Entwicklung und Evaluation automatischer Verfahren zur Anonymisierung von Gerichtsentscheidungen

Axel Adrian, Nathan Dykes, Stephanie Evert, Philipp Heinrich und Michael Keuchen *

Abstract

Dieser Beitrag beschäftigt sich mit der Entwicklung und Evaluation von Verfahren zur automatischen Anonymisierung von Gerichtsentscheidungen. Wir stellen zunächst den Prozess der Erstellung eines Goldstandards vor, der auf Grundlage von Richtlinien, die der aktuellen Rechtslage folgen, manuell erstellt wurde. Dieser dient als Grundlage für Training und Evaluation eines von uns entwickelten neuronalen Sprachmodells zur automatischen Annotation von sensiblen Textstellen. Neben unserem eigenen Prototyp (LeAK) werden drei verfügbare Tools (TAB, OpenRedact und A-Tool) vorgestellt.

I. Einleitung

Aufgrund des Demokratieprinzips, des Rechtsstaatsprinzips, des Gewaltenteilungsgrundsatzes und des Justizgewährungsanspruchs besteht nach höchstrichterlicher Rechtsprechung eine Pflicht zur Veröffentlichung von veröffentlichungswürdigen Gerichtsentscheidungen.[1] Transparenz und Nachvollziehbarkeit der Entscheidungspraxis der Gerichte sind ein Kernbestand von Demokratie und Rechtsstaatlichkeit und sind ganz allgemein nicht nur für Rechtspraxis und Rechtswissenschaft, sondern auch für die Öffentlichkeit wichtig, um das eigene rechtliche Verhalten anpassen und die Gerichtspraxis einschätzen und beurteilen zu können.[2] Daneben ist die Verfügbarkeit von Entscheidungen ein wesentlicher Baustein für zukünftige Anwendungen der künstlichen Intelligenz (KI) im großen Spektrum von Legal Tech.[3] Was liegt näher, als Gerichtsentscheidungen als Trainingsdaten für KI-Anwendungen zu nutzen? Um hier erfolgreich zu sein, werden sehr große Datenmengen benötigt.[4] In Deutschland wurden jedoch beispielsweise von allen Gerichtsbarkeiten in den Jahren 2011 bis 2020 nur ca. 2,3% der

* Der Autor Prof. Dr. Axel Adrian ist Notar und Honorarprofessor für Rechtstheorie und Rechtsgestaltung an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU); der Autor Nathan Dykes, M.A., ist wissenschaftlicher Mitarbeiter und Doktorand an der FAU; die Autorin Prof. Dr. Stephanie Evert ist Inhaberin des Lehrstuhls für Korpus- und Computerlinguistik an der FAU; der Autor Philipp Heinrich, M.Sc., ist wissenschaftlicher Mitarbeiter und Doktorand an der FAU; der Autor Michael Keuchen, Ass. Jur., ist wissenschaftlicher Mitarbeiter und Doktorand an der FAU.

[1] BVerwG NJW 1997, 2694 (2695); BVerfG NJW 2015, 3708 (3710); BGH NJW 2017, 1819 (1819 f.).

[2] BVerwG NJW 1997, 2694 (2695); BGH NJW 2017, 1819 (1820); *Huff*, Justiz und Öffentlichkeit, 1996, S. 8 f; *Lederer*, Open Data, 2015, S. 53.

[3] *Keuchen/Deuber* RDi 2022, 189 (190 f.).

[4] *Adrian/Schröder/Maier*, in: Adrian/Evert/Kohlhase/Zwickel, Digitalisierung von Zivilprozess und Rechtsdurchsetzung, 2022, S. 199 ff.; *Keuchen/Deuber* RDi 2022, 229 (235); *Bilski/Schmid* NJOZ 2019, 657.

grundsätzlich veröffentlichungsfähigen Gerichtentscheidungen in öffentlich zugänglichen Rechtsprechungsportalen von Bund und Ländern publiziert.[5]

