# Automatische Anonymisierung von Gerichtsurteilen: Eine Vision scheint realisierbar

*unedited manuscript — please refer to the published paper for reliable citations*

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

*Seit über zwei Jahren arbeiten wir im Rahmen eines Forschungsprojektes mit dem Bayerischen Staatsministerium der Justiz an rechtlichen und technischen Fragen zur Möglichkeit einer automatischen Anonymisierung von Urteilen. Nach den bisherigen Ergebnissen scheint es nun im Bereich des Möglichen zu liegen, mit maschinellen Verfahren vollautomatisch Gerichtsurteile zu anonymisieren. Dabei erreichte Breite und Tiefe der maschinell durchgeführten Anonymisierung bietet einen Schutz vor Deanonymisierung, der mindestens so hoch erscheint wie die bisherige manuelle Anonymisierung in den Justizverwaltungen in Deutschland. Hier werden die bisherigen Ergebnisse und der von uns entwickelte Prototyp vorgestellt.*

## 1. Einleitung

Der vorliegende Beitrag stellt die bisherigen Ergebnisse unseres Forschungsprojekts zur Evaluation der Möglichkeiten einer automatischen Anonymisierung von Gerichtsentscheidungen mit dem Bayerischen Staatsministerium der Justiz dar. Das im April 2020 begonnene Forschungsprojekt an der Friedrich-Alexander-Universität Erlangen-Nürnberg wurde bereits auf der IRIS 2021 erstmals vorgestellt.[1] Mittlerweile hat die interdisziplinäre Projektarbeit einige zuversichtliche Forschungsergebnisse zu Tage gefördert und einen Software-Prototypen zur automatischen Anonymisierung hervorgebracht. Nach einem kurzen Abriss der Ziele in der Politik (2.), der bisherigen Veröffentlichungspraxis (3.) soll das Forschungsprojekt, mit der zugrundeliegenden Datenbasis und Methodik, der beteiligten Disziplin der Rechtswissenschaft sowie der Korpus- und Computerlinguistik und des derzeit schon funktionierenden Prototyps (4.) vorgestellt werden.

## 2. Ziele der Politik – Anonymisierung und Pseudonymisierung

Mittlerweile hat die Politik in Deutschland Ziele vorgegeben, die erfordern, dass die Aufgabe einer möglichst vollautomatischen Anonymisierung von Gerichtsentscheidungen durch maschinelle Verfahren erfüllt werden kann. Im Koalitionsvertrag der Koalitionsparteien in Deutschland von 2021 mit dem Titel „Mehr Fortschritt wagen“ wurde auf Seite 106 wörtlich ausgeführt: „Gerichtsentscheidungen sollen grundsätzlich in anonymisierter Form in einer Datenbank öffentlich und maschinenlesbar verfügbar sein.“ Damit wurde das Ziel formuliert, dass (möglichst alle?) Gerichtsentscheidungen zu veröffentlichen sind. Die Veröffentlichung kann mindestens zwei verschiedenen Zwecken dienen, was im Wort „maschinenlesbar“ zum Ausdruck kommt.

Zum einen kann der Zweck der Veröffentlichung schlicht Transparenz staatlichen Handelns sein, hier namentlich der Rechtsprechung, was auch den Zugang der Bürger und der Wissenschaft zum Recht verbessert. Dies ist zu begrüßen, weil jeder in einem gewaltenteiligen Rechtsstaat nachlesen können sollte, was vor den Gerichten in welcher Weise entschieden wird, um die Rechtsregeln zur Kenntnis nehmen und sein Verhalten darauf einrichten können sollte. Hierzu würde eine Anonymisierung in der Weise ausreichen, dass zu verbergende Textstellen weggelassen oder geschwärzt oder mit randomisierten Initialen maskiert werden. Im Idealfall sollte dabei die jeweilige Rolle der Textstelle für das Verständnis beibehalten werden, also man sollte noch erkennen können, wenn es sich z.B. um dieselbe Person bei der anonymisierten Nennung einer Initiale handelt.

---

[1] *ADRIAN/EVERT/KEUCHEN/HEINRICH/DYKES*, Anonymisierung von Gerichtsurteilen – Eine wesentliche Voraussetzung für E-Justice –. In: Schweighofer/Kummer/Saarenpää/Eder/Hanke (Hrsg.), Cybergovernance - Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS 2021, Bern 2021, S. 137–147.

Zum anderen kann damit auch bezweckt werden, künftig die veröffentlichten Gerichtsentscheidungen als Trainingsdaten für maschinelle Lernverfahren zu nutzen.[2] Diese Idee ist im Kontext von künstlicher Intelligenz (KI) nicht neu. Immer wieder wird eine freie Verfügbarkeit von Urteilen sowie anderer juristischer Daten, wie Schriftsätze, als eine wesentliche Voraussetzung für die Entwicklung komplexer Legal-Tech-Tools betont.[3] Zwar existieren in Deutschland schon ein paar wenige Leuchtturmprojekte, jedoch sind diese in Bezug auf breite Anwendbarkeit und Komplexität beschränkt. In Deutschland fehlen bis heute z.B. gut funktionierende Predictive Tools, wie diese in den USA bereits eingesetzt werden[4], oder gar Software zur automatischen Erstellung von Urteilsentwürfen auf Grundlage von Gerichtsentscheidungen.[5] Sobald Urteile auch als Trainingsdaten genutzt werden sollen, reicht eine „bloße" Anonymisierung nicht aus, da die maschinellen Lernverfahren aufgrund „realitätsferner" Schwärzung oder Maskierung in Form von Initialen sachwidrige Zusammenhänge erlernen würden, die in „echten" Urteilen nicht vorkommen. Notwendig ist eine realistische Pseudonymisierung.[6] Eine solche vollautomatisch mit maschinellen Verfahren zu erzeugen, ist uns noch nicht gelungen. Im Folgenden geht es daher um unseren derzeitigen Prototyp, der eine automatische Anonymisierung durchführt, bei der die sensiblen Textstellen durch randomisierte Initialen maskiert werden, wobei die jeweilige Information einer Textstelle erhalten bleibt. Wenn es sich z.B. um dieselbe Person handelt, werden immer dieselben Initialen verwendet.

