Robustheit und Domänenanpassung bei der automatischen Anonymisierung von Gerichtsentscheidungen

Erkenntnisse anhand des Beispiels der automatischen Anonymisierung von Gerichtsentscheidungen von verschiedenen Instanzen und Rechtsgebieten

Der vorliegende Beitrag stellt das vom BMBF geförderte Projekt *AnGer* vor, das die Entwicklung eines vollautomatischen KI-Systems zur Anonymisierung von Gerichtsentscheidungen erforscht. Neben der Diskussion der generellen Machbarkeit werden die notwendigen Schritte zur praktischen Umsetzung des Vorhabens präsentiert. Im Fokus steht dabei die Robustheit des Systems, die wir als zweckgebundene Übertragbarkeit an neue (sprachliche) Domänen verstehen. Diese Perspektive schlägt die Brücke zwischen normativen Rechtsbegriffen der KI-Verordnung und der Computerlinguistik, in der das Problem als „Domänenanpassung“ bekannt ist, für welche wir Ergebnisse umfassender Experimente präsentieren.

Technischer Sachverhalt

Die automatische Anonymisierung von Gerichtsentscheidungen ist eine Anwendung der Computerlinguistik. Dafür werden insbesondere vortrainierte große Sprachmodelle benutzt, welche zunächst auf umfangreichen Textdaten trainiert und anschließend durch sog. Finetuning auf eine Aufgabestellung spezialisiert werden. Gemäß den Vorgaben der KI-Verordnung müssen solche KI-Systeme hohe Anforderungen an Genauigkeit und Robustheit erfüllen. Genauigkeit wird im Kontext der Anonymisierung insbesondere als Recall verstanden, also als der Anteil korrekt erkannter sensibler Textstellen an allen tatsächlich sensiblen Stellen. Robustheit bezieht sich im Sinne der Computerlinguistik auf die Übertragbarkeit der Leistungsfähigkeit auf sprachliche Domänen, die nicht mit der Trainingsdomäne übereinstimmen. Im konkreten Anwendungsfall definiert sich die sprachliche Domäne durch die Kombination aus gerichtlicher Instanz (bspw. Amtsgericht) und Rechtsgebiet (bspw. Verkehrsrecht).

I. Einleitung

Trotz der rechtlichen Verpflichtung, gerichtliche Entscheidungen der Öffentlichkeit zugänglich zu machen, werden in Deutschland derzeit lediglich zwei bis drei Prozent aller Urteile veröffentlicht.[1] Ein zentraler Grund dafür liegt in der aufwändigen und kostenintensiven – an den Gerichten manuell durchgeführten – Anonymisierung, die erforderlich ist, um sensible Informationen der betroffenen Parteien zu schützen. Angesichts dieser Herausforderungen zielt das Projekt *AnGer* (*Anonymisierung von Gerichtsentscheidungen für die digitale Justiz*) darauf ab, KI-Systeme zur vollautomatischen Anonymisierung zu entwickeln, die eine skalierbare und rechtssichere Veröffentlichung von Urteilen ermöglichen sollen.

Die Entwicklung solcher Systeme steht vor hohen Anforderungen: Als KI-Systeme Nach Erwg. 61 S. 3 KI-VO soll sich die Einstufung als hochrisikoreich zwar „nicht auf KI-Systeme erstrecken, die für rein begleitende Verwaltungstätigkeiten bestimmt sind, und die die tatsächliche Rechtspflege in Einzelfällen nicht beeinträchtigen, wie die Anonymisierung oder Pseudonymisierung gerichtlicher Urteile, Dokumente oder Daten, die Kommunikation zwischen dem Personal, Verwaltungsaufgaben

oder die Zuweisung von Ressourcen." Diese Erwägungen stehen allerdings im Widerspruch zum eindeutigen Wortlaut in Anhang III Nr. 8 lit. a KI-VO zur Bestimmung eines Hochrisiko-KI-Systems, worin ausdrücklich von „Justizbehörden" und der „Anwendung des Rechts auf konkrete Sachverhalte" die Rede ist und demnach nicht nur Gerichte und ihre Rechtsprechungstätigkeit erfasst sind.[2] Der Widerspruch ergibt sich durch einen Vergleich mit den anderen Aufzählungen im Anhang III; diese erfassen ebenso Behörden wie Strafverfolgungs-, Asylbehörden usw, dh die Argumentation, dass nur eine „ungefährliche" Justizverwaltungstätigkeit vorliegt, trägt nicht. Vielmehr ist vergleichbar, dass die Gefährdungslage für Grundrechtsverletzungen bei der Anonymisierung von Gerichtsentscheidungen gleichermaßen gegeben ist, wie bei den anderen genannten Verwaltungstätigkeiten. Entgegenstehende Erwägungsgründe sind zudem nicht wie entgegenstehende Gesetze zu behandeln, sondern funktionieren als eine argumentative Hürde,[3] welche unserer Ansicht nach für den Fall der Anonymisierung überwunden werden kann. Letztlich spricht Erwg. 61 auch seinem Wortlaut nach nicht von einem Zwang, welcher zudem überwunden werden könnte, sondern nur von einem Sollen. Daher antizipieren wir für den Anwendungsfall der automatischen Anonymisierung von Gerichtsurteilen sicherheitshalber, dass es sich um ein Hochrisiko-KI-System handelt.

Neben der Einhaltung von Vorgaben wie Genauigkeit, Robustheit und Cybersicherheit (Artikel 15 KI-VO) ist insbesondere die Zweckbestimmung der Systeme von entscheidender Bedeutung. Während wir zeigen können, dass eine vollautomatische Anonymisierung grundsätzlich mit hoher Genauigkeit realisierbar ist, bleibt eine vollständige Robustheit nur schwer zu garantieren. Insbesondere geht der Nachweis der Anpassungsfähigkeit an neue oder abweichende Daten – eine Herausforderung, die in der Computerlinguistik unter Robustheit subsumiert und dort auch als „Domänenanpassung" bezeichnet wird – mit technischen und begriffstheoretischen Schwierigkeiten einher.

Im vorliegenden Beitrag klären wir die Begriffe Robustheit und Domänenanpassung sowohl aus der Perspektive der Computerlinguistik als auch im Kontext der KI-VO. Basierend darauf präsentieren wir experimentelle Ergebnisse, die die Leistungsfähigkeit und Grenzen unserer KI-Systeme zur vollautomatischen Anonymisierung unter verschiedenen Bedingungen beleuchten. Unser Ziel ist es, eine fundierte Grundlage für die Weiterentwicklung solcher Systeme zu schaffen, die sowohl den rechtlichen Anforderungen als auch praktischen Bedürfnissen gerecht werden.

II. Projektbeschreibung

Das vom BMBF geförderte Forschungsprojekt *AnGer* erforscht seit 2023 in einem interdisziplinären Team aus Jurist:innen und Computerlinguist:innen an der Friedrich-Alexander-Universität Erlangen-Nürnberg das Potential eines automatischen Systems zur Anonymisierung von Gerichtsurteilen.[4] Ziel des Projekts ist es unter anderem, technisch und somit tatsächlich zu ermöglichen, dass „Gerichtsentscheidungen […] grundsätzlich in anonymisierter Form in einer Datenbank öffentlich und maschinenlesbar verfügbar sein [sollen]", wie im noch aktuellen Koalitionsvertrag[5] gefordert wird. Andererseits ist der Zugang zu einer Vielzahl von Urteilen Voraussetzung für die Entwicklung von weiteren juristischen KI-Anwendungen, die Urteile als Trainingsdaten für maschinelle Lernverfahren nutzen müssen.