Eine wesentliche Ursache, warum nur so wenige Entscheidungen veröffentlicht werden, ist die Notwendigkeit vor einer Veröffentlichung sicherzustellen, dass keine Rechte der beteiligten natürlichen und juristischen Personen verletzt werden.[6] Zu den zu schützenden Rechtspositionen der Beteiligten und Betroffenen gehören insbesondere das Recht auf informationelle Selbstbestimmung (Art. 2 Abs. 2 GG i.V.m. Art. 1 Abs. 1 GG)[7], das Unternehmenspersönlichkeitsrecht, die Betriebs-, Geschäfts-, Steuer-[8], und Sozialgeheimnisse (§ 30 AO und § 35 SGB I) und der allgemeine Datenschutz.[9] Diese Rechte müssen zumindest durch eine effektive Anonymisierung geschützt werden, bevor Entscheidungen veröffentlicht werden dürfen.[10] Die Anonymisierung ist die Erkennung und Unkenntlichmachung kritischer Textstellen (wie Personennamen, Adressen, Datumsangaben, Kennzeichen sowie weitere (indirekt) identifizierende Merkmale). Damit Informationen als anonym im rechtlichen Sinne gelten, dürfen die zu schützenden Personen nicht oder nur mit einem großen Aufwand bestimmbar sein.[11] Das ist dann der Fall, wenn die personenbezogenen Daten so verändert werden, dass Einzelangaben über persönliche und sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft, gegebenenfalls unter Heranziehung von Zusatzwissen aus allgemein zugänglichen Quellen, einer bestimmten oder bestimmbaren Person zugeordnet werden kann.[12] Daher ist derzeit eine sehr aufwändige, weil nur manuell durchgeführte, Anonymisierung von Gerichtsentscheidungen vorzunehmen, so dass die Zahl an Veröffentlichungen durch personelle und sächliche Kapazitätsgrenzen der Justizverwaltungen erheblich beschränkt wird. Zugleich darf die Anonymisierung aber nicht so weit reichen, dass Lesbarkeit und Verständlichkeit der Entscheidung nicht mehr gegeben sind. Deshalb wird es häufig erforderlich sein, statt einer reinen „Schwärzung“, insbesondere zur Unterscheidung von Beteiligten eine Pseudonymisierung vorzunehmen. Die Pseudonymisierung ist die Maskierung der kritischen Textstellen durch Ersetzung mit realistischen Fantasiebezeichnungen. Anhand dieser Beschreibung und Herausforderungen, einen gerechten Ausgleich zwischen öffentlichen Informationsinteressen und berechtigten Interessen der Beteiligten und Betroffen zu finden, zeigt sich das Problem einer effizienten Anonymisierung von Gerichtsentscheidungen und ist damit eine der Hauptursachen, warum die Digitalisierung von rechtswissenschaftlichen Anwendungen im Vergleich zum Digitalisierungsstand in anderen Disziplinen stark hinterherhinkt. Auch um später Gerichtsentscheidungen als Trainingsdaten für KI-Anwendungen nutzen zu können, muss eine Pseudonymisierung und nicht nur eine Anonymisierung erfolgen.

---

[5] *Keuchen/Deuber* RDi 2022, 229 (233).

[6] *Adrian/Dykes/Evert/Heinrich/Keuchen/Proisl*, in: Adrian/Evert/Kohlhase/Zwickel, Digitalisierung von Zivilprozess und Rechtsdurchsetzung, S. 173.

[7] BVerfG NJW 1984 419 (422); *Nöhre* MDR 2019, 136 (136 f.); BGH NJW 2018 3123.

[8] *Haupt* DStR 2014, 1025 (1029).

[9] *Adrian/Evert/Keuchen/Heinrich/Dykes*, in: Schweighofer/Kummer/Saarenpää/Eder/Hanke, Cybergovernance – Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS, 2021, S. 137 (142 f.).

[10] OLG Karlsruhe Beschluss vom 22.12.2020, 6 VA 24/20 = GRUR-RS 2020, 37423, Rn. 36–39; *Nöhre* MDR 2019, 136 (138 f.); BVerwG NJW 1997, 2694 (2695).

[11] OLG Karlsruhe Beschluss vom 22.12.2020, 6 VA 24/20 = GRUR-RS 2020, 37423, Rn. 36–39; VGH Baden-Württemberg MMR 2011, 277 (278).

[12] OLG Karlsruhe Beschluss vom 22.12.2020, 6 VA 24/20 = GRUR-RS 2020, 37423, Rn. 36–39; VGH Baden-Württemberg MMR 2011, 277 (278); *Adrian/Evert/Keuchen/Heinrich/Dykes*, in: Schweighofer/Kummer/Saarenpää/Eder/Hanke, Cybergovernance – Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS, S. 137 (141); vgl. Erwg. 26 DSGVO.

Wir arbeiten daher seit 2020 im Rahmen eines Forschungsprojektes mit dem Bayerischen Staatsministerium der Justiz an rechtlichen und technischen Fragen zur Möglichkeit einer automatischen Anonymisierung von Urteilen mit Hilfe computerlinguistischer Verfahren (LeAK = Legal Anonymization Kit).[13] Die Zwischenergebnisse des ersten Projektziels, nämlich die Entwicklung und detaillierte Evaluation eines Software-Prototypen für die automatische Erkennung sensibler Textstellen, wurden bereits veröffentlicht.[14]