Trainingsdaten, die aus realistisch pseudonymisierten Urteilen bestehen, würden künftig typischerweise für Mustererkennung eingesetzt werden, also für Verfahren sog. subsymbolischer KI. Hier soll vorsorglich erwähnt werden, dass die Finanzverwaltungen das wohl derzeit beste in Deutschland funktionierende „Legal-Tech-Tool" einsetzen. Nach § 155 Abs. 4 AO dürfen bereits ausschließlich automationsgestützte Steuerfestsetzungen erfolgen. Im Jahr 2019 wurden so bundesweit (ohne Nordrhein-Westfalen) 10% aller Einkommensteuerveranlagungen, also rund 2,2 Millionen Steuerbescheide, vollautomatisch durch maschinelle Verfahren erlassen.[7] Dies allerdings weniger mithilfe maschineller Lernverfahren und Mustererkennung, also subsymbolischer KI-Systeme, sondern indem Expertensysteme zum Einsatz kommen, also symbolische KI-Verfahren.[8]

## 3. Unverändert geringe Veröffentlichungspraxis bei Gerichtsentscheidungen

Derzeit ist nur ein sehr geringer Anteil der Gerichtsentscheidungen öffentlich frei zugänglich verfügbar. Eine neuste empirische Untersuchung hat gezeigt, dass für den Zeitraum 2011 bis 2020 in den öffentlich und kostenlos zugänglichen Rechtsprechungsportalen der 16 Bundesländer und des Bundes im Schnitt

---

2 *WAGNER*, Legal Tech und Legal Robots, 2. Aufl., Wiesbaden 2020, S. 25; *KEUCHEN/DEUBER*, Öffentlich zugängliche Rechtsprechung für Legal Tech – Eine rechtliche und empirische Betrachtung im Lichte des DNG – Teil 1, RDi 2022, S. 189 (S. 190).

3 *HARTUNG*, Quantitative legal research in Germany. In: Vogl (Ed.), Research Handbook on Big Data Law, Cheltenham/Northampton 2021, S. 228 (S. 231); *WERTHMANN*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Künstliche Intelligenz und Robotik, München 2020, § 22 Rn. 79 ff. Bereits im Beitrag *ADRIAN,* Der Richterautomat ist möglich – Semantik ist nur eine Illusion?, RECHTSTHEORIE 2017, S. 77 ff. wurde gefordert nicht nur Urteile, sondern grundsätzlich auch alle Schriftsätze, die historisch zum Erlass des jeweiligen Urteils führten, in die Datenbasis von Legal-Tech-Tools einzubeziehen. Dies ist mit Überlegungen aus der Methodenlehre, der Wissenschaftstheorie und der Rechtsphilosophie zu begründen, was in *ders.,* Grundprobleme einer juristischen (gemeinschaftsrechtlichen) Methodenlehre, Berlin 2009, *ders.,* Grundzüge einer allgemeinen Wissenschaftstheorie auch für Juristen, Berlin 2014, *ders.,* Wie wissenschaftlich ist die Rechtswissenschaft?, RECHTSTHEORIE 2010, S. 521 ff., dargelegt wurde.

4 *VOGL,* Changes in the US Legal Market Driven by Big Data/Predictive Analytics and Legal Platforms. In: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, Beck, München 2018, S. 53 ff; *ASHLEY,* Artificial Intelligence and legal Analytics, Cambridge University Press, 2017, S. 107, 109, 111, 125 f.; *FREUDENTHALER*, Case-based-Reasoning (CBR): Grundlagen und ausgewählte Anwendungsbeispiele des fallbasierten Schließens, AkademikerVerlag, Saarbrücken 2008; *BENCH-CAPON/SARTOR*, A Model of Legal Reasoning with Cases Incorporating Theories and Values, Artificial Intelligence 150, 2003, S. 97 ff.

5 Siehe zur prinzipiellen Möglichkeit dieser Vision *ADRIAN,* Der Richterautomat ist möglich – Semantik ist nur eine Illusion?, RECHTSTHEORIE 2017, S. 77 ff., wobei die juristische Methodenlehre nicht unmittelbar als Vorbild dienen kann, was in *ADRIAN,* Juristische Methodenlehre – Ein Vorbild für verantwortungsvolle Digitalisierung?. In: Schweighofer/Hötzendorfer/Kummer/Saarenpää (Hrsg.), Verantwortungsbewusste Digitalisierung, Tagungsband des 23. Internationalen Rechtsinformatik Symposiums IRIS 2020, Bern 2020, S. 41–48 dargelegt wurde.

6 Wir verwenden hier die Begriffe „Anonymisierung" und „Pseudonymisierung" nicht im Sinne der DSGVO. Unter Anonymisierung verstehen wir das bloße Unkenntlichmachen von identifizierenden Textstellen, selbst wenn diese durch eindeutige Bezeichner oder Initialen ersetzt werden, die mit Hilfe eines geheimen Schlüssels wieder den ursprünglichen natürlichen oder juristischen Personen zugeordnet werden können (was in der DSGVO-Terminologie eine Pseudonymisierung darstellen würde). Die von uns angestrebte realistische Pseudonymisierung verwendet hingegen grammatisch korrekte und natürlich klingende Ersetzungen, die aber nicht mehr den ursprünglichen Personen zugeordnet werden können (und somit im Sinne der DSGVO eine Anonymisierung darstellen).

7 Bundestagsdrucksache 19/19733, S. 9 Tz. 14.

8 *ADRIAN/BARTHEL*, Expertensysteme im Bereich der Steuerverwaltung – Vorbild bei der Realisierung eines künftigen digitalen Justizportals?, In: Adrian/Kohlhase/Evert/Zwickel (Hrsg.), Digitalisierung von Zivilprozess und Rechtsdurchsetzung, Berlin 2022, S. 101–115.

lediglich 2,3% der bereinigten Gerichtsentscheidungen veröffentlicht wurden.[9] Diese geringe Veröffentlichungsquote setzt einen seit Jahrzehnten bestehenden Trend fort.[10] In rechtlicher Hinsicht müsste aber viel mehr veröffentlicht werden, da in Deutschland die Rechtsprechung bereits 1997 eine Veröffentlichungspflicht von Entscheidungen aus dem Rechtsstaatsgebot einschließlich der Justizgewährungspflicht, dem Demokratiegebot und dem Grundsatz der Gewaltenteilung abgeleitet hat.[11] Gleichzeitig müssen die Interessen der Beteiligten und Betroffenen geschützt werden. In der damit verbundenen Notwendigkeit der Anonymisierung liegt der entscheidende Grund, dass die Urteile durch die Justizverwaltungen nur in geringer Zahl veröffentlicht werden, da die Anonymisierung aktuell immer noch aufwendig, uneinheitlich und insbesondere manuell erfolgt.[12]