Ein erster Prototyp für die vollautomatische Anonymisierung wurde bereits im Vorgängerprojekt *LeAK* (*Legal Anonymization Kit*; gefördert vom BayStMJ im Zeitraum 2020–2023) entwickelt. Dieser erzielt bei einer In-Domain-Evaluation 99% Recall[6] auf Hochrisikostellen (Namen, Adressen und weitere direkte Identifikatoren). Die generelle technische Möglichkeit einer vollautomatischen Anonymisierung wurde somit nachgewiesen.[7]

II.1. Domänen und Datenmaterial

Die in *LeAK* zur Verfügung gestellten Daten beschränkten sich auf Amtsgerichtsurteile in zwei Rechtsgebieten – Mietrecht und Verkehrsrecht – im Gesamtumfang von circa 1 Million Token[8], siehe Tabelle 1. Es zeigte sich, dass KI-Systeme zur automatischen Anonymisierung, welche nur auf einem Rechtsgebiet trainiert werden, einem auf beiden Rechtsgebieten trainierten System unterlegen sind.

Zur Erforschung der Domänenanpassungsfähigkeit des KI-Systems stehen im Projekt *AnGer* nun Entscheidungen von Oberlandesgerichten zur Verfügung. Während bei *LeAK* lediglich Urteile vorlagen, umfassen diese Entscheidungen zusätzlich Beschlüsse, Hinweisbeschlüsse und Verfügungen. Diese Entscheidungsarten unterscheiden sich gerade im textlichen Aufbau und der rechtlichen Komplexität von den Amtsgerichtsurteilen. Für die in Kapitel IV. vorgestellten Experimente gehen wir davon aus, dass die Kombination aus Rechtsgebiet und Instanz eine sprachliche Domäne darstellt. Die Notwendigkeit einer domänenspezifischen Anpassung ist eine der zentralen Hypothesen, die unser Forschungsprojekt untersucht. Tabelle 1 zeigt eine Übersicht über den vollständigen Datensatz aufgeteilt nach Instanz und Rechtsgebiet. Die Aufteilung nach Rechtsgebieten entspricht dabei in etwa auch der Aufteilung der Kammern der Oberlandesgerichte.

Domäne	Anz. Dokumente	Anz. Sätze	Anz. Token
Amtsgericht			
Mietrecht	247	24930	399822
Verkehrsrecht	323	33726	552906
Oberlandesgericht			
Allgemeine Zivilsachen	103	15113	384715
Bankensachen	86	8869	213321
Bausachen	92	13256	281604
Beschwerdeverfahren	48	4489	110884
Familiensachen	62	7333	156503
Handelssachen	217	40508	1015706
Immaterialgüter	102	19994	535408
Kostensachen	26	2007	45730
Schiedssachen	42	6084	141702
Verkehrsunfallsachen	98	9287	247597
Kapitalanlagesachen	87	13257	357505

Tabelle 1: Datenmaterial aufgeteilt nach Instanz und Rechtsgebiet.

II.2. Manuelle Anonymisierung: Erstellung des Goldstandards

Für die Entwicklung eines automatischen Systems zur Anonymisierung von Gerichtsurteilen sowie zur Evaluation jeglicher Anonymisierungsverfahren (manueller oder automatischer Natur) wird ein Goldstandard benötigt, dh ein Korpus von Urteilen, in dem die zu anonymisierenden Stellen korrekt markiert wurden. Hierfür beschäftigen wir studentische Hilfskräfte, deren Aufgabe es ist, unabhängig voneinander alle sensiblen Stellen zu identifizieren und zusätzlich den Grund für die Anonymisierung sowie das Risikoniveau der Stelle zu vermerken.

Grundlage der Arbeit der Annotator:innen bilden ausführliche Annotationsrichtlinien. Diese wurden zu Beginn des Projekts auf Basis von rechtsdogmatischen Grundlagen zur Anonymisierung, wie zB dem allgemeinen Persönlichkeitsrecht, der DSGVO, § 30 AO, GeschGehG, Unternehmenspersönlichkeitsrecht, §§ 203 f. StGB, § 35 SGB I, erstellt und werden fortgehend anhand von Rückmeldungen der Annotator:innen angepasst und verbessert.[9] In den Guidelines werden mehr als 20 Informationskategorien unterschieden: Namen natürlicher und juristischer

Personen oder Marken sowie Adressangaben, die überwiegend ein hohes Deanonymisierungsrisiko bergen und besonders zuverlässig anonymisiert werden müssen; Datumsangaben, die leicht automatisch identifiziert werden können; Fahrzeugnummern, die ein hohes Risiko für Deanonymisierung bergen; allgemeine identifizierende Merkmale und Aktenzeichen, die zwar oft für sich genommen unkritisch sind, aber zusammen mit anderen Informationen zur Deanonymisierung beitragen können und daher ein mittleres Risiko bergen; sonstige Informationen, die ein niedrigeres Potential haben, zur Deanonymisierung beizutragen. Nachdem der Datensatz um Gerichtsentscheidungen von Oberlandesgerichten aus mehr als zehn Rechtsgebieten erweitert wurde, mussten auch die Richtlinien, vor allem im Bereich der Immaterialgüter und der Abgrenzung von Literaturangaben, die keine Annotation erfordern, angepasst werden.

In einem ersten Schritt werden alle Urteile von mindestens vier Annotator:innen unabhängig voneinander annotiert. Untersuchungen haben gezeigt, dass vier bis sechs unabhängige Annotator:innen ausreichen, um ein Korpus von 1 Million Token nahezu perfekt zu anonymisieren.[10] Dabei ist die Übereinstimmung der Annotator:innen untereinander bezüglich den meisten Kategorien sehr hoch; lediglich bei den allgemeinen identifizierenden Merkmalen findet sich ein hohes Maß an Subjektivität. Etwaige Unstimmigkeiten, auch aufgrund von Flüchtigkeitsfehlern und Auslegungsspielräumen bei Grenzfällen, werden in einem nächsten Schritt von mindestens einer weiteren Hilfskraft aufgelöst. Diese entscheidet, welche Annotation den Richtlinien entspricht (ein Vorgang, der sich „Adjudikation“ nennt). Das Ergebnis ist ein konfliktfreies Annotationslayer, das alle zu anonymisierenden Stellen im jeweiligen Text enthält und das als „Goldstandard“ bezeichnet wird.