Es existieren, soweit ersichtlich, derzeit jedenfalls auch bereits zwei Anwendungen, die eine automatische Anonymisierung von Gerichtsurteilen versprechen bzw. erforschen wollen, nämlich "Text Anonymization Benchmark" (TAB)[15] und OpenRedact[16]. Daneben wird in der deutschsprachigen Gerichtspraxis der Schweiz derzeit bereits an einzelnen Gerichten teilweise das A-Tool[17] eingesetzt, eine Anwendung, die jedenfalls eine Unterstützung der manuell anonymisierenden Richter und Richterinnen anbietet. Flächendeckende automatische Lösungen in Deutschland, die in der Gerichtspraxis eingesetzt würden, existieren allerdings noch nicht, wenn auch bereits interessante Forschungsergebnisse zu einer automatischen Anonymisierung vorliegen. So gibt es einerseits Arbeiten auf Basis von bereits anonymisierten Daten, die kritische Textstellen rein anhand des umgebenden Textmaterials erkennen.[18] Andere Ansätze beziehen Metadaten aus den Gerichtsverwaltungssystemen ein und nutzen diese daneben als Grundlage für eine Anonymisierung.[19] Das alleinige Stützen der Anonymisierung auf solche Metadaten aus Gerichtsverwaltungssystemen ist jedenfalls unzureichend.

Im vorliegenden Beitrag sollen nun wenigstens beispielhaft die oben genannten Anwendungen TAB, OpenRedact und A-Tool kurz dargestellt und überblicksartig mit den Erkenntnissen aus unserem Forschungsprojekt, bei dem mit unveränderten, also noch nicht anonymisierten, Urteilen und Schriftsätzen gearbeitet wird, verglichen werden. Dabei soll auch die Frage geklärt werden, wie überhaupt die Leistungsfähigkeit und Zuverlässigkeit von solchen automatischen Anonymisierungssystemen gemessen, wie solche Anwendungen also evaluiert werden können.

II. Rechtsgrundlagen der Anonymisierung (Richtlinien)

Bei der Frage, wie überhaupt eine Evaluation automatischer Systeme durchgeführt werden kann, ist wichtig zu verstehen, dass es bis heute keine einheitlichen und flächendeckenden Richtlinien der Justizverwaltungen gibt, nach denen festgelegt wäre, welche Textstellen zu anonymisieren sind.[20]

---

[13] *Adrian/Evert/Keuchen/Heinrich/Dykes*, in: Schweighofer/Kummer/Saarenpää/Eder/Hanke, Cybergovernance – Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS, S. 137-147.

[14] *Adrian/Dykes/Evert/Heinrich/Keuchen/Proisl*, in: Adrian/Evert/Kohlhase/Zwickel, Digitalisierung von Zivilprozess und Rechtsdurchsetzung, 2022, S. 173-197.

[15] *Pilán/Lison/Øvrelid/Papadopoulou/Sánchez/Batet*, The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization, 2022, abrufbar unter: https://arxiv.org/pdf/2202.00443v1.pdf. Alle Internetquellen wurden zuletzt am 18.08.2022 abgerufen.

[16] https://openredact.org.

[17] https://www.balo.ai/a-tool.

[18] *Glaser/Schamberger/Matthes*, Eighteenth International Conference for Artificial Intelligence and Law, 2021, S. 205 ff.

[19] *Dévaud/Kummer*, in: Hürlimann/Kettiger, Anonymisierung von Urteilen, 2021, S. 61 ff.

[20] *Adrian/Dykes/Evert/Heinrich/Keuchen/Proisl*, in: Adrian/Evert/Kohlhase/Zwickel, Digitalisierung von Zivilprozess und Rechtsdurchsetzung, S. 173 (176).

Einzelne Gerichte oder Gerichtsverwaltungen arbeiten mit internen Richtlinien,[21] die sich aber untereinander in Bezug auf den Anonymisierungsumfang, die zu verwendenden Anonymisierungstechniken und den Arbeitsablauf stark unterscheiden. Solche einheitlichen und umfassenden Richtlinien müssen also erst entwickelt werden, um einen (rechtlichen) Maßstab für eine (technisch-mathematische) Evaluation überhaupt erst zu schaffen. Dies setzt zuallererst die Klärung von verschiedenen Rechtsfragen voraus, die nach den relevanten gesetzlichen Vorschriften und den Erkenntnissen aus der Rechtsprechung zu beantworten sind. Hierzu wurde in unserem Forschungsprojekt ausführlich geforscht. Wie bereits gezeigt, existieren diverse Rechtspositionen, die bei einer Anonymisierung zu schützen sind. Daneben ist der gesetzliche Maßstab für die Anonymisierungsqualität sehr hoch, da nur mit einem unverhältnismäßig großen Aufwand eine Deanonymisierung möglich sein darf. Bei der Frage, wann ein unverhältnismäßiger Aufwand zur Deanonymisierung nötig wäre, gilt ein relativer bzw. subjektiver Maßstab aus Sicht der verantwortlichen Stelle; d.h. Sonder- oder Zusatzwissen Dritter oder Vierter führt nicht allein deswegen zu einer Identifizierbarkeit, weil es objektiv besteht. Dennoch genügt es bereits, wenn dieses Wissen aus zulässigen öffentlich und allgemein zugänglichen Quellen herangezogen werden kann, was wiederum ein enormes Gefahrenpotenzial schafft, dem die Anonymisierung standhalten muss. Frei verfügbares Wissen sowie spezifische Verfahrensinformationen aus Medienberichten lassen sich leicht und schnell mit Informationen aus dem Urteil verknüpfen und ermöglichen eine Deanonymisierung.[22]

