Forschungsprojekt

Automatische Anonymisierung von Gerichtsurteilen

EDV-Gerichtstag 2022

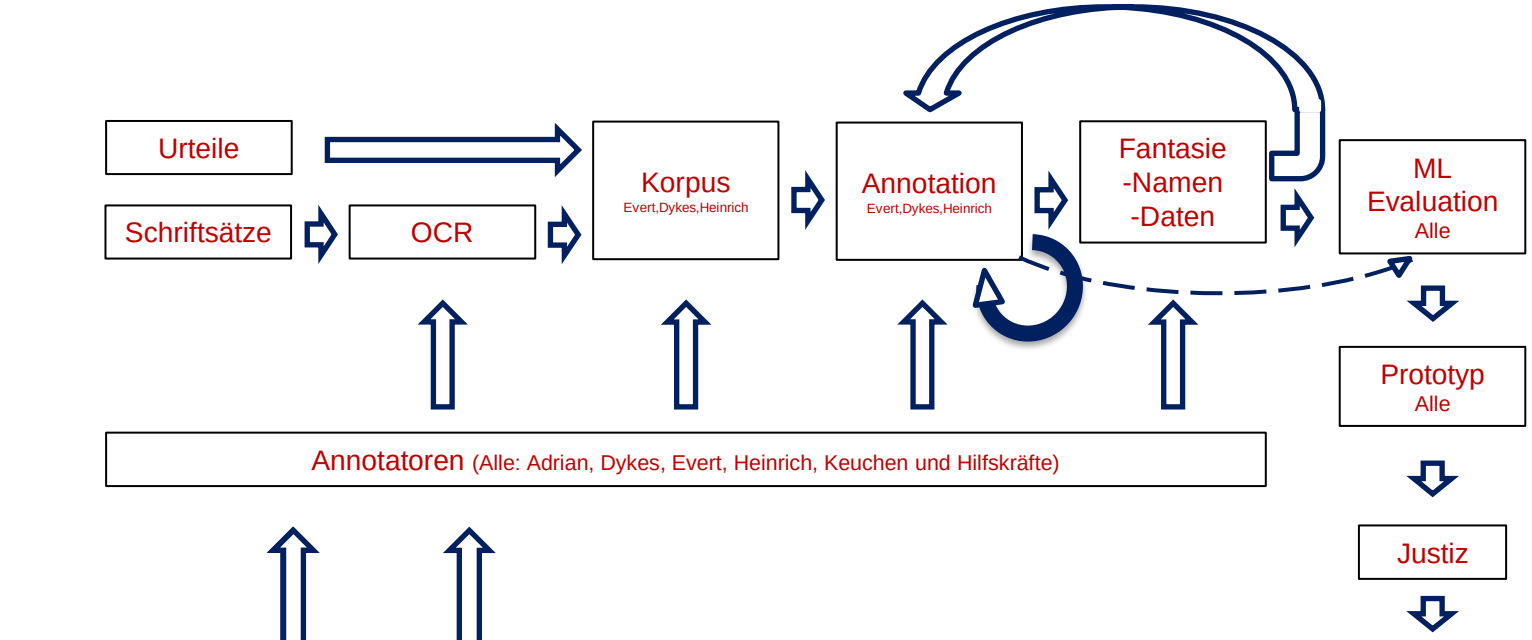
Saarbrücken, 15. September 2022

Prof. Dr. Axel Adrian
Prof. Dr. Stephanie Evert

www.str2.rw.fau.de/lehrstuhl/honorarprofessor/
www.linguistik.fau.de/team/lead

Konzeption

Konzeption und Zusammenspiel von Recht und Technik



Was ist zu anonymisieren?

Adrian, Keuchen

- Grundrecht auf informationelle Selbstbestimmung.
- DSGVO → personenbezogene Daten.
- § 30 AO → Steuergeheimnis.
- GeschäftsgeheimnisG, Unternehmenspersönlichkeitsrecht
- §§ 203 f. StGB; §§ 17 ff. UWG → Betriebs- / Geschäftsgeheimnis.
- § 30a AO a.F. → Bankgeheimnis (abgeschafft).
- § 35 SGB I → Sozialgeheimnis
- Etc. ?