Zur Erstellung eines Datensatzes, mit dem auch außerhalb besonders geschützter Räume Experimente durchgeführt werden können, wird der fertige Goldstandard von den Hilfskräften pseudonymisiert, dh alle annotierten sensiblen Textstellen werden durch realistische Phantasienamen und andere frei erfundene Angaben ersetzt. Es wird dabei darauf geachtet, dass durch die Pseudonymisierung der Informationsgehalt der Textstellen und des Urteils nicht verändert wird. Beispielsweise werden Vornamen, Familiennamen und andere Eigennamen innerhalb eines Urteils immer durch die gleichen Pseudonyme ersetzt; Datumsangaben werden so verändert, dass der Tatbestand und Verfahrensablauf in sich konsistent bleiben. Auch dieser Schritt erfolgt auf Grundlage eigens dafür erstellter Pseudonymisierungsrichtlinien. Da die Zuordnung von Pseudonymen zu Klarnamen nach Abschluss der Bearbeitung unwiderruflich gelöscht wird, handelt es sich bei dem resultierenden Datensatz um vollständig anonymisierte Daten im Sinne der DSGVO. Der Goldstandard unterliegt damit nicht mehr dem Datenschutz und kann somit für korpuslinguistische Untersuchungen sowie als Trainingskorpus für die Entwicklung eines automatischen Systems auf Hochleistungs-Rechensystemen genutzt werden.

II.3. Annotationssoftware

Die Annotations- und Pseudonymisierungsarbeiten finden in einer eigens hierfür entwickelten Web-Applikation statt, welche den gesamten Prozess in einer einheitlichen Arbeitsumgebung abbildet. Neben der Speicherung der manuellen Eingaben zur Annotation und Pseudonymisierung unterstützt die Applikation die automatische heuristische Auflösung von Diskrepanzen zwischen Annotator:innen im Vorfeld der Adjudikation (so werden bspw. einfache Mehrheitsentscheidungen getroffen und Abweichungen zwischen Risikokategorien auf den Median aufgelöst) sowie die Erstellung von Vorschlägen für die Pseudonymisierung.

Für den produktiven Einsatz bei Gerichten wurde das Toolkit zu einem Software-Demonstrator weiterentwickelt, der eine semi-automatische Anonymisierung und Pseudonymisierung von Word- und PDF-Dateien unterstützt. Die Dokumente werden hierbei von einem angepassten Sprachmodell (siehe Abschnitt II.4.) automatisch annotiert und der bzw. die Anwender:in am Gericht übernimmt die

Rolle des bzw. der Adjudikator:in – kann also weitere Annotationen hinzufügen und bestehende modifizieren oder löschen. Im Anschluss wird das Dokument automatisch pseudonymisiert, wobei sich die Pseudonymisierung momentan für die meisten Kategorien auf eine Erstellung von (konsistent randomisierten) Abkürzungen beschränkt.

II.4. Automatische Anonymisierung: Finetuning von Sprachmodellen

Viele Tools zur automatischen Anonymisierung arbeiten meist nur semi-automatisch und bieten damit lediglich eine Unterstützung für den immer noch manuellen Anonymisierungsprozess. Diese Tools schlagen Textstellen vor, die anonymisiert werden sollten; menschliche Bearbeiter:innen müssen im Anschluss den gesamten Text durchlesen und die vorgeschlagenen Textstellen manuell überprüfen sowie weitere Textstellen identifizieren – oder aber sie verlassen sich in ihrer eigenen Verantwortung blind auf die Vorschläge des Tools. Im Gegensatz dazu soll eine vollautomatische Anonymisierung alle sensiblen Textstellen vollständig und ohne menschliche Eingabe identifizieren und kategorisieren. Nur eine vollautomatische Anonymisierung skaliert auf die Veröffentlichung von mehreren Millionen Gerichtsentscheidungen.

Zu den sensiblen Informationen, die von dem automatischen Tool erkannt werden müssen, gehören insbesondere Datums- und Ortsangaben sowie die Namen von natürlichen und juristischen Personen. Diese Aufgabe ist in der Computerlinguistik als „Named Entity Recognition“ (NER) bekannt. Es existieren bereits Standardwerkzeuge für NER. Allerdings können diese in unserem Anwendungsfall nicht ohne weitere Anpassungen und Trainingsschritte eingesetzt werden können, da sie nicht alle relevanten Arten sensibler Informationen erkennen (insbesondere keine allgemeinen identifizierenden Merkmale) und die genaue Definition der zu erkennenden Textstellen oft nicht unseren (juristisch hergeleiteten) Annotationsrichtlinien entspricht. Bei den meisten computerlinguistischen Anwendungen ist man mit einem weitaus niedrigeren Recall zufrieden als für eine automatische Anonymisierung erforderlich.

Im Laufe der letzten Jahre hat sich in der computerlinguistischen Forschung gezeigt, dass vortrainierte große Sprachmodelle sehr erfolgreich auf NER und ähnliche Aufgabenstellungen angewendet werden können. Diese Ansätze eignen sich vor allem gut für Fälle, in denen nur eine begrenzte Menge von Trainingsdaten zur Verfügung steht. Im Rahmen des Projekts führen wir dementsprechend Experimente mit vortrainierten Open-Source-Sprachmodellen durch und passen diese durch weiteres Training auf unserem anonymisierten und pseudonymisierten Goldstandard an die spezifische Aufgabenstellung an. In der Computerlinguistik bezeichnet man diese Vorgehensweise auch als „Finetuning“, bei der ein bereits trainiertes Grundmodell auf eine spezifische Domäne oder einen bestimmten Datensatz spezialisiert wird.

Für die Evaluation der angepassten Modelle ist zum einen der oben erwähnte Recall hilfreich – dh wie viele der zu anonymisierenden Textstellen vom Modell richtig erkannt wurden – zum anderen der positive Vorhersagewert (Precision) – dh welcher Anteil der vom Modell identifizierten Textstellen tatsächlich korrekt ist. Es ist anzumerken, dass die beiden Maßzahlen in wechselseitiger Abhängigkeit stehen. Das bedeutet, dass bei einer Verbesserung des Recall eine schlechtere Precision in Kauf genommen werden muss. Für die Auswertung der automatischen Anonymisierung ist primär der Recall ausschlaggebend, da bereits eine einzige übersehene Textstelle (zB ein Personenname, eine Wohnanschrift oder ein Kfz-Kennzeichen) für eine Deanonymisierung ausreichen kann – wohingegen zu viele anonymisierte Stellen höchstens zur Unverständlichkeit des Textes führen. Wir konzentrieren uns bei der Darstellung in Kapitel IV. daher insbesondere auf den Recall der entwickelten Systeme.

III. Robustheit in Computerlinguistik und KI-VO

Welche Anforderungen muss ein KI-System erfüllen, um als robust zu gelten? Wir diskutieren im vorliegenden Kapitel zunächst, was Robustheit in der Computerlinguistik bedeutet und argumentieren anschließend, dass die KI-VO in der Risikoanalyse nicht von einem – wie in der Computerlinguistik üblich – statistisch geprägten Robustheitsbegriff ausgeht. Es ist dahingehend zielführender, die computerlinguistische Robustheit von KI-Systemen im Kontext der Domänenanpassung (engl. Domain Adaptation) zu betrachten, welche ein Problem der Zweckbestimmung darstellt.