Demnach müssen die Annotationsrichtlinien darauf abzielen, sämtliche kritische Wörter, Sätze und Textpassagen im Urteil zu erfassen, die möglicherweise einer späteren Entscheidung zur Anonymisierung, Pseudonymisierung, Entfernung größerer Teile aus dem Urteil oder einem vollständigen Unterbleiben der Entscheidungsveröffentlichung dienen. Bei der Risikobewertung (hoch, mittel, niedrig) und Entscheidung einer Annotationswürdigkeit muss nicht nur die jeweilige konkrete Textstelle in den Blick genommen werden, sondern diese ist im Gesamtkontext des Urteils zu betrachten. Besonders gefährlich und meist nicht auf den ersten Blick zu erkennen sind Umstände, die nur in der Kombination entweder von Fakten allein aus dem Urteil oder in Kombination mit externem Wissen, eine Identifizierbarkeit der Personen erlauben (sog. cross-referencing).[23]

Daraus folgt, dass nach den Annotationsrichtlinien im Zweifel viele Textstellen annotiert werden müssen. Damit im Nachhinein aber eine Bewertung und Verwendung der Annotationen für die Trainingsdaten vorgenommen werden kann, muss zusätzlich eine Kategorisierung erfolgen. So müssen die verschiedenen annotierten Informationen in Risikokategorien, wie hoch bei Namen und Adressen, mittel bei Datumsangaben und niedrig bei indirekten Informationen eingestuft werden. Natürlich kann wegen der Umstände des Einzelfalls eine andere Einstufung des Risikos erforderlich sein, weil es sich beispielsweise um eine seltene oder einmalige Konstellation handelt. Daneben verlangen die Richtlinien noch die Angabe, ob die Information zum Verständnis der Entscheidung essenziell ist. Dieses Informationserhaltungsinteresse kann später bei der Entscheidung über die Anonymisierung einbezogen werden und gegebenenfalls das Eingehen eines geringen Deanonymisierungsrisikos rechtfertigen.

---

[21] *Van Opijnen/Peruginelli/Kefali/Palmirani*, On-line Publication of Court Decisions in the EU, 2017, S. 73, abrufbar unter: https://bo-ecli.eu/uploads/deliverables/Deliverable%20WS0-D1.pdf.

[22] *Vokinger/Mühlematter*, Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(banken), Jusletter 2.9.2019, S. 16.

[23] BeckOK DatenSR/*Schild*, 37. Edition 01.08.2021, Grundlagen und bereichsspezifischer Datenschutz, Syst. E., Rn. 58; *Bieri*, in: Hürlimann/Kettiger, Anonymisierung von Urteilen, S. 1, 13.

III. Goldstandard (manuelle Annotation)

Wenn die rechtlichen Anforderungen analysiert und festgelegt sind, müssen Schritte erfolgen, die eine technische Evaluation der Systeme ermöglichen. Hierzu muss ein sog. Goldstandard erstellt werden, d.h. es müssen — einfach ausgedrückt — Texte maschinenlesbar erfasst und manuell (gegebenenfalls mit Hilfe eigens dafür entwickelter Tools) perfekt anonymisiert werden. Da davon auszugehen ist, dass menschliche Annotatorinnen und Annotatoren nicht absolut fehlerfrei arbeiten (d.h., dass sie sensible Textstellen übersehen oder auch überflüssige Stellen als sensibel markieren können), müssen die Texte von mehreren Menschen annotiert werden, um sicherzustellen, dass keine nach den rechtlichen Anforderungen zu anonymisierende Textstelle übersehen wurde — dass also perfekt anonymisierte Texte vorliegen.

Dieser Goldstandard kann dann als Grundlage dienen, um die Leistungsfähigkeit maschineller Verfahren bewerten zu können. Die Leistungsfähigkeit maschineller Verfahren kann somit auch direkt mit der Leistungsfähigkeit perfekt geschulter Menschen verglichen werden und ebenso kann die manuelle Leistungsfähigkeit eines Menschen bei der Erfüllung der Anonymisierungsaufgabe gemessen werden. An dieser Stelle ist anzumerken, dass die Erstellung der Anonymisierungsrichtlinien und die Entwicklung der Annotationstools — sowie die manuelle Erstellung des Goldstandards — ein iterativer Prozess ist: die einzelnen Schritte können nicht einfach linear aufeinander erfolgen. So erfordert eine Überarbeitung der Richtlinien aus offensichtlichen Gründen eine nochmalige Annotation der bereits annotierten Texte. Andererseits tauchen bei der Erschließung größerer Textmengen von Text unweigerlich weitere Problemfälle auf, die in den Richtlinien behandelt werden müssen; insbesondere, wenn Texte aus anderen Rechtsgebieten in den Goldstandard aufgenommen werden sollen.