Unser Team

Wissenschaftliche Leitung:

Prof. Dr. Axel Adrian
Prof. Dr. Stephanie Evert

Wissenschaftliche Mitarbeiter:

Nathan Dykes	(Linguist)
Philipp Heinrich	(Mathematiker)
Michael Keuchen	(Jurist)
Dr. Thomas Proisl (ab 04/21)	(Deep Learning)

Wissenschaftliche Hilfskräfte:

Zwischen 10-20 Personen

Scandienstleister:

BZB – Behindertenzentrum Boxdorf gemeinnützige GmbH
Boxdorfer Werkstatt und Tagesförderstätte

BayStMJ:

Ministerialdirigent Heinz-Peter Mair
Ministerialrat Dr. Martin Wachter
Regierungsrat Dr. Christoph Freimuth

Datenschutz:

Jeder Mitarbeiter hat
- Verpflichtungsvereinbarung unterzeichnet,
- Merkblatt erhalten und quittiert, sowie
- Protokoll dazu unterzeichnet

und wurde persönlich hoheitlich verpflichtet.

Zusammenarbeit mit BayStMJ und Amtsgerichten war perfekt.
Vielen Dank!

Veröffentlichung unserer bisherigen Ergebnisse

„**Manuelle und automatische Anonymisierung von Urteilen**“ von Axel Adrian, Natalie Dykes, Stefan Evert, Philipp Heinrich, Michael Keuchen, Thomas Proisl in Adrian/Evert/Kohlhase/Zwickel, Digitalisierung von Zivilprozess und Rechtsdurchsetzung, Berlin 2022, Seite 173 ff.

Mitte November erscheint in Heft 4 der LTZ unser Aufsatz „**Entwicklung und Evaluation automatischer Verfahren zur Anonymisierung von Gerichtsentscheidungen**“ von Axel Adrian, Nathan Dykes, Stephanie Evert, Philipp Heinrich und Michael Keuchen

Videovortrag zu unserem Forschungsprojekt im Auftrag des Bayerischen Staatsministeriums der Justiz samt dazugehörige Präsentation: Automatische Anonymisierung von Gerichtsurteilen

https://media.video.taxi/embed/EpcNDnZ3Xalr

Rechtliche Grundlagen zum Thema Anonymisierung von Gerichtsurteilen

Koalitionsvertrag „Mehr Fortschritt wagen“ von 2021 (S 106.)

Justiz

[…]

Gerichtsentscheidungen sollen grundsätzlich in anonymisierter Form in einer Datenbank öffentlich und maschinenlesbar verfügbar sein.

Ausgangssituation

- Einerseits rechtliche Pflicht zur Veröffentlichung von Gerichtsentscheidungen.

- Andererseits rechtliche Pflicht zum Schutz vor Preisgabe personenbezogener, personenbeziehbarer und sonstiger vertraulicher Informationen von beteiligten natürlichen und juristischen Personen (Kläger, Beklagte, Sachverständige, Zeugen, Betroffene, Anwälte, Richter, etc.) (VG Düsseldorf vom 23.11.2020, 29 K 13336/17; Ausnahme bei Marken, geschäftlichen Zeichen und Registernummer der Marken in markenrechtlichen Streitigkeiten nach OLG Frankfurt GRUR-RR 2020, 64, 67).

- Die manuelle Anonymisierung durch Justizverwaltung ist aufwendig und personalintensiv.