III.1. Robustheit in der Computerlinguistik

In der Computerlinguistik wird ein KI-Modell in der Regel für eine spezifische Domäne trainiert; das bedeutet, dass sowohl Trainings- als auch Testdaten aus derselben Domäne stammen. Die momentan erfolgreichsten computerlinguistische Ansätze – große Sprachmodelle (engl. Large Language Model) – basieren auf tiefen künstlichen neuronalen Netzen, die eine große Menge an hochwertigen Trainingsdaten benötigen, um gute und zuverlässige Ergebnisse zu erzielen. Solche Trainingsdaten stehen allerdings nicht immer im erforderlichen Umfang zur Verfügung, sodass sich die Sprachmodelle aufgrund ihrer Komplexität zu sehr an die Trainingsdaten anpassen (sog. Überanpassung). Eine Überanpassung des Sprachmodells auf eine Domäne entsteht nun nicht ausschließlich durch Datenmangel, sondern auch durch mangelnde Datenqualität: KI-Systeme sind anfällig gegenüber Verzerrungen (engl. Bias) sowie Scheinkorrelationen.[11] Speziell bei NER-Anwendungen kann die sogenannte Namensregelmäßigkeit zu Überanpassung führen, was die Generalisierung auf andere Domänen, bei denen Namen oder Datumsangaben in einem anderen Format vorkommen, erschweren kann.[12]

Da viele Modelle anfällig für Veränderungen in Anwendungsdaten sind, selbst wenn sie bei domänenspezifischen Testdaten gute Ergebnisse erzielen, kann in der Computerlinguistik erst dann von einem „robusten“ automatischen Verfahren die Rede sein, wenn das trainierte System bei unbekannten Daten, die sich signifikant von den ursprünglichen Trainingsdaten unterscheiden (zB von einer anderen Instanz oder aus einem anderen Rechtsgebiet stammen), konsistent verlässliche Ergebnisse erzielt. Hier spricht man in der Forschung auch von Domänengeneralisierung.[13] Generell dürfen die Ausgaben eines robusten KI-Systems nicht von Transformationen der Eingabedaten beeinträchtigt werden.[14]

Außerdem sollen solche Modelle ebenfalls bei anderen „Störungen“ – wie bspw. Rechtschreibfehler, Zeichenfehler oder Tippfehler durch Nutzende – weiterhin zuverlässige Ausgaben erzeugen können. Ein Versagen des Systems kann oft schwerwiegende Folgen haben, etwa wenn ein Modell für die Anonymisierung von Gerichtsentscheidungen nach einer Änderung des Dokumentformats nicht mehr zuverlässig funktioniert. Störungen können aber auch durch gezielte Angriffe (engl. adversarial attacks) verursacht werden. Dabei werden die Eingaben absichtlich so verändert, dass das betroffene KI-System mit hoher Wahrscheinlichkeit falsche Entscheidungen trifft. Ein Beispiel hierfür ist die sogenannte Divergenzattacke des Google-Forschungsteams auf das von OpenAI entwickelte ChatGPT. Während ihres Experiments wurde ChatGPT aufgefordert, das Wort „Poem“ unendlich oft zu generieren. Ab einem bestimmten Punkt legte ChatGPT jedoch seine Trainingsdaten offen, anstatt weiter „Poem“ zu generieren.[15]

Auch wenn gezielte Angriffe bei der Anonymisierung von Gerichtsentscheidungen eher unwahrscheinlich sind, verdeutlichen solche Beispiele die Wichtigkeit von robusten KI-Systemen. Dementsprechend beschäftigt sich die computerlinguistische Forschung in den letzten Jahren vermehrt mit Ansätzen zur Verbesserung der Robustheit. Dazu gehört unter anderem „Data Augmentation", ein Ansatz, bei dem die Datenmenge durch die Generierung von künstlichen Daten aus den vorhandenen Trainingsdaten erweitert wird. Eine Variante dieser Methode ist das Training mit zusätzlichen verrauschten Daten (engl. perturbation training), bei dem den ursprünglichen Daten „verrauschte" Instanzen hinzugefügt werden, um die Robustheit des Modells gegenüber unsauberen oder fehlerhaften Daten zu verbessern. Ein weiterer vielversprechender Ansatz zur Verbesserung der Robustheit ist das sogenannte adversarial training, bei dem gezielt schwierige oder manipulierte Eingaben während des Trainingsprozesses generiert und integriert werden, um das Modell gegen potenzielle Angriffe oder Verzerrungen zu schützen.[16]

Abschließend ist zu betonen, dass es trotz der Relevanz der Thematik in der aktuellen computerlinguistischen und KI-Forschung noch kein einheitliches Vorgehen zur Bestimmung der Robustheit gibt. In der Regel werden die gängigen Metriken wie Precision und Recall auf einem Evaluierungsdatensatz verwendet. In einigen Fällen kommen auch die Erfolgsraten von Experimenten mit gezielten Angriffen zum Einsatz. Insgesamt bleibt jedoch die Frage offen, wie Robustheit von KI-Modellen während des Trainings direkt und ohne weitere aufwändige Schritte bewertet und sichergestellt werden kann.[17] Es ist auch darauf hinzuweisen, dass es nahezu unmöglich ist, ein System gegen alle denkbaren Störungen sowie Attacken zu verteidigen. Demnach kann eine uneingeschränkte Robustheit nicht garantiert werden.

III.2. Robustheit in der KI-VO

In der KI-VO wird der Rechtsbegriff der Robustheit anders verstanden: Robustheit wird in Art. 15 KI-VO gemeinsam mit Genauigkeit und Cybersicherheit genannt; durch Abgrenzung zu den beiden Begriffen kann sich dem Verständnis von Robustheit genähert werden.

„Genauigkeit" ist weder in der KI-VO selbst noch in den Erwägungsgründen definiert. In einem vorherigen Aufsatz haben wir argumentiert, dass Genauigkeit einzelfallspezifisch und quantitativ anhand eines Goldstandards gemessen werden muss.[18] Auch wenn die Verordnung davon ausgeht, dass die Systeme vor den Benchmarks entwickelt werden,[19] so leistet der Vergleich mit dem Goldstandard jedenfalls die Messbarkeit der Leistungsfähigkeit des Systems, sodass ein fundierter Diskurs über die Genauigkeitsanforderungen eines Systems überhaupt stattfinden kann. Anders als die Genauigkeit des Systems – also die Leistungsfähigkeit des Systems bei zweck- und bestimmungsgemäßem Gebrauch – betreffen „Cybersicherheit" und „Robustheit" die Leistungsfähigkeit bei unerwarteten oder unvorhergesehenen Ereignissen. Cybersicherheit ist der Schutz des Systems vor böswilligen Dritten.[20] Die Schutzziele der Cybersicherheit sind dabei Verfügbarkeit, Vertraulichkeit, Authentizität und Integrität.[21] Robustheit schließlich meint schwerpunktmäßig die Minimierung von Risiken bezüglich der Grenzen des Systems.[22]

Ein Anhaltspunkt dafür, ob eine Fehlfunktion eines Systems eine Cybersicherheits- oder eine Robustheitsproblematik ist, kann aus der Richtung der Problemquelle abgeleitet werden. Problemquellen von außerhalb des Systems betreffen die Cybersicherheit – Problemquellen, die systemimmanent sind, betreffen die Robustheit des Systems. Gemäß Art. 15 Abs. 4 KI-VO müssen „Hochrisiko-KI-Systeme … so widerstandsfähig [gemeint ist wohl robust] wie möglich gegenüber Fehlern, Störungen oder Unstimmigkeiten sein, die innerhalb des Systems oder der Umgebung, in der das System betrieben wird, insbesondere wegen seiner Interaktion mit natürlichen Personen oder anderen Systemen, auftreten können." Die Lesart von „Robustheit" als „technische Robustheit"[23] wird

dadurch unterstützt, dass die EU in einer Untersuchung zu Robustheit und Erklärbarkeit davon ausgeht, dass Robustheit ein Fehlervermeidungsmaß ist.[24] Durch den offenen Ansatz des Anwendungsbereichs ist nachvollziehbar, dass die KI-VO in der Risikoanalyse nicht von einem statistisch geprägten Robustheitsbegriff wie dem der Computerlinguistik ausgeht. Diese Begriffsüberschneidung ist zwar unglücklich, entspricht aber auch zumindest einer der Verwendungen von Robustheit in der Informatik.