Da a priori unklar ist, wie schwierig das Auffinden aller zu anonymisierenden Stellen in einem Fließtext ist, wie konsistent Annotatorinnen und Annotatoren in ihren Entscheidungen sind („Intra-Annotator-Reliabilität“) und wie gut sie gegenseitig übereinstimmen („Inter-Annotator-Reliabilität“), ist es notwendig, Urteile sowohl mehrfach vom gleichen Menschen als auch von mehreren Menschen bearbeiten zu lassen. In unserem eigenen Projekt (siehe Abschnitt VI.) wurde daher zu Beginn jedes Urteil von fünf, im weiteren Verlauf von vier, geschulten Hilfskräften unabhängig voneinander annotiert.[24] Es zeigte sich dabei insbesondere, dass die (paarweise) Inter-Annotator-Reliabilität bei Vorliegen von ausführlichen Richtlinien sehr hoch ist. Die Standardmaßzahl zur Berechnung der Intra- und bzw. paarweisen Inter-Annotator-Reliabilität ist Cohen’s kappa, d.h. die prozentuale Übereinstimmung korrigiert um die Übereinstimmung, die bei rein zufälligen Entscheidungen anzunehmen wäre. Gängige Maßzahlen zur Evaluation der Intra-Annotator- bzw. paarweisen Inter-Annotator-Reliabilität sind Krippendorff's alpha und Cohen's kappa. Diese Werte können hier allerdings nicht angewandt werden, da es sich nicht um eine einfache Klassifikation vorgegebener Objekte handelt: die Anzahl der zu annotierenden Textstellen ist ja nicht von vornherein bekannt. Eine geeignetere Evaluation zur Beurteilung der manuellen Annotation sind daher die in Abschnitt IV. vorgestellten Maßzahlen *Precision* und *Recall* gegen den finalen Goldstandard.

Trotz sorgfältiger und mehrfacher Annotation bleibt offensichtlich immer ein Restrisiko, dass eine sensible Textstelle von allen Hilfskräften übersehen wurde. Die vier- bzw. fünffache Annotation der Urteile ist eine

[24] *Adrian/Dykes/Evert/Heinrich/Keuchen/Proisl*, in: Adrian/Evert/Kohlhase/Zwickel, Digitalisierung von Zivilprozess und Rechtsdurchsetzung, S. 173 (181 f.).

Vorsichtsmaßnahme, um sicherzustellen, dass alle zu annotierenden Textstellen mit hoher Wahrscheinlichkeit gefunden werden. Zur Erfassung der Wahrscheinlichkeit, dass weitere Hilfskräfte weitere sensible Textstellen finden würden, kann ein einfaches statistisches Modell eingesetzt werden.[25] Wir gehen dazu vereinfachend davon aus, dass alle Annotatorinnen und Annotatoren zufällig und unabhängig voneinander zu annotierenden Textstellen mit Ausfallwahrscheinlichkeit *q* übersehen und alle Textstellen gleich schwierig zu erkennen sind; wir gehen also insgesamt von reinen Flüchtigkeitsfehlern aus (und nicht etwa intrinsisch unterschiedlich schwierig zu entscheidenden bzw. leicht zu übersehenden Stellen).[26]

Unter diesen Modellannahmen ist die Anzahl der verbleibenden *False Negatives* (Textstellen, die fälschlicherweise nicht markiert wurden) im Goldstandard leicht abschätzbar. Für die vorliegenden Daten wurde ermittelt, dass vier Personen ausreichend sind, um einen nahezu perfekten Goldstandard mit ca. 1 Million Wörtern zu erstellen; das Modell liefert hierfür die Abschätzung, dass eine fünfte Person im Schnitt weniger als eine weitere zu anonymisierende Stelle finden würde.

IV. Bewertungskriterien für Anonymisierungsverfahren

Um nun die Leistungsfähigkeit und Zuverlässigkeit von automatischen Verfahren (oder auch von menschlichen Annotatorinnen und Annotatoren) messen zu können, werden typischerweise die folgenden beiden wichtigsten Maßzahlen betrachtet: dies ist erstens die Sensitivität (auch *Recall* genannt), um zu messen, welcher Anteil der tatsächlich zu anonymisierenden Stellen vom System gefunden werden und zweitens der positive Vorhersagewert (auch *Precision* genannt), um zu bestimmen, welcher Anteil der vom System gefundenen Stellen tatsächlich zu anonymisieren sind.[27] Sind beispielsweise in einem Text 100 Stellen zu anonymisieren und findet ein System davon 95 (*True Positives*), wobei es fälschlicherweise zusätzliche zehn Stellen (*False Positives*) markiert, so hat dieses System 95/100 = 95% Recall und 95/105 ≈ 90% Precision.