## 4. Forschungsprojekt zur automatischen Anonymisierung von Gerichtsentscheidungen

Um die rechtlichen und technischen Probleme der Anonymisierungsaufgabe zu lösen, forschen wir insbesondere mit Hilfe korpus- und computerlinguistischer Verfahren für das Bayerische Staatsministerium der Justiz seit April 2020 an der Möglichkeit einer im Idealfall vollautomatischen[13] Anonymisierung von Urteilen. Ein erstes Ziel des Projekts, nämlich die Entwicklung und detaillierte Evaluation eines Software-Prototyp für die automatische Erkennung sensibler Textstellen, wurde erreicht. Hierfür wurden aufwendig Trainingsdaten erstellt, indem eine manuelle Kennzeichnung dieser Stellen (Annotation und Adjudikation) in Urteilen und Schriftsätzen erfolgte, die auch in Kategorien eingeteilt und bezüglich ihres Risikoniveaus bewertet wurden. Es wurde ein Goldstandard erstellt, indem jeder einzelne Text von sechs verschiedenen Bearbeiter:innen analysiert wurde. Wenn auch noch nicht alle Schriftsätze zu den Urteilen annotiert wurden, so konnte ein funktionsfähiger Prototyp erstellt und evaluiert werden, der eine zielgerichtete informationserhaltende vollautomatische Anonymisierung von Urteilen von Amtsgerichten im Wohnraummiet- und Verkehrsrecht mit guten Ergebnissen durchführt. Ein Vergleich des Prototyps mit anderen Tools zur Anonymisierung zeigt, dass die Ergebnisse bislang von keinem anderen kommerziellen oder wissenschaftlichen Projekt erreicht werden.[14]

### 4.1. Recht und Anonymisierung – Deanonymisierungsexperimente

Wie die Judikative eine Veröffentlichung für obligatorisch erachtet, so verlangt sie ebenso den Schutz von den berechtigten Interessen der Beteiligten und Betroffenen.[15] Eine indirekte Pflicht zur Anonymisierung ergibt sich aus einer Vielzahl an Rechtsvorschriften und -positionen, die berührt sein können, wenn die Informationen aus einer Entscheidung ungehindert veröffentlicht werden würden.[16] Zu solchen Schutzvorschriften gehören bspw. das Recht auf informationelle Selbstbestimmung (Art. 2 Abs. 2 GG i.V.m. Art. 1 Abs. 1 GG), das Unternehmenspersönlichkeitsrecht, die Betriebs-, Geschäfts-, Steuer-, und Sozialgeheimnisse (§ 30 AO und § 35 SGB I) und der allgemeine Datenschutz.[17] Dem Schutz unterliegen natürliche und juristische Personen.[18]

---

[9] *KEUCHEN/DEUBER*, Öffentlich zugängliche Rechtsprechung für Legal Tech – Eine rechtliche und empirische Betrachtung im Lichte des DNG – Teil 2, RDi 2022, S. 229 (S. 230 f.).

[10] *KUNTZ*, Quantität gerichtlicher Entscheidungen als Qualitätskriterium juristischer Datenbanken, JurPC Web-Dok. 12/2006, Abs. 34: dort ist die Rede von 0,27 – 4,95% für die Zeit von 2000 bis 2004. Eine Zusammenfassung zu noch älteren Untersuchungen findet sich bei *KEUCHEN/DEUBER*, Öffentlich zugängliche Rechtsprechung für Legal Tech – Eine rechtliche und empirische Betrachtung im Lichte des DNG – Teil 2, RDi 2022, S. 229 (S. 233).

[11] BVerwG 26. Februar 1997 - 6 C 3.96; BGH 05. April 2017 - IV AR (VZ) 2/16; BVerfG 14. September 2015 – 1 BvR 857/15.

[12] *HARTUNG*, Quantitative legal research in Germany. In: Vogl (Ed.), Research Handbook on Big Data Law, Cheltenham/Northampton 2021, S. 228 (S. 233); *ADRIAN/DYKES/EVERT/HEINRICH/KEUCHEN/PROISL*, Manuelle und automatische Anonymisierung von Urteilen. In: Adrian/Kohlhase/Evert/Zwickel (Hrsg.), Digitalisierung von Zivilprozess und Rechtsdurchsetzung, Berlin 2022, S. 173 (S. 176).

[13] Bisherige Software Lösungen zur (semi)automatisierten Anonymisierung erfordern zumeist eine exakte manuelle Eingabe der kritischen Textinformation sowie des Pseudonyms. Hierzu mit Lösungsbeispiel: *DÉVAUD/KUMMER*, (Semi-)Automatische Anonymisierung von Entscheiden, Jusletter IT 23. Februar 2017, S. 2 ff.

[14] ADRIAN/DYKES/EVERT/HEINRICH/KEUCHEN, Entwicklung und Evaluation automatischer Verfahren zur Anonymisierung von Gerichtsentscheidungen, im Erscheinen in LTZ 2022 Heft 4.

[15] BVerwG 26. Februar 1997 - 6 C 3.96; VGH Mannheim 10. Juli 2020 - 2 S 623/20; OLG Frankfurt a.M. 19. September 2019 – 20 VA 21/17.

[16] So bereits *ADRIAN/EVERT/KEUCHEN/HEINRICH/DYKES*, Anonymisierung von Gerichtsurteilen – Eine wesentliche Voraussetzung für E-Justice –. In: Schweighofer/Kummer/Saarenpää/Eder/Hanke (Hrsg.), Cybergovernance - Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS 2021, Bern 2021, S. 142 ff.

[17] Vgl. so auch zur Schweiz *BIERI*, Das Handwerk der Urteilsanonymisierung. In: Hürlimann/Kettiger (Hrsg.), Anonymisierung von Urteilen, Basel 2021, S. 1 (S. 5).

[18] *NÖHRE*, Anonymisierung und Neutralisierung von veröffentlichungswürdigen Gerichtsentscheidungen, MDR 2019, S. 136 (S. 137 f.).