III.3. Robustheit und Domänenanpassung

Während Robustheit die Fähigkeit eines Modells beschreibt, unter veränderten oder herausfordernden Bedingungen (zB verrauschte Daten, veränderte Verteilungen) konsistent gute Leistungen zu erbringen, bezieht sich Domänenanpassung auf die Fähigkeit des Modells, auf Daten aus einer Zieldomäne gut zu funktionieren, die sich wesentlich von den Trainingsdaten (der Quelldomäne) unterscheiden. Beide Aspekte sind miteinander verwandt: ein robustes Modell ist besser in der Lage, mit Veränderungen in den Bedingungen umzugehen, die sich bei der Übertragung auf eine Zieldomäne ergeben; umgekehrt setzt die Domänenanpassung auch Techniken wie adversarial training ein, mit denen die Robustheit des Modells verbessert wird. Domänenanpassung geht aber über die allgemeine Robustheit hinaus, indem es die spezifischen Besonderheiten der Zieldomäne berücksichtigt. Dabei kommen spezielle Techniken wie das Finetuning mit einer kleinen Menge manuell annotierter Trainingsdaten aus der Zieldomäne.[25] Um gleichzeitig Robustheit und effektive Domänenanpassung zu erreichen, sind sowohl robuste Pretraining-Strategien als auch adaptive Mechanismen, die auf spezifische juristische Domänen zugeschnitten sind, erforderlich. Zukünftige Arbeiten könnten zudem die Wechselwirkung zwischen Robustheit und Domänenanpassung in mehrsprachigen oder sprachenübergreifenden juristischen Datensätzen untersuchen, um die Anwendbarkeit von KI-Systemen in verschiedenen Rechtsordnungen weiter auszubauen.

Im Rahmen der KI-VO lässt sich Domänenanpassung nicht wie in der Computerlinguistik als Robustheitsproblem, sondern besser als Zweckbestimmungsproblem verstehen. Nach Art. 3 Nr. 12 KI-VO ist für die Zweckbestimmung auch relevant, in welchen Umständen und unter welchen Bedingungen das System verwendet werden soll. Daher sind Domänen Teil des Zwecks eines Systems. Die Zweckbestimmung des Systems ist eine der Dokumentationspflichten für Hochrisiko-KI-Systeme nach Art. 13 III b) i) KI-VO. In unseren Experimenten (siehe Kapitel IV.) zeigt sich, dass das automatische System nicht den erforderlichen sehr hohen Recall erreicht, ohne auf der Zieldomäne nachtrainiert zu sein. Deshalb sind Rechtsgebiet und Instanz in diesem Fall wesentliche Faktoren für die Beschreibung der Zweckbestimmung.

Neben den Pflichten aus Art. 13 KI-VO ist zu berücksichtigen, dass schon bei der Auswahl der Daten eine Pflicht zur Dokumentation und Begründung aus Art. 10 KI-VO besteht. Einerseits spielt die korrekte Auswahl und Auswertung der vorhandenen Trainingsdaten ebenfalls nach Art. 10 Abs. 2 KI-VO eine unmittelbare Rolle für die Zweckbestimmung; bzw. stehen Datenauswahl und Bestimmung des Zwecks in einer Wechselwirkung. Aus den vorhandenen Daten müssen Einschränkungen des Verwendungszwecks herausgearbeitet werden oder umgekehrt (Abs. 3).

IV. Experimente zur Domänenanpassung

In diesem Abschnitt beschreiben wir unsere Experimente der vollautomatischen Anonymisierung von Gerichtsurteilen auf Basis verschiedener vortrainierter Sprachmodelle, insbesondere dem deutschen GottBERT[26] und dem multilingualen XLM-RoBERTa[27] von der Plattform Huggingface.[28] Die Sprachmodelle werden in einer Multitask-Learning-Architektur für die Anonymisierungsaufgabe angepasst. Das bedeutet, dass drei weitere lineare Schichten hinzugefügt werden, welche als NER-

Tagger fungieren und jeweils für die Erkennung der zu anonymisierenden Textstellen, der jeweils zugehörigen Informationsklasse sowie des Risikoniveaus verantwortlich sind.

Für Training und Evaluation werden die oben beschriebenen Goldstandards aus amtsgerichtlichen sowie oberlandesgerichtlichen Gerichtsentscheidungen herangezogen und jeweils in 50% Trainingsdaten, 25% Development-Daten (für die Optimierung der Modellparameter) sowie 25% Testdaten (für die finale Evaluation) aufgeteilt. Eine Optimierung der Hyperparameter mit Hilfe der Development-Daten wird dabei jeweils vor dem eigentlichen Training durchgeführt. Es sei an dieser Stelle daran erinnert, dass wir eine Domäne als die Kombination aus einem Rechtsgebiet und einer Instanz definieren (bspw. Mietrechturteile der Amtsgerichte). Die Evaluation der Modelle wird hierbei auf Textstellenebene ausgeführt, wobei kleinere Abweichungen für die Berechnung des Recalls zugelassen werden – eine vom Modell anonymisierte Textstelle wird also als korrekt gewertet, auch wenn sie mehr Token abdeckt als im Goldstandard annotiert.[29]

Die Experimente sind wie folgt aufgebaut: Die Sprachmodelle werden jeweils auf AG- und OLG-Entscheidungen trainiert und auf den gleichen Daten evaluiert (In-Domain-Evaluation). Außerdem wird noch ein weiteres Modell auf einen kombinierten Datensatz aus beiden Instanzen feinangepasst. Alle Modelle werden zusätzlich für die Untersuchung der Domänenanpassung auf den oberlandesgerichtlichen Testdaten validiert.

Sprachmodell	Training	Evaluation	Precision	Recall	Recall PII
GottBert	AG	AG	97.35	96.75	99.03
	AG	OLG	87.28	90.88	86.04
	OLG	OLG	92.16	94.05	97.54
	AG+OLG	OLG	91.84	94.08	97.54
XLM-RoBERTa	AG	AG	96.98	96.55	99.12
	AG	OLG	83.70	89.79	93.47
	OLG	OLG	90.69	94.41	97.99
	AG+OLG	OLG	91.90	93.93	96.99

Tabelle 2: Evaluationsergebnisse der Experimente mit GottBERT und XLM-RoBERTa mittels Precision, Recall sowie Recall für direkte Identifikatoren (Recall PII).