Prinzipiell stehen diese beiden Größen in wechselseitiger Abhängigkeit, d.h. eine Erhöhung von Recall ist meist einfach zu erreichen, indem eine niedrigere Precision in Kauf genommen wird. Für hochsensible Daten, wie sie in unserem Projekt vorliegen, ist insbesondere der Recall ausschlaggebend: Eine übersehene Stelle (beispielsweise der Name eines Zeugen) stellt bei der Anonymisierung eine weit schwerwiegende Grundrechtsverletzung dar, als eine zusätzliche – d.h. unnötigerweise – anonymisierte Stelle. Sehr geringe Precision führt allerdings aus offensichtlichen Gründen zur Unlesbarkeit und Unverständlichkeit des Urteils.

---

[25] *Heinrich/Evert/Dykes*, Annotator agreement in the anonymization of court decisions, Proceedings of Corpus Linguistics Conference, 2021, abrufbar unter: https://corpora.linguistik.uni-erlangen.de/data/cl2021_377_heinrich.html.

[26] *q=0,02* bedeutet beispielsweise, dass jede Hilfskraft 2% aller Textstellen übersieht. Das Modell vernachlässigt damit auch, dass unterschiedliche Hilfskräfte unterschiedliche Leistungsfähigkeit aufweisen.

[27] *Sebastiani,* An axiomatically derived measure for the evaluation of classification algorithms, in: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, 2015, S. 11-20.

V. Überblick über die Anwendungen TAB, OpenRedact und A-Tool

Der kürzlich als Preprint veröffentlichte Text Anonymization Benchmark ist vergleichbar mit dem Ansatz unseres Forschungsprojekts. TAB befasst sich mit der Anonymisierung von englischsprachigen Urteilen des Europäischen Gerichtshofs für Menschenrechte (EGMR). Im Gegensatz zu unserem Projekt (siehe Abschnitt VI.), bei dem auf höchste Datenqualität abgezielt wird, wurden hier jedoch Quantität und Schnelligkeit in den Mittelpunkt gestellt. In einem semiautomatischen Annotationsprozess wurden potenziell zu anonymisierende Textstellen zunächst automatisch von computerlinguistischen Standardwerkzeugen (Named-Entity-Recognition von spaCy) vorgeschlagen und bei der anschließenden manuellen Überprüfung nur in ca. 24% der Fälle verändert. Nur kleine Teile des Goldstandards wurden von mehreren Annotatorinnen und Annotatoren bearbeitet, deren Übereinstimmung oft sehr niedrig liegt. Es überrascht daher nicht, dass der für eine automatische Anonymisierung eingesetzte Prototyp die vorher automatisch vorgeschlagenen Textstellen mit hoher Genauigkeit wieder auffindet. Eine Veröffentlichung dieses Goldstandards ist überhaupt nur deshalb möglich, weil der EGMR die öffentlichen Informationsinteressen gegenüber dem Schutz der privaten Interessen größtenteils überwiegen lässt und sich grundsätzlich nicht zur Anonymisierung verpflichtet sieht. Nur ausnahmsweise, auf besonderen und begründeten Antrag eines Beteiligten nach Art. 47 Nr. 4 EGMRVerfO, wird über eine Anonymisierung entschieden und damit vom extensiven Prinzip der Öffentlichkeit (Art. 6 Abs. 1 S. 2, 44 III EMRK) abgewichen. Dabei müssen aktiv vom Antragenden die möglichen Auswirkungen in Folge der Veröffentlichung dargelegt werden, wie beispielsweise negative Folgen für die (psychische) Gesundheit, Gefahr vor Vergeltungsmaßnahmen oder eine beeinträchtigte Entwicklung von Kindern.[28] Das grundsätzliche Regel-Ausnahme-Verhältnis von Anonymisierung der Beteiligten von Amts wegen und Öffentlichkeitsgewähr ist daher ein umgekehrtes im Vergleich zum deutschen Recht, obwohl selbst Art. 8 EMRK das Recht auf Achtung des Privat- und Familienlebens kodifiziert. Geboten erscheint hier vielmehr auch eine Anonymisierung, die letztendlich nicht im Widerspruch zum Öffentlichkeitsgebot steht, sondern schützenswerte Interessen der Beteiligten wahrt.[29] Im durch unser Projekt untersuchten staatlichen (deutschen) Bereich gelten daher viel strengere datenschutzrechtliche Anforderungen, die durch unsere Forschungsfragen und unsere Goldstandards repräsentiert werden müssen.