Um zu bestimmen, wann ein Text als anonym gilt, kann auf das Konzept der DSGVO zurückgegriffen werden. Notwendig ist die Beseitigung des Personenbezugs, durch Beseitigung direkter (Name, Anschrift etc.) wie indirekter (deskriptive Merkmale zu Beruf, Gesundheit etc.) Identifikatoren einer Person.[19] Ob dies der Fall ist, ist eine normative Fragestellung, die aus dem Blickwinkel der Gefahr einer Deanonymisierung zu beantworten ist. Eine solche Gefahr muss nicht absolut ausgeschlossen sein, aber eine Deanonymisierung darf nur mit unverhältnismäßig großem Aufwand an Zeit, Kosten und Arbeitskraft sowie unter Heranziehung von Zusatzwissen aus allgemein zugänglichen Quellen möglich sein.[20] Dieses Maß der Anonymität, das letztendlich die Wahrscheinlichkeit der Identifizierbarkeit beschreibt,[21] lässt sich vergleichbar in Erwägungsgrund 26 DSGVO wiederfinden. Demnach sind bei der Feststellung der direkten wie indirekten Identifizierbarkeit einer Person alle Mittel, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, zu berücksichtigen. Bei der Feststellung, ob Mittel nach allgemeinem Ermessen wahrscheinlich zur Identifizierung der natürlichen Person genutzt werden, sollen alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand herangezogen werden, wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und zu erwartende technologische Entwicklung einzubeziehen sind.[22]

Das allgemein zugängliche oder nach bestimmten rechtlichen Maßnahmen legal verfügbare Zusatzwissen[23] dient dem Angreifer, um Überschneidungen in den Informationen zu entdecken, damit die Personen unter möglichst sicherem Ausschluss von Doppelgängern mit einer hinreichenden Sicherheit deanonymisiert werden können.[24] Um diese Gefahr und normative Parameter zu erfassen, sind empirische Daten von Nöten, welche die tatsächlichen Gefahren und Risiken mit dem rechtlichen Maß des „unverhältnismäßig großen Aufwands“ aufzeigen.[25] Damit bestehen gewisse Unsicherheiten hinsichtlich des Deanonymisierungsrisikos und folglich der Identifizierbarkeit und Anonymität.[26]

Neuere empirische Studien zeigen, wie angreifbar der aktuelle Standard der Anonymisierung ist, da bereits mit äußerst geringfügigen – also nicht unverhältnismäßigem – Aufwand die Anonymität aufgehoben werden kann. Drei Ursachen lassen sich herauskristallisieren. Erstens sind herkömmliche Anonymisierungstechniken wie die Verwendung von Initialen höchst anfällig.[27] Zumindest sollte eine Randomisierung der Initialen vorgenommen werden.[28] Zweitens werden regelmäßig zu wenige – insbesondere indirekte – Identifikationsmerkmale anonymisiert, so dass mittels *cross-referencing* bzw. *linkage* eine Zusammenführung von mehreren indirekten Identifikatoren und schlussendlich eine direkte Identifikation möglich ist.[29] Drittens nimmt das frei verfügbare Wissen enorm zu, so dass hier leicht eine

---

[19] *BIERI*, Das Handwerk der Urteilsanonymisierung. In: Hürlimann/Kettiger (Hrsg.), Anonymisierung von Urteilen, Basel 2021, S. 1 (S. 7–9).

[20] OLG Karlsruhe 22. Dezember 2020 - 6 VA 24/20; VGH Baden-Württemberg 23. Juli 2010 - 1 S 501/10. Vgl. so auch zur Schweiz *BIERI*, Das Handwerk der Urteilsanonymisierung. In: Hürlimann/Kettiger (Hrsg.), Anonymisierung von Urteilen, Basel 2021, S. 1 (S. 5 f.).

[21] *RIEDL*, De-Anonymisierung als grundsätzliche, nicht ausschliessbare Option. In: Hürlimann, Daniel/Kettiger, Daniel (Hrsg.), Anonymisierung von Urteilen, Basel 2021, S. 31 (S. 43).

[22] Vgl. so auch Art. 29-Datenschutzgruppe, Stellungnahme 5/2014 zu Anonymisierungstechniken, 10. April 2014, Aktenzeichen 082914 0829/14/DE WP216, S. 6.

[23] So ist ebenso zum Zusatzwissen einzubeziehen, ob mit rechtlichen Mitteln das Wissen erlangt werden kann. Vgl. zu IP-Adressen EuGH 19. Oktober 2016 - C 582/14; BGH 16. Mai 2017 - VI ZR 135/13.

[24] *ADRIAN/DYKES/EVERT/HEINRICH/KEUCHEN/PROISL*, Manuelle und automatische Anonymisierung von Urteilen. In: Adrian/Kohlhase/Evert/Zwickel (Hrsg.), Digitalisierung von Zivilprozess und Rechtsdurchsetzung, Berlin 2022, S. 173 (S. 177 f.).

[25] So bereits die Überlegung *HORNUNG/WAGNER*, Der schleichende Personenbezug, CR 2019, S. 565 (S. 574); *ADRIAN/EVERT/KEUCHEN/HEINRICH/DYKES*, Anonymisierung von Gerichtsurteilen – Eine wesentliche Voraussetzung für E-Justice –. In: Schweighofer/Kummer/Saarenpää/Eder/Hanke (Hrsg.), Cybergovernance - Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS 2021, Bern 2021, S. 137–147. *DEUBER/KEUCHEN/CHRISTIN*, Assessing Anonymity Techniques Employed in German Court Decisions: A De-Anonymization Experiment, erscheint in: 32nd USENIX Security Symposium, 2023

[26] *RIEDL*, De-Anonymisierung als grundsätzliche, nicht ausschliessbare Option. In: Hürlimann, Daniel/Kettiger, Daniel (Hrsg.), Anonymisierung von Urteilen, Basel 2021, S. 31 (S. 48 f.).

[27] *DEUBER/KEUCHEN/CHRISTIN*, Assessing Anonymity Techniques Employed in German Court Decisions: A De-Anonymization Experiment, erscheint in: 32nd USENIX Security Symposium, 2023. Allgemein *KARG*, Anonymität, Pseudonyme und Personenbezug revisited?, DuD 2015, S. 520.