Tabelle 2 zeigt, dass die Modelle bei der In-Domain-Evaluation, also bei Training und Evaluation innerhalb der gleichen Domäne, die zu anonymisierenden Textstellen bereits sehr gut identifizieren können, insbesondere direkte Identifikatoren (personally identifying information, PII). Hingegen ist bei den domänenübergreifenden Experimenten (hier AG auf OLG) festzustellen, dass die Übertragung auf eine andere Domäne im Vergleich zum In-Domain-Training (hier OLG auf OLG) deutlich schlechter abschneidet. Demnach kann man festhalten, dass die auf AG-Entscheidungen feinangepassten Modelle trotz sehr guter In-Domain-Leistung keine domänenübergreifende Robustheit gewährleisten können. Ein naheliegender Grund ist, dass die AG-Modelle ausschließlich auf Gerichtsentscheidungen aus Miet- und Verkehrsrecht spezialisiert sind, während der OLG-Datensatz viele andere Rechtsgebiete umfasst. Diese Vermutung wird von der Aufschlüsselung in einzelne Rechtsgebiete (vgl. Tabelle 3) bestätigt. Insbesondere besteht bei Rechtsgebieten wie Beschwerdeverfahren, Immaterialgütern oder Schiedssachen erheblicher Verbesserungsbedarf. Für andere Domänen wie Banken-, Handel- und Kostensachen hat das AG-GottBERT Modell hingegen bei hochsensiblen Textstellen bereits hervorragende Recall-Werte von mindestens 99% erreicht.

OLG-Rechtsgebiete	Precision	Recall	Recall PII	Recall für AG+OLG	Recall PII für AG+OLG

Allg. Zivilsachen	89.22	93.61	97.91	96.26	99.30
Bankensachen	92.89	93.83	100.0	96.51	100.00
Bausachen	94.48	97.13	94.86	98.16	96.37
Beschwerdeverfahren	80.58	95.91	93.83	95.57	98.77
Familiensachen	86.15	92.63	95.40	94.02	98.28
Handelssachen	84.90	95.08	99.07	97.63	100.00
Immaterialgüter	78.65	77.36	83.90	83.11	87.80
Kostensachen	85.53	94.67	100.0	98.00	100.00
Schiedssachen	90.24	85.84	97.87	94.56	98.94
Verkehrsunfallsachen	86.02	85.35	95.88	89.90	98.24

Tabelle 3: Evaluationsergebnisse des AG-GottBERT Modells in den einzelnen Rechtsgebiete aus den oberlandesgerichtlichen Entscheidungen. Zusätzlich werden die Ergebnisse mit OLG-Domänenanpassung (AG+OLG) präsentiert.

Die mit AG+OLG bezeichneten Spalten in Tabelle 3 zeigen die Ergebnisse nach einer Domänenanpassung (in Form von Finetuning auf dem kombinierten Trainingsdatensatz). Erst hier werden über die meisten Rechtsgebiete hinweg zufriedenstellende Werte für eine vollautomatische Anonymisierung erreicht, bei den hochsensiblen PII sogar oft Recall-Werte über 98%. Die nach wie vor schlechten Ergebnisse für Immaterialgüter legen nahe, dass in diesem Rechtsgebiet keine ausreichende Menge von Trainingsdaten für eine erfolgreiche Domänenanpassung vorliegt. Insgesamt bestätigen die Experimente unsere Hypothese, dass für jedes Rechtsgebiet und jede Instanz eine spezifische Domänenanpassung erforderlich ist, um eine vollautomatische Anwendbarkeit der KI-Modelle zu garantieren.

Bei ähnlichen computerlinguistischen Aufgabestellungen hat sich gezeigt, dass Finetuning eines vortrainierten Sprachmodells bereits bei einer kleinen Menge an Trainingsdaten gute Ergebnisse liefert. Um die Frage näher zu beleuchten, wie viele Trainingsdaten für ein erfolgreiches Finetuning benötigt werden, haben wir zusätzlich Lernkurven unserer Modelle bestimmt, welche die des Modells in Abhängigkeit von der Größe des verwendeten Trainingsdatensatzes zeigen. Der Einfachheit halber basieren diese Lernkurven auf einer Evaluation mit seqeval[30], so dass die Werte nicht direkt mit den obenstehenden Tabellen vergleichbar sind (da seqeval im Gegensatz zu unseren Evaluationskriterien kleine, unkritische Abweichungen zwischen Modellvorhersage und Goldstandard nicht akzeptiert). Außerdem werden nur Lernkurven für In-Domain-Anwendung der Modelle gezeigt.

Abbildung 1 zeigt, dass beide Sprachmodelle bereits mit nur zehn Prozent der AG-Trainingsdaten einen Recall von über 90% erreichen. Im Gegensatz dazu weisen die Lernkurven der OLG-Daten noch auf einige Herausforderungen hin. Obwohl ein Recall über 90% mit etwa 20 Prozent der Daten erzielt werden kann, schwanken die Kurven doch deutlich. Das kann vor allem daran liegen, dass diese Gerichtsinstanz mehr Rechtsgebiete zu betreuen hat, also mehr Domänen beinhaltet, die wiederum über weniger Trainingsdaten verfügen.

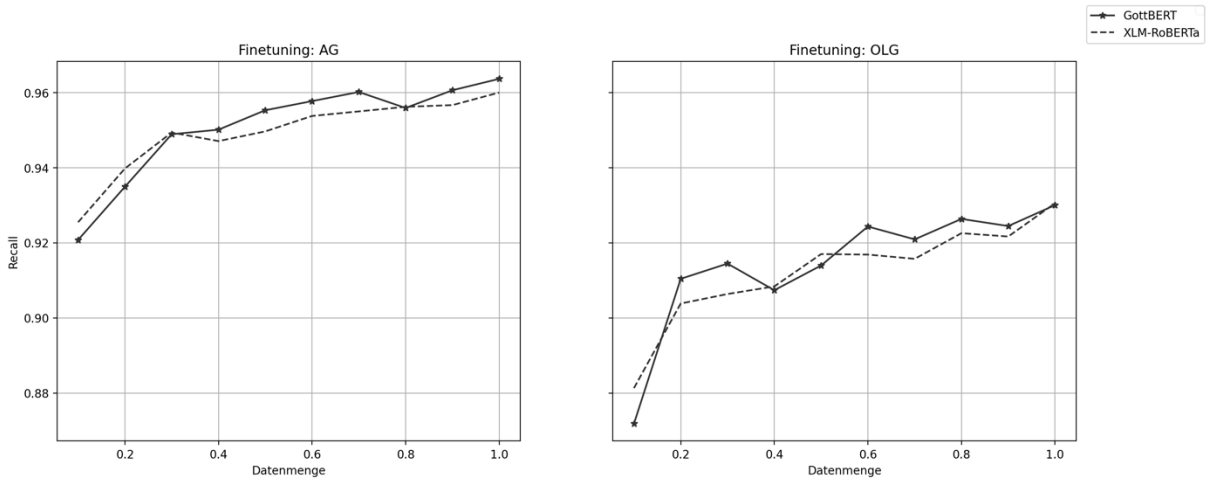

Abbildung 1: Lernkurve für das Finetuning auf AG und OLG-Daten mit GottBERT und XLM-RoBERTa.