OpenRedact ist ein Open-Source-Tool zur Anonymisierung von deutschsprachigen Texten. Dabei werden besonders Behörden und Gerichte als Zielgruppe in den Blick genommen. Ausdrücklich sollen Urteile oder auch sonstige behördliche Texte z.B. bei etwaigen Anfragen auf Grund von Informationsfreiheitsgesetzen mit dem Tool anonymisiert werden können. Die Webseite verspricht eine einfache Anonymisierung von Dokumenten in unterschiedlichen Formaten (Microsoft Word, PDF, Plain-Text) mittels Drag & Drop in einer Web-App bzw. RESTful-Anfragen an die API oder auch über Nutzung des Command-Line-Interfaces des zugrundeliegenden Backends. Die Technologie zur Schwärzung basiert hier hauptsächlich auf einer Kombination von regulären Ausdrücken und Named-Entity-Recognition. Diese Komponenten müssen jedoch auf den jeweiligen Anwendungsfall zugeschnitten werden, was im Verantwortungsbereich des Nutzers liegt; Outof-the-box liegt die Performance für unseren Goldstandard bei ca. 80% Recall und etwas weniger als 50% Precision. Ein entsprechender Disclaimer (dass die Software für keine kritischen Aufgaben

[28] *Ernst*, Transparenz in der Judikative, 2021, S. 172 f. m.w.N.

[29] *Grabenwarter/Pabel*, Europäische Menschenrechtskonvention, 7. Aufl., 2021, § 24 Rn. 111; *Lutschouing*, Entscheidungsveröffentlichung im Zivilprozess, 2021, S. 108-112 m.w.N.

verwendet werden sollte) ist im Software-Repository vermerkt. Es ist hervorzuheben, dass Workflow und Frontend zwar weit produktreif entwickelt sind, eine aktive Weiterentwicklung, d.h. insbesondere Training auf weiteren Daten, nicht stattfindet. Der grundsätzliche Ansatz, auch indirekte Identifikatoren zu anonymisieren, ist zwar vorhanden, jedoch hat OpenRedact keinen Zugriff auf umfangreiche manuell annotierte Daten, die für ein entsprechendes Training notwendig wären.

Die Firma BALO.AI bietet, wie erwähnt, Anonymisierungslösungen für schweizerische Gerichte an. Ihr A-Tool markiert Textstellen in den Entscheidungen direkt in Word (Word-Plugin) und schlägt Platzhalter vor.[30] Hierbei werden die sensiblen Begriffe teilweise vorab über die im Verwaltungssystem hinterlegten Daten definiert (insb. Namen, Geburtsdaten, Anschriften der Beteiligten) und mittels einfachem „Suchen und Ersetzen" anonymisiert. Indirekte Identifikatoren, die nicht als Metadaten hinterlegt sind, können hier nicht anonymisiert bzw. pseudonymisiert werden. Die Leistung eines solchen Tools hängt offensichtlich stark davon ab, welche Daten im Verwaltungssystem hinterlegt sind, wie viel Zeit und Sorgfalt in die manuelle Ausweitung dieser Begriffe geht und wie konsistent die Erwähnungen im Fließtext sind. Eine systematische Evaluation des A-Tools liegt u.a. deswegen nicht vor; die kommerzielle Webseite verspricht lediglich, dass mit dieser Lösung „bis zu 70% schneller anonymisier[t]" werden kann.

VI. Unser Projekt: *Legal Anonymization Kit* (LeAK)

Unser eigenes Projekt, LeAK, konzentriert sich auf die automatische Anonymisierung und Pseudonymisierung von Urteilen aus Miet- und Verkehrsrecht. Zu diesem Zweck wurden zunächst umfangreiche Datensätze manuell annotiert. Die Datensätze im vorliegenden Projekt umfassen ca. 250 Urteile (400.000 Wörter) aus dem Mietrecht und ca. 320 Urteile (ca. 550.000 Wörter) aus dem Verkehrsrecht. Zur Erstellung dieser Textbasis wurde jedes Urteil von mindestens vier geschulten Annotatorinnen und Annotatoren bezüglich sensiblen Textstellen gesichtet. Die Hilfskräfte folgten dabei den oben aufgeführten detailliert ausgearbeiteten Annotationsrichtlinien. Eine unabhängige Adjudikatorin bzw. ein unabhängiger Adjudikator löst zum Ende noch etwaige Unstimmigkeiten auf, was in einer Endversion der Daten, dem Goldstandard, resultiert. Verglichen mit dem im Annotationsprozess entstandenen Goldstandard erreichen die Annotatorinnen und Annotatoren Recall-Werte von über 95%, wobei die individuelle Leistung offensichtlich sowohl von der Einarbeitungszeit abhängt als auch zwischen den zeitgleich eingearbeiteten Hilfskräften schwankt.