[28] *ADRIAN/DYKES/EVERT/HEINRICH/KEUCHEN/PROISL*, Manuelle und automatische Anonymisierung von Urteilen. In: Adrian/Kohlhase/Evert/Zwickel (Hrsg.), Digitalisierung von Zivilprozess und Rechtsdurchsetzung, Berlin 2022, S. 173 (S. 178).

[29] *DEUBER/KEUCHEN/CHRISTIN*, Assessing Anonymity Techniques Employed in German Court Decisions: A De-Anonymization Experiment, erscheint in: 32nd USENIX Security Symposium, 2023; *VOKINGER/MÜHLEMATTER*, Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(banken), Jusletter 2. September 2019, S. 16; *SCHILD*, in: Beck`scher Online-Kommentar Datenschutzrecht, Wolff/Brink (Hrsg.), 41. Edition 01.08.2022, Grundlagen und bereichsspezifischer Datenschutz, Syst. E., Rn. 58.

Verknüpfung mit Medienberichten, anderen Gerichtsentscheidungen oder offiziellen Datenbanken möglich ist, da diese ebenso wenig konsistent anonymisiert sind.[30]

Mit Blick auf die Anforderungen an die Anonymisierungsleistung eines Tools zur automatischen Anonymisierung kann aus den diversen rechtlichen Regelungen einerseits und den in der Rechtspraxis festgestellten Deanonymisierungsgefahren andererseits folgendes festgehalten werden. Es müssen sowohl direkte wie indirekte Identifikatoren zuverlässig erkannt und anonymisiert werden. Wegen der Deanonymisierungsgefahren und der uneingeschränkten Pflicht zum Schutz der personenbezogenen Daten muss sichergestellt sein, dass die Fehlerrate verschwindend klein ist, bevor im großen Umfang Entscheidungen publiziert werden. Zu bedenken ist, dass selbst eine Fehlerrate von nur 1% bei ca. 1,6 Millionen veröffentlichungsfähigen Entscheidungen in einem Jahr[31] zu 16.000 unzureichend anonymisierten Urteilen führen könnte und Rechtsverletzungen in dieser Zahl pro Jahr im Raum stehen. Insgesamt führt dies zu sehr hohen Anforderungen an die Qualität und Leistungsfähigkeit eines Tools zur Anonymisierung.

### 4.2. Erstellung eines Goldstandards

Für die Entwicklung maschineller Lernverfahren zur automatischen Anonymisierung sowie zur Evaluation jeglicher Anonymisierungsverfahren (manueller oder automatischer Natur) wird ein *Goldstandard* benötigt, d.h. ein Korpus von Urteilen, in dem alle zu anonymisierenden Stellen korrekt markiert wurden. Hierfür beschäftigen wir studentische Hilfskräfte, deren Aufgabe es ist, unabhängig voneinander die sensiblen Stellen zu identifizieren und zusätzlich den Grund für die Anonymisierung sowie das Risikoniveau der Stelle zu vermerken.[32] Zuvor werden die Hilfskräfte in mehreren Probeläufen auf dem vorhandenen Goldstandard geschult.

Das Datenmaterial umfasst im *Kernkorpus* 570 Urteile mit einem Gesamtumfang von knapp 1 Million Token:

- 247 Urteile aus 227 Akten zum Mietrecht, insgesamt 399.822 Token (ca. 1669 Token/Urteil),
- 323 Urteile aus 320 Akten zum Verkehrsrecht, insgesamt 552.906 Token (ca. 1782 Token/Urteil).

Schwankungen in der Länge der Urteile sind v.a. durch die unterschiedliche Länge des Tatbestands und der Entscheidungsgründe bedingt. Das strukturell über alle Urteile hinweg sehr ähnliche Rubrum umfasst typischerweise knapp 115 Token, die Rechtsbehelfsbelehrung ca. 385 Token.

Alle Urteile wurden am Anfang des Projekts von mindestens vier unterschiedlichen Annotator:innen unabhängig voneinander annotiert. Etwaige Unstimmigkeiten wurden in einem separaten Durchgang von einer weiteren Hilfskraft aufgelöst (ein Vorgang, der sich Adjudikation nennt). Das Resultat ist der Goldstandard, welcher zur Bewertung der einzelnen Annotator:innen wie auch für Training und Evaluation maschineller Verfahren zur Erkennung von sensiblen Stellen genutzt werden kann.[33] Alle Urteile wurden ebenfalls in einem iterativen Prozess vollständig, realistisch und sinnerhaltend pseudonymisiert. Wie bereits an anderer Stelle bemerkt[34], ist die Übereinstimmung der Annotator:innen untereinander bzgl. der meisten Kategorien sehr hoch; lediglich bei den sonstigen identifizierenden Merkmalen findet sich – wie zu erwarten – ein hohes Maß an Subjektivität.

---

[30] *DEUBER/KEUCHEN/CHRISTIN*, Assessing Anonymity Techniques Employed in German Court Decisions: A De-Anonymization Experiment, erscheint in: 32nd USENIX Security Symposium, 2023.

[31] Die Zahl ist entnommen aus *KEUCHEN/DEUBER*, Öffentlich zugängliche Rechtsprechung für Legal Tech – Eine rechtliche und empirische Betrachtung im Lichte des DNG – Teil 2, RDi 2022, S. 229 (S. 233).

[32] Dieses Vorgehen wird in der Computerlinguistik als „Tagging" bezeichnet; die zuweisbaren Begründungskategorien werden in diesem Zusammenhang „Tags" genannt. Das Tagset mit derzeit ca. 20 Tags in sechs Oberkategorien und entsprechende Annotationsrichtlinien werden von uns stetig iterativ weiterentwickelt.

[33] *HEINRICH/DYKES/EVERT*, Annotator agreement in the anonymization of court decisions, Corpus Linguistics 2021, zeigt: Vier bis sechs unabhängige Annotator:innen sind ausreichend, um einen Text mit 1 Million Token nahezu perfekt zu anonymisieren: ein:e weitere:r Annotator:in würde im Schnitt weniger als eine weitere sensible Textstelle erkennen.

[34] *ADRIAN/EVERT/KEUCHEN/HEINRICH/DYKES*, Anonymisierung von Gerichtsurteilen – Eine wesentliche Voraussetzung für E-Justice –. In: Schweighofer/Kummer/Saarenpää/Eder/Hanke (Hrsg.), Cybergovernance - Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS 2021, Bern 2021, S. 137–147.