Für Abbildungen 2 und 3 wurden die zu anonymisierenden Textstellen in hohe, mittlere und niedrige Risikoniveaus unterteilt und einzeln evaluiert. Beide Abbildungen zeigen deutlich, dass PII (hohes Risiko) in beiden Datensätzen über alle Modelle hinweg sehr gut erkannt wurden. Hier wurden bei allen Modellen ab ca. 30% der Trainingsdaten sehr gute Recall-Werte erreicht. Dagegen wachsen die Lernkurven bei Textstellen mit mittlerem sowie niedrigem Risiko sehr langsam an, so dass vor allem bei OLG-Entscheidungen weitere Verbesserungen nötig sind. Insbesondere deuten die steigenden Kurven hierbei auf eine mögliche Verbesserung hin. Vor allem scheinen die vorliegenden Trainingsdaten noch nicht auszureichen, um optimale Resultate zu erzielen.

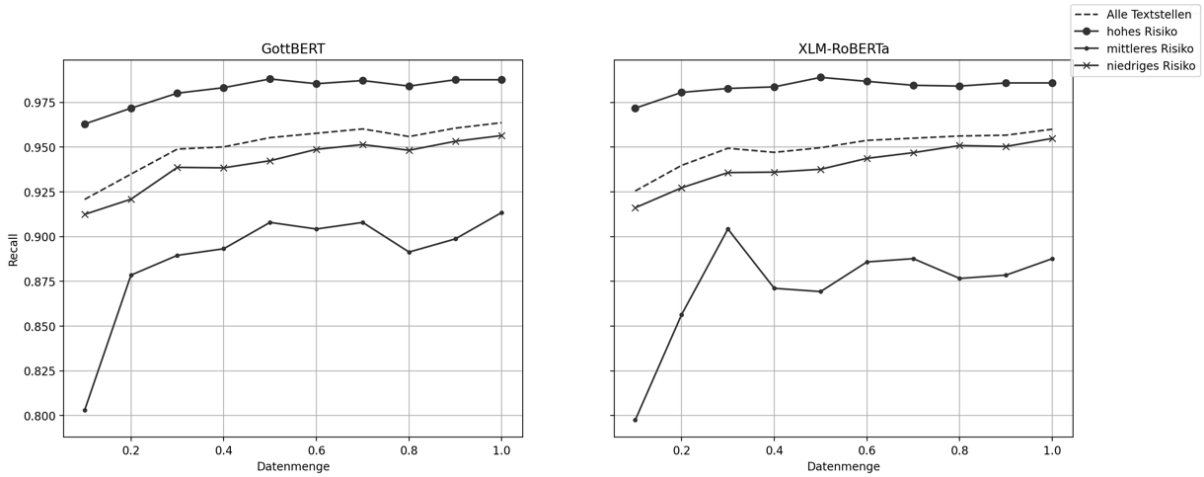

Abbildung 2: Lernkurve auf AG-Goldstandard nach verschiedenen Risikoniveaus unterteilt.

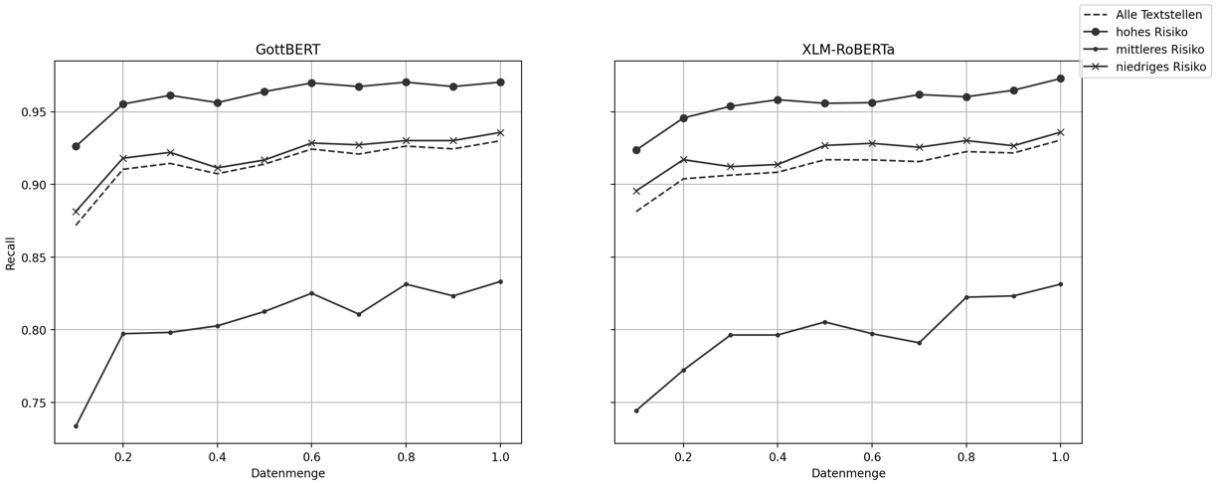

Abbildung 3: Lernkurve auf OLG-Goldstandard nach verschiedenen Risikoniveaus unterteilt.

V. Fazit

Der vorliegende Beitrag führt in das vom BMBF geförderte Projekt *AnGer* ein, das sich der Entwicklung vollautomatischer KI-Systeme zur Anonymisierung von Gerichtsurteilen widmet. Wir können zeigen, dass eine vollautomatische Anonymisierung technisch machbar ist – ein entscheidender Schritt, um langfristig sowohl neu entstehende Urteile als auch den umfangreichen Altbestand effizient anonymisieren zu können.

Gleichzeitig weist unsere Analyse darauf hin, dass die in der KI-VO geforderte Robustheit von Hochrisiko-KI-Systemen nur bedingt mit dem computerlinguistischen Begriff von Robustheit übereinstimmt. Eine zweckgebundene Robustheit erfordert vielmehr die kontinuierliche Überwachung der Leistungsfähigkeit des Systems auf spezifischen Zieldomänen. Unsere Experimente zur Domänenanpassung unterstreichen dabei die zentrale Rolle von Datenumfang und -qualität: Sprachmodelle, die auf Datensätzen aus einer anderen Domäne als der Zielanwendung trainiert wurden, zeigen eine deutlich eingeschränkte Anonymisierungsfähigkeit.

Unsere Ergebnisse unterstreichen die Notwendigkeit, auch in Zukunft auf einen hochwertigen Goldstandard zurückzugreifen, der von geschulten Fachkräften erstellt wird. Dieser Goldstandard bleibt unerlässlich, um die Qualität und Verlässlichkeit vollautomatischer Anonymisierungssysteme sicherzustellen, den regulatorischen Anforderungen gerecht zu werden und die Qualität von KI-Systemen messen zu können. Dabei ist anzumerken, dass für ein breit einsetzbares System ggf. sehr große Goldstandards erstellt werden müssen, damit für alle Rechtsgebiete ausreichende Mengen von Trainingsdaten zur Verfügung stehen.

