyse		
Textmining-Technik		
Thema		
Trend		
Überwachung		
Verbesserung		
Verwendung		
Vorhersagemodell		
Vorteil		
Werkzeug		
Zusammenhang		

Text 2:

Textmining ist ein datengetriebener Ansatz, der sich auf die Extraktion von Informationen aus Texten konzentriert. Mit Hilfe von automatisierten Verfahren wie der maschinellen Lernmethoden, kann Textmining dazu beitragen, große

Mengen von Textdaten zu analysieren, um Muster, Trends und Erkenntnisse zu identifizieren. Es gibt viele Anwendungen von Textmining, einschließlich der Analyse von Social-Media-Posts, Kundenfeedback, medizinischen Aufzeichnungen, und vielem mehr. Typische Techniken des Textminings umfassen die Entfernung von Stoppwörtern, die Tokenisierung, die Part-of-Speech-Tagging und die Entitätsanalyse. Eine wichtige Anwendung von Textmining ist die Sentimentanalyse, welche es ermöglicht, die Stimmung oder Meinung einer Person oder einer Gruppe von Personen zu identifizieren. Insgesamt bietet das Textmining eine leistungsstarke Möglichkeit, um Einblicke und Erkenntnisse aus großen Textmengen zu gewinnen.

- 114 Wörter, 816 Zeichen, 929 mit Leerzeichen
- abzüglich Stoppwörtern 56 Wörter, Anteil echter informationstragender Wörter 49,12%

Nomen (32)	Verben (6)	Adjektive (7)	Präpositionen
Analyse	analysieren	automatisiert	einschließlich
Ansatz	ermöglichen	datengetrieben	
Anwendung (2)	gewinnen	groß (2)	
Aufzeichnung	identifizieren (2)	leistungsstark	
Einblick	konzentrieren	maschinell	
Entfernung	umfassen	medizinisch	
Entitätsanalyse		typisch	
Erkenntnis (2)			
Extraktion			
Gruppe			
Hilfe			
Information			
Kundenfeedback			
Lernmethode			
Meinung			
Menge			
Möglichkeit			
Muster			
Part-of-Speech-Tagging			
Person (2)			
Sentimentanalyse			
Social-Media-Posts			

Stimmung			
Stoppwort			
Technik			
Text			
Textdaten			
Textmining (6)			
Textmenge			
Tokenisierung			
Trend			
Verfahren			

Text 3:

Textmining ist ein Prozess, der automatisierte Methoden verwendet, um strukturierte Informationen aus unstrukturierten Texten zu extrahieren. Es geht darum, Muster und Trends in großen Textdatenmengen zu identifizieren und zu analysieren. Die Technik wird in verschiedenen Bereichen eingesetzt, wie z.B. in der Marktforschung, im Kundenfeedback-Management, im Bereich der medizinischen Forschung und in der Sicherheitsüberwachung.

Durch Textmining können Unternehmen und Organisationen wertvolle Erkenntnisse gewinnen, die ihnen helfen, Entscheidungen zu treffen und ihre Strategien zu optimieren. Die Technik nutzt Algorithmen und statistische Methoden, um große Mengen an Textdaten zu durchsuchen und zu klassifizieren.

Zu den Herausforderungen bei der Anwendung von Textmining gehören unter anderem die Sprachbarrieren, die menschliche Sprache komplex und oft mehrdeutig machen. Die Erstellung von aussagekräftigen Datenmodellen und die Berücksichtigung von Datenschutzrichtlinien sind ebenfalls wichtige Aspekte, die berücksichtigt werden müssen. Trotz dieser Herausforderungen bleibt Textmining ein wertvolles Werkzeug für Unternehmen und Organisationen, um tiefere Einblicke in die Textdaten zu gewinnen, die sie sammeln.

- 152 Wörter, 1076 Zeichen, 1227 mit Leerzeichen
- abzüglich Stoppwörtern 72 Wörter, Anteil echter informationstragender Wörter 47,37%

Nomen (33)	Verben (14)	Adjektive (12)
Algorithmus	analysieren	automatisiert
Anwendung	berücksichtigen	aussagekräftig
Aspekt	durchsuchen	groß (2)
Bereich (2)	einsetzen	komplex
Berücksichtigung	extrahieren	medizinisch
Datenmodelle	gehören	mehrdeutig
Datenschutzrichtlinie	gewinnen (2)	menschlich
Einblick	helfen	statistisch
Entscheidung	identifizieren	strukturiert
Erkenntnis	klassifizieren	unstrukturiert
Erstellung	optimieren	verschieden
Forschung	sammeln	wertvoll (2)
Herausforderung (2)	treffen	
Information	verwenden	
Kundenfeedback-Management		
Marktforschung		
Menge		
Methode (2)		
Muster		
Organisation (2)		
Prozess		
Sicherheitsüberwachung		
Sprache		
Sprachbarriere		
Strategie		
Technik (2)		
Text		
Textdaten (2)		
Textdatenmenge		
Textmining (3)		
Trend		
Unternehmen (2)		
Werkzeug		

10 Verzeichnis der Abkürzungen

API	Application Programming Interface
bspw.	beispielsweise
bzgl.	bezüglich
FN	Fußnote
ggü.	gegenüber
GPT	generative pre-training transformer
i.d.R.	in der Regel
i.S.v.	im Sinne von
KI	künstliche Intelligenz
NLP	natural language processing
NMT	neural machine translation
o.S.	ohne Seitenangabe
PPO	proximal policy optimization
SEO	search engine optimization
SFT	supervised fine tuning
SOTA	state of the art
u.Ä.	und Ähnliche/r/s
u.U.	unter Umständen
vgl.	vergleiche
WMT	workshop in machine translation
z.B.	zum Beispiel
z.T.	zum Teil