### 4.3. Überblick und Evaluation automatischer Anonymisierungsverfahren

Im Verlauf des Forschungsprojekts wurden zunächst verschiedene Ansätze zur automatischen Anonymisierung anhand einer Vorabversion des Goldstandards evaluiert. Hierzu wurde eine zu diesem Zeitpunkt bereits pseudonymisierte Teilmenge der Mietrecht-Urteile herangezogen.[35] Der beste Ansatz wurde schließlich noch auf den nicht pseudonymisierten Originaldaten (Wohnraummiet- sowie Verkehrsrecht) evaluiert.

Wir verwenden die bei Klassifikationsaufgaben üblichen Evaluationsmaße *Precision* und *Recall*. Der Recall *R* quantifiziert die „Trefferquote" des Systems, d.h. den Anteil der zu anonymisierenden Textstellen im Goldstandard, der vom automatischen System gefunden wurde. Ein hoher Recall bedeutet, dass das System den Großteil der zu anonymisierenden Textstellen gefunden hat, ein niedriger Recall bedeutet, dass viele relevante Textstellen vom System übersehen wurden. Die Precision *P* quantifiziert die „Genauigkeit" des Systems, d.h. welcher Anteil der vom System gefundenen Textstellen tatsächlich zu anonymisieren ist. Eine hohe Precision bedeutet, dass ein Großteil dieser Textstellen zurecht vom System vorgeschlagen wurde, eine niedrige Precision bedeutet, dass das System viele irrelevante Textstellen zur Anonymisierung auswählt (und damit bspw. zu viele Textstellen geschwärzt werden). Recall und Precision hängen voneinander ab: Eine hohe Precision wird in der Regel durch niedrigeren Recall erkauft und umgekehrt. In unserem Fall ist besonders ein hoher Recall wichtig. Automatische Systeme sollen idealerweise alle zu anonymisierenden Textstellen finden; dafür können niedrigere Precision-Werte, d.h. unnötigerweise zur Anonymisierung markierte Textstellen, in Kauf genommen werden. Der sogenannte $F_1$-Wert kombiniert Precision und Recall in einen Globalwert (das harmonische Mittel aus Precision und Recall), der oft zum Ranking unterschiedlicher Systeme verwendet wird.Es ist zudem anzumerken, dass wir hier eine strenge Evaluation zeigen: eine zu anonymisierende Textstelle gilt nur als korrekter Treffer, wenn sie *exakt* gefunden wurde, d.h. mit ihrem genauen Anfang und Ende. Wird auch nur ein Token zu viel oder zu wenig erkannt, wird die Textstelle nicht als korrekt gewertet.

Da es sich bei direkten Identifikatoren der Hochrisiko-Kategorie in den Urteilen hauptsächlich um Namen natürlicher und juristischer Personen sowie Adressen handelt, ist es naheliegend, sog. *Named Entity Recognizer* (NER) einzusetzen. Diese Tools werden für verschiedene Sprachen „Off-the-Shelf", d.h. direkt einsatzbereit, angeboten. Sie haben allerdings zwei Schwachstellen: Erstens entspricht nur eine Teilmenge der zu anonymisierenden Textstellen – Namen natürlicher und juristischer Personen, Ortsangaben sowie Datumsangaben – den von solchen Modellen annotierten *Named Entities*. Zweitens fassen Off-the-Shelf-Modelle die Grenzen der annotierten Textstellen oft enger als unsere Annotationsrichtlinien. So erzielt Flair[36], das zu den aktuell besten verfügbaren NER-Modellen für deutsche Texte gehört, nur enttäuschende Ergebnisse auf den pseudonymisierten Mietrechturteilen (erste Zeile in Tabelle 1). Auch eine für juristische Texte angepasste Version erreicht nur leicht bessere Werte. Schließlich wurde das Open-Source-Tool OpenRedact[37] evaluiert, das für eine semiautomatische Anonymisierung u.a. deutscher Gerichtsurteile entwickelt wurde und speziell angepasste NER-Standardwerkzeuge mit regelbasierten Verfahren kombiniert. Auch dieses Tool erreicht in unserer Evaluation mit 81% Recall noch keine für eine vollautomatische Anonymisierung akzeptable Zuverlässigkeit.

Ein zweiter Standardansatz der Computerlinguistik ist, maschinelle Lernverfahren gezielt für die Aufgabenstellung der Anonymisierung zu trainieren. Hierfür wurden sog. NER-Tagger speziell auf unserem pseudonymisierten Goldstandard trainiert. Der Goldstandard wurde dazu zufällig in Trainings- und Testdaten aufgeteilt. Wir führten Experimente mit zwei Werkzeugen durch, die auf traditionellen maschinellen Lernverfahren (CRF = Conditional Random Fields in OpenNLP[38]) bzw. auf Deep Learning (Sequenztagger von Riedl und Padó[39]) beruhen. Damit konnten merklich bessere Evaluationsergebnisse erreicht werden, die aber mit einem Recall von maximal 83% immer noch unzureichend für eine automatische Anonymisierung sind.

---

[35] Experimente sind leichter auf den pseudonymisierten Daten durchzuführen, da diese keine sensiblen Stellen mehr enthalten und daher den geschützten Raum verlassen dürfen.

[36] https://github.com/flairNLP/flair (alle Internetquellen zuletzt abgerufen am 29.10.2022)

[37] https://openredact.org

[38] https://opennlp.apache.org

[39] *RIEDL/PADÓ*, A named entity recognition shootout for German, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 2), 2018, S. 120–125.

| | Alle Textstellen | | | Recall nach Risikoniveau | | |
|---|---|---|---|---|---|---|
| System | Precision | Recall | $F_1$ | *hoch* | *mittel* | *niedrig* |
| Standard-NER (Flair) | 0.14 | 0.12 | 0.13 | 0.39 | 0.31 | 0.01 |
| Legal-NER (Flair) | 0.26 | 0.16 | 0.19 | 0.42 | 0.28 | 0.05 |
| OpenRedact | 0.49 | 0.81 | 0.61 | 0.87 | 0.82 | 0.78 |
| 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 |