Schnell gelesen…

- „Robustheit“ von computerlinguistischen KI-Systemen ist als zweckgebundene Übertragbarkeit in neue (sprachliche) Domänen zu verstehen
- Bei der Domänenanpassung ist zu beachten, dass die Domäne Teil der Zweckbestimmung eines KI-Systems ist
- Die Grenzen der Übertragbarkeit in neue Domänen sind daher bei Entwicklung und Einsatz des Systems zu testen und als Grenze des Einsatzzwecks transparent zu machen
- Finetuning von Large Language Models ist der erfolgreichste Ansatz für die vollautomatische Anonymisierung von Gerichtsentscheidungen
- Aber: Umfang und Qualität von Trainingsdaten sind der maßgebliche Schlüssel zum Erfolg

[1] *Keuchen*/*Deuber*, RDi 2022, 229 (230 f.).
[2] *Adrian*/*Evert*/*Heinrich*/*Keuchen*, Auslegung des KI-VO-E zur Evaluation von Verfahren der Künstlichen Intelligenz am Beispiel der automatischen Anonymisierung von Gerichtsentscheidungen, in: Schweighofer, Erich/Eder, Stefan/Costantini, Frederico/Schmautzer, Felix/Pfister, Jonas (Hrg.), Sprachmodelle: Juristische Papageien oder mehr? – Tagungsband des 27. Internationalen Rechtsinformatik Symposions IRIS 2024, S. 205 – 215 (S. 208).
[3] *Gumpp,* ZfPW 2022, 446 (455) mwN
[4] Die Webseite der Forschungsgruppe ist auffindbar unter dem folgenden Link: https://www.linguistik.phil.fau.de/projects/leak-anger/ (zuletzt aufgerufen am 12.11.2024).
[5] Koalitionsvertrag von 2021-2025 „Mehr Fortschritt wagen", S. 85, heruntergeladen von: https://www.spd.de/fileadmin/Dokumente/Koalitionsvertrag/Koalitionsvertrag_2021-2025.pdf (zuletzt aufgerufen am 12.11.2024).
[6] *Recall* entspricht dem Anteil der korrekt als sensibel erkannten Stellen an allen tatsächlich sensiblen Stellen.
[7] Hintergründe und Ziele der automatischen Anonymisierung ausführlicher in: *Adrian*/*Dykes/Evert*/*Heinrich*/*Keuchen*, Automatische Anonymisierung von Gerichtsurteilen – Eine Vision scheint realisierbar, in: Schweighofer, Erich/Zanol, Jakob/Eder, Stefan (Hrsg.), Rechtsinformatik als Methodenwissenschaft des Rechts – Tagungsband des 26. Internationalen Rechtsinformatik Symposions IRIS 2023, S. 211–220.
[8] In der Computerlinguistik ist es üblich, die Größe eines Korpus in sogenannten Token zu messen; diese umfassen neben den eigentlichen Wörtern auch Satzzeichen, Zahlen, Abkürzungen, usw als selbständige Einheiten und bilden die Grundlage für die computerlinguistische Verarbeitung der Texte. Die Anzahl Token ist typischerweise circa 10% größer als die Anzahl Wörter.
[9] Ausführlicher in: *Adrian*/*Dykes*/*Evert*/*Heinrich*/*Keuchen*, Anonymisierung von Gerichtsurteilen – Eine wesentliche Voraussetzung für E-Justice –, in: Schweighofer, Erich/Kummer, Franz/Saarenpää, Ahti/Eder, Stefan/Hanke, Philip (Hrsg.), Cybergovernance - Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS 2021, S. 137 – 149.
[10] *Dykes*/*Evert*/*Heinrich*, Annotator agreement in the anonymization of court decisions, Corpus Linguistics 2021.
[11] *Shah*/*Schwartz*/*Hovy,* Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview, in: Jurafsky/Chai/Schluter/Tetreault (Hrgs.). Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online 2020.
[12] *Lin*/*Lu*/*Tang*/*Han*/*Sun*/*Wei*/*Jing*, A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?, in: Webber/Cohn/He/Liu (Hrgs.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online 2020; *Ma*/*Wang*/*Zhou*/*Zhang*/*Huang*, Towards Building More Robust NER datasets: An Empirical Study on NER Dataset Bias from a Dataset Difficulty View, in: Bouamor/Pino/Bali (Hrgs.), Conference on Empirical Methods in Natural Language Processing, Singapore 2023.
[13] *Wang*/*Wang*/*Yang*, Measure and Improve Robustness in NLP Models: A Survey. in: Carpuat/de Marneffe/Vladimir/Ruiz (Hrsg.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, United States 2022, S. 4569.
[14] *Moradi*/*Samwald*, Evaluating the Robustness of Neural Language Models to Input Perturbations, in: Moens/ Huang/Specia/Wen-tau (Hrsg.), Conference on Empirical Methods in Natural Language Processing, Online 2021.
[15] *Nasr*/*Carlini*/*Hayase*/*Jagielski*/*Cooper*/*Ippolito*/*Choquette-Choo*/*Wallace*/*Tramèr*/*Lee*, Scalable extraction of training data from (production) language models, arXiv preprint arXiv:2311.17035, S. 9.
[16] *Simoncini/Spanakis*. SeqAttack: On Adversarial Attacks for Named Entity Recognition. in: *Adel/Shi* (Htgs.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online and Punta Cana, Dominican Republic, 2021.
[17] *Omar*/*Choi*/*Nyang*/*Mohaisen*, Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions, IEEE Access (10).
[18] *Adrian*/*Evert*/*Heinrich*/*Keuchen*, Auslegung des KI-VO-E zur Evaluation von Verfahren der Künstlichen Intelligenz am Beispiel der automatischen Anonymisierung von Gerichtsentscheidungen, in: Schweighofer, Erich/Eder, Stefan/Costantini, Frederico/Schmautzer, Felix/Pfister, Jonas (Hrg.), Sprachmodelle: Juristische Papageien oder mehr? – Tagungsband des 27. Internationalen Rechtsinformatik Symposions IRIS 2024, S. 205 – 215.

[19] So in § 15 II KI VO; zudem Erwg. 74 S. 6.
[20] Erwg. 76 KI-VO.
[21] Art. 46 CSA.
[22] Erwg. 75 KI-VO, vgl. auch *Martini/Wendehorst*, KI-VO Kommentar, 2024, Art. 15 Rn. 44.
[23] "Technische Robustheit", Wortlaut des Erwg. 75 KI-VO.
[24] *Hamon/Junklewitz/Sanchez*, Robustness and Explainability of Artificial Intelligence - From technical to policy solutions, EUR 30040.
[25] *Wilson/Cook,* A survey of unsupervised deep domain adaptation, ACM Transactions on Intelligent Systems and Technology (TIST), 11(5), 2020.
[26] *Scheible/Frei/Thomczyk/He/Tippmann/Knaus/Jaravine/Kramer/Boeker*, GottBERT: a pure German Language Model, in: Al-Onaizan/Bansal/Chen (Hrgs.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, United States 2024.
[27] *Conneau/Khandelwal/Goyal/Chaudhary/Wenzek/Guzmán/Grave/Ott/Zettlemoyer/Stoyanov*, Unsupervised Cross-lingual Representation Learning at Scale, in: Jurafsky/Chai/Schluter/Tetreault (Hrgs.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online 2020.
[28] *Wolf*, Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771.
[29] Solche Textstellen reduzieren jedoch Precision, da „zu viel" annotiert wird.
[30] https://github.com/chakki-works/seqeval (zuletzt aufgerufen am 02.12.2024).