- Nur etwa 2,3 % aller Gerichtsentscheidungen in Deutschland werden derzeit durchschnittlich veröffentlicht. Überwiegend letztinstanzliche und obergerichtliche Entscheidungen. Tatsacheninstanzen und Eingangsinstanzen sind unterrepräsentiet (Keuchen/Deuber RDi 2022, 229 (233) haben eigene Berechnungen angestellt und kommen auf 2,3 %; „Weniger als 2 %“ werden typischerweise genannt, wenn COUPETTE/FLECKNER, Quantitative Rechtswissenschaft, JZ 2018, S. 379 (S. 381) auch nur von einer sehr kleinen Teilmenge sprechen, oder HARTUNG mitteilt, dass 99 % aller Urteile nicht in digitaler Form verfügbar sind: www.netzpiloten.de/werkzeuge-daten-gerichtsurteile (abgerufen am 26.10.2020). Siehe zu älteren Statistiken KUNTZ, Quantität gerichtlicher Entscheidungen als Qualitätskriterium juristischer Datenbanken, JurPC Web-Dok. 12/2006, Abs. 34: dort ist die Rede von 0,27 – 4,95 %.)

- Ziele: (1) Zugang Bürger zum Recht (nur Veröffentlichung) => Anonymisierung und/oder
  (2) Trainingsdaten für KI-Anwendungen => realistische Pseudonymisierung

- Mittel: möglichst automatische Verfahren zur Anonymisierung und Pseudonymisierung

Rechtsbegriff der Anonymität

Anonymisierung erfolgt indem Verknüpfbarkeit von Merkmalsausprägungen mit einem Individuum verunsichert wird, durch Vergrößerung der Menge an möglichen Merkmalsträgern dieser Merkmalsausprägungen.

Für die (indirekte) Identifizierbarkeit, sind alle objektiven Faktoren wie die Kosten der Re-Identifizierung und der dafür erforderliche Zeitaufwand und Arbeitseinsatz zu berücksichtigen, wobei dafür die zum Zeitpunkt der Anonymisierung verfügbare Technologie und die technologischen Entwicklungen maßgeblich sind (Erwg. 15, 26 DSGVO).

So kann bspw. ein Wohnort nicht nur über die Adresse dargestellt werden, sondern über deskriptive Merkmale, wie das (einzige) rote Haus in einem bestimmten kleinen Dorf. Hier lauert eine häufig unterschätzte Gefahr in Urteilen aufgrund von kostengünstigen Verknüpfungsmöglichkeiten eine De-Anonymisierung herbeizuführen

Anonymität ist dann gewährleistet, wenn Zeit-, Kosten- und Technik-Aufwand zu groß, als dass mit Re-Identifizierung zu rechnen ist (OLG Karlsruhe vom 22.12.2020, 6 VA 24/20; VGH Baden-Württemberg vom 23.7.2010, 1 S 501/10).

Verfahren zur De-Anonymisierung und daraus resultierende Gefahren einer Re-Identifizierung sind noch unzureichend erforscht. Es sind empirische Daten zum Arbeits- und Zeiteinsatz sowie zum verfügbaren Zusatzwissen notwendig.

Umsetzung/Tag-Set

Identifikatoren – Textstellen die eine Identifizierung ermöglichen

Direkte Identifikatoren:

- Namen (natürliche und juristische Personen)
- Adressangaben
- Geburtsdaten
- …

Indirekten Identifikatoren:

- Berufsangaben
- Titel
- Gesundheitsdaten
- deskriptive Angaben (örtliche Verhältnisse, Betriebsinformationen)
- einzigartige Merkmale (das einzige rote Haus im kleinen Dechsendorf)
- …

Entwickeltes Tag-Set auf Basis der juristischen Dogmatik

- **Formales**
  - Aktenzeichen
  - Gericht
- **Natürliche Person**
  - Name
  - Juristische Funktionsträger
  - Sonstiges identifizierendes Merkmal
- **Juristische Person**
  - Name
  - Sonstiges identifizierendes Merkmal
- **Adresse**
  - Ortsangabe
  - Sonstiges identifizierendes Merkmal
- **Kontaktdaten**
- **Datum**
  - Weltwissen
  - Sachverhalt
  - Prozessgeschichte
- **Fahrzeug**
  - Kennzeichen
  - Fahrgestellnummer
  - Sonstiges identifizierendes Merkmal
- **Sonstiges**

Weitere Optionen bei allen Tags:
- Risikoniveau (hoch/mittel/niedrig)
- Informationserhaltung notwendig
- unsicher/diskussionswürdig

Ergebnis zur Frage, wann