**Tabelle 1: Evaluation der korrekten Erkennung von Textstellen**

In jüngster Zeit sind einige computerlinguistische Publikationen erschienen, bei denen vortrainierte neuronale Sprachmodelle erfolgreich auf komplexe Tagging-Aufgaben angewendet wurden, selbst wenn für die Zielaufgabe nur relativ kleine Mengen an Trainingsdaten verfügbar waren. Angesichts dieser aktuellen Entwicklung führten wir Experimente mit dem vortrainierten neuronalen Sprachmodellen GottBERT[40] durch. Das per Fine-Tuning an die Anonymisierungsaufgabe angepasste GottBERT-Modell wurde mit einem zusätzlichen Layer als NER-Tagger trainiert und erzielte in der Evaluation auf den pseudonymisierten Urteilen zum Mietrecht die mit Abstand besten Ergebnisse. So lässt ein Recall von 90% eine vollautomatische Anonymisierung in greifbare Nähe rücken. Textstellen mit hohem Risiko, also direkte Identifikatoren, sind damit sogar bereits zu 96% abgedeckt. Dieser hohe Recall wurde allerdings durch Abstriche bei der Precision erkauft, die lediglich 80% beträgt. Das CRF-Modell (OpenNLP) wählte eine andere Balance mit 88% Precision, so dass beide Ansätze insgesamt die gleiche F-Score von 84% erzielten. Da für unsere Zwecke der Recall eine wesentlich wichtigere Rolle spielt, wurde der GottBERT-Ansatz als guter Ausgangspunkt für die weitere Forschungsarbeit gewählt.

Eine abschließende Evaluation des auf pseudonymisierten Urteilen trainierten GottBERT-Modells zeigte einen unverändert hohen Recall von 90% sowohl auf weiteren, nicht pseudonymiserten Urteilen wie auch auf den Urteilen zum Verkehrsrecht in unserem Goldstandard. Dies belegt einerseits, dass Training und Evaluation auf realistisch pseudonymisierten Urteilen einen validen Ansatz darstellt. Andererseits weckt es die Hoffnung, dass automatische Anonymisierungsverfahren zu einem gewissen Grad auch rechtsgebietsübergreifend eingesetzt werden können.

### 4.4. Fine-Tuning eines neuronalen Sprachmodells

Ausgehend von diesen vielversprechenden Ergebnissen führten wir im weiteren Verlauf des Projekts umfangreiche Experimente mit dem Fine-Tuning neuronaler Sprachmodelle auf Basis des pseudonymisierten Kernkorpus von 570 pseudonymisierten Urteilen durch. Unter anderem wurden verschiedene vortrainierte neuronale Sprachmodelle für deutsche Texte als Alternativen zu GottBERT erprobt und schließlich das Modell gbert-base[41] ausgewählt.

Zu diesem Zweck wurde das Korpus zufällig in 50% Trainingsdaten, 25% Development-Daten (für die Optimierung der Modelle) und 25% Testdaten (für die hier dargestellte abschließende Evaluation) aufgeteilt. Damit standen insgesamt ca. 474.000 Token Trainingsdaten zur Verfügung sowie ca. 239.000 Token Testdaten. Die Evaluation erfolgte wieder streng auf Ebene ganzer Textstellen, aber wie im englischsprachigen Text Anonymization Benchmark[42] werden dabei kleine Abweichungen zugelassen: Wird statt *Adrian* die größere Textstelle *Zeuge Adrian* vom Modell gefunden, so gilt diese für die Recall-Berechnung als korrekt (da die sensible Information tatsächlich anonymisiert wurde), nicht aber für die

---

[40] *SCHEIBLE/THOMCZYK/TIPPMANN/JARAVINE/BOEKER*, GottBERT: a pure German language model, 2020, abrufbar unter: https://arxiv.org/abs/2012.02110.

[41] https://huggingface.co/deepset/gbert-base; siehe auch *CHAN/SCHWETER/MÖLLER,* German's next language model, Proceedings of the 28th International Conference on Computational Linguistics, 2020, S. 6788–6796.

[42] *PILÁN/LISON/ØVRELID/PAPADOPOULOU/SÁNCHEZ/BATET*, The Text Anonymization Benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. Computational Linguistics, 2022.

Precision-Berechnung. Umgekehrt wird *Kellerabteil* statt *Kellerabteil Nr. 44* für Precision als korrekt gezählt (da keine überflüssigen Token anonymisiert wurden), nicht aber für Recall. Zusätzlich zur Erkennung aller zu anonymisierender Textstellen wurden Modelle für die Identifikation der jeweiligen Informationskategorie und die Vorhersage des Risikoniveaus trainiert und schließlich in ein kombiniertes Modell zusammengeführt.

Insgesamt erreicht das kombinierte Modell einen Recall von $\boldsymbol{R = 96{,}91\%}$ bei einer Precision von $\boldsymbol{P = 96{,}60\%}$, was insgesamt einen F-Wert von $\boldsymbol{F_1 = 96{,}75\%}$ ergibt. Es konnte gegenüber den obigen Ergebnissen also eine erhebliche Verbesserung des Recalls um fast 7 Prozentpunkte erzielt werden. Die Precision stieg sogar um mehr als 16 Prozentpunkte an und ist gleichauf mit Recall, was die praktische Einsatzfähigkeit verbessert. Von den Hochrisikostellen im Goldstandard werden sogar **98,85%** korrekt erkannt. Besonders hohe Anonymisierungsqualität wird dabei erwartungsgemäß im Rubrum und in der Rechtsbehelfsbelehrung erreicht (Recall fast 100%), während insbesondere Tatbestand (durchschnittlicher Recall ca. 92% im Mietrecht und ca. 95% im Verkehrsrecht) und Entscheidungsgründe (Recall ca. 94% im Miet- sowie Verkehrsrecht) mit ihrer hohen Anzahl sonstiger identifizierender Merkmale deutlich schwieriger zu anonymisieren sind.

Wir zeigen die Evaluationsergebnisse für die automatische Erkennung der einzelnen Informationskategorien in Form einer Konfusionsmatrix (Abbildung 1), welche die vorhergesagte Informationskategorie (*Predicted*) mit der tatsächlichen Kategorie (*Gold*) vergleicht. Textstellen, die vom Modell nicht erkannt wurden, sind als FN abgekürzt, falsch vorgeschlagene Textstellen als FP. ANY steht für wenige Textstellen, bei denen das kombinierte Modelle keine Informationskategorie vorhersagen konnte. Die stärker umrandeten Felder entlang der Diagonale entsprechen den *True Positives*, also korrekt erkannten und klassifizierten Textstellen.