In einem letzten Schritt wurden alle erkannten Textstellen manuell durch realistische Pseudonyme ersetzt. Die manuelle Erstellung dieser pseudonymisierten Texte ist aufwändig, wird aber durch automatische Vorschläge unterstützt. Bei der Pseudonymisierung ist insbesondere darauf zu achten, dass wiederkehrende Entitäten mit den gleichen Pseudonymen ersetzt werden, auch wenn die Schreibweise bzw. Darstellung bei mehreren Vorkommen im Urteil unterschiedlich sein kann (beispielsweise verschiedene Namens- oder Datumsformate). Diese konsistente Handhabung ist essentiell, um die Lesbarkeit des pseudonymisierten Urteils zu gewährleisten.

[30] *Dévaud/Kummer*, in: Hürlimann/Kettiger, Anonymisierung von Urteilen, S. 61-70.

Die pseudonymisierten Datensätze dienen letztendlich als Trainings- und Evaluationsdaten für die automatische Anonymisierung mithilfe maschineller Lernverfahren. Die untenstehende Tabelle zeigt, dass mit einem Deep-Learning-Ansatz auf Basis des vortrainierten neuronalen Sprachmodells GottBERT[31] bereits 90% aller zu anonymisierenden Textstellen erkannt werden konnten, während NER-Standardwerkzeuge (NER = Named Entity Recognition), die speziell auf die Erkennung von Eigennamen gemünzt sind, nur einen kleinen Prozentsatz der relevanten Textstellen erkennen.

	Alle Textstellen			Nach Risiko		
System	Precision	Recall	F_1	*hoch*	*mittel*	*niedrig*
Standard-NER (Flair)	0.14	0.12	0.13	0.39	0.31	0.01
OpenNLP	**0.88**	0.80	0.84	0.85	0.45	0.83
Riedl & Padó	0.80	0.83	0.82	0.90	0.52	0.85
Fine-tuned GottBERT	0.80	**0.90**	**0.84**	**0.96**	0.80	0.89

Die Erkennung der zu anonymisierenden Stellen ist erfreulicherweise besonders zuverlässig für Textstellen mit hohem Deanonymisierungsrisiko (überwiegend direkte Identifikatoren wie Personennamen und Adressangaben), bei denen sogar schon ein Recall von 96% erzielt wurde. Das so trainierte Modell wurde auch mit gleichen Ergebnissen auf nicht pseudonymisierten Urteilen getestet, was die Validität des pseudonymisierten Goldstandards als Trainingsdaten nachweist. Im weiteren Verlauf unseres Projekts konnte – u.a. durch Vergrößerung der verfügbaren Trainingsdaten und verschiedene Optimierungen – der Recall insgesamt sogar bereits auf über 95% gesteigert werden.

VII. Fazit und Ausblick

Durch die dargestellte Möglichkeit zur Evaluation von automatischen Anonymisierungsverfahren können Aussagen über die flächendeckenden Einsatzmöglichkeiten solcher Verfahren in den Justizverwaltungen getroffen und verschiedene Systeme miteinander verglichen werden. Unsere bisherigen Evaluationsexperimente zeigen im Vergleich zu anderen Systemen bereits gute Ergebnisse unserer Anonymisierungsverfahren, insbesondere im Detektieren von Merkmalen mit hohem Risiko. Soweit ersichtlich, werden diese guten Ergebnisse bislang von keinem anderen System erreicht.

In der verbleibenden Projektphase wollen wir auch noch die dem Urteil zugehörigen und vorangegangenen Schriftsätze als Trainingsdaten mit einbeziehen und erhoffen uns dadurch eine weitere Performancesteigerung. Diese Hoffnung wird genährt durch die These, dass entsprechend dem hermeneutischen Ansatz[32] in der juristischen Methodenlehre, die Gerichte auch sonstige identifizierende Merkmale im Urteil vielfach aus den Schriftsätzen herauslesen. Kommen bestimmte Merkmale häufig in den Schriftsätzen vor, so kann daraus geschlossen werden, dass diese Merkmale auch identifizierende Wirkungen haben könnten und dass es sich um entscheidungserhebliche Tatsachen handeln könnte. Dann sollte das Tool die Möglichkeit eröffnen manuell einzugreifen um dem Richter, der Richterin die

[31] *Scheible/Thomczyk/Tippmann/Jaravine/Boeker*, GottBERT: a pure German language model, abrufbar unter https://arxiv.org/abs/2012.02110.

[32] *Adrian* Rechtstheorie 2017, 77 (102, 105, 109).

Entscheidung zu überlassen, ob eine Anonymisierung erfolgen soll oder nicht, um im Hinblick auf die Verständlichkeit der Entscheidung die Textstelle trotz allem zu erhalten. Daher arbeiten wir auch an der Erforschung eines für die Justizverwaltung optimal zugeschnittenen Frontends.

Schließlich wollen wir in den kommenden Monaten auch die Leistungsfähigkeit unseres automatischen Systems zur automatischen Anonymisierung und Pseudonymisierung anhand weiterer 1.600 OLG-Urteile verschiedenster Rechtsgebiete evaluieren.