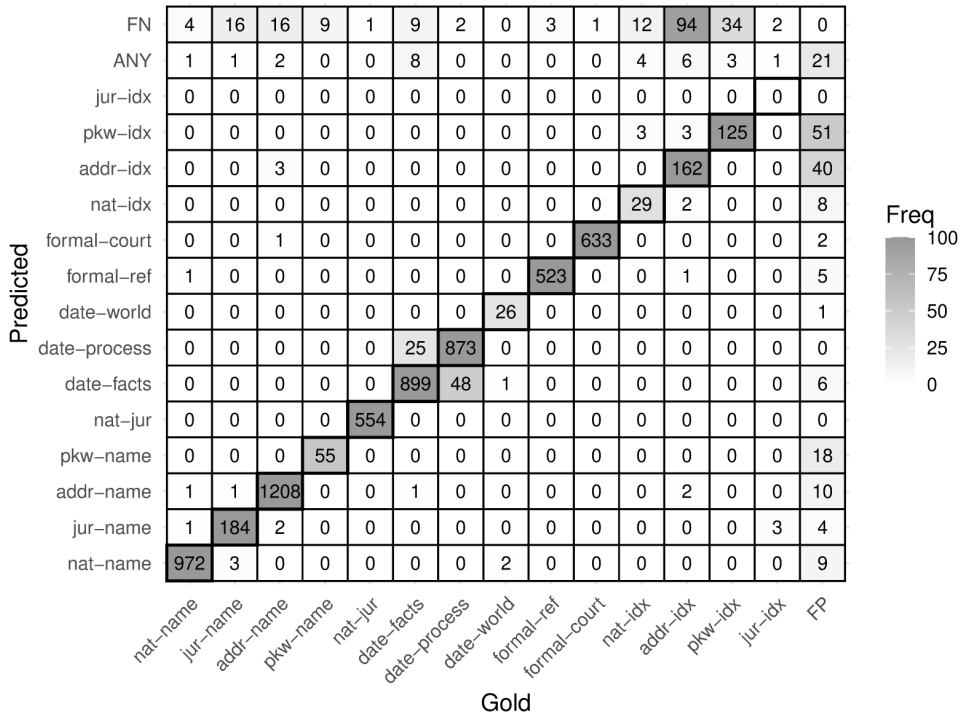


**Abbildung 1: Konfusionsmatrix des kombinierten Modells**

Die Konfusionsmatrix zeigt, dass Namen natürlicher und juristischer Personen (*nat-name*, *jur-name*), Justizangehörige (*nat-jur*), Aktenzeichen (*formal-ref*) und Gerichtsort (*formal-court*) schon fast perfekt erkannt werden (kaum Einträge jenseits der Diagonale). Wie erwartet besteht erheblicher Verbesserungsbedarf bei sonstigen identifizierenden Merkmalen – bspw. 94 FN bei den identifizierenden Merkmalen einer Adresse (*addr-idx*). Bei der relativ großen Anzahl von Fehlklassifikationen für Datumsangaben handelt es sich überwiegend um harmlose Verwechslungen zwischen den Unterkategorien (25 Angaben zum Sachverhalt wurden als Angaben zum Prozess klassifiziert, 48 Angaben zum Prozess als Angaben zum Sachverhalt). Auch einige andere Verwechslungen – z.B. zwischen verschiedenen Arten

identifizierender Merkmale – sind eher harmlos. Diese Beobachtungen legen nahe, dass auch die gezielte Anonymisierung von bestimmten, durch Anwender:innen ausgewählten Informationskategorien ohne wesentliche Qualitätseinbußen realisiert werden kann.

## 5. Zusammenfassung und Ausblick

Mit einem Recall von insgesamt knapp 97% (und sogar knapp 99% auf Hochrisikostellen) anonymisiert der von uns entwickelte Prototyp in einer Art und Weise, die eine vollautomatische Anonymisierung von Urteilen realistisch erscheinen lässt. Tatsächlich ist der Prototyp bereits zuverlässiger als geschulte Annotator:innen, besonders wenn nur eine Person eingesetzt wird. Dennoch stellt sich die Frage: Kann die Lücke zu 100% Recall geschlossen werden? Eine Möglichkeit besteht in der Hinzunahme von weiterem Datenmaterial. Da uns auch die zu den Urteilen zugehörigen Schriftsätze zur Verfügung stehen, können diese nach Digitalisierung und Annotation nutzbar gemacht werden. Für das Mietrecht wurden dazu 1172 Schriftsätze aus 213 Akten im Gesamtumfang von 1,1 Millionen Token annotiert (aufgrund von OCR-Problemen bspw. bei handschriftlich verfassten Dokumenten konnten nicht alle Schriftsätze verwendet werden). Zudem stehen uns 1348 unannotierte Schriftsätze im Umfang von knapp 1,5 Millionen Token aus dem Verkehrsrecht zur Verfügung. Die einfachste Herangehensweise ist nun, die annotierten Schriftsätze als zusätzliche Trainingsdaten zu nutzen. Auf diese Weise konnten wir den Recall beim Mietrecht um ca. 1,5 Prozentpunkte erhöhen – und dies bei gleichzeitgier Erhörung der Precision um knapp 1 Prozentpunkt. Allerdings ist dies offensichtlich mit viel manueller Arbeit verbunden. Eine weitere Möglichkeit ist die Zuhilfenahme von unannotierten Daten. Bei dieser Möglichkeit kann bspw. der Kontext für jeden Satz aus dem Urteil um ähnliche Sätze aus den Schriftsätzen erweitert werden; die Satzähnlichkeit wird dabei mittels Sentence-BERT bestimmt.[43] Auch in diesem Setting konnte der Recall verbessert werden; wie zu erwarten allerdings in einem geringeren Umfang von ca. 0,5 Prozentpunkten. Hier ist noch einiges an Forschungsarbeit notwendig, um ein nahezu perfektes System zu schaffen. Der in der Rechtspraxis vorherrschende Standard wird durch den Prototyp aber allemal übertroffen.

[43] *REIMERS/GUREVYCH*, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, S. 3982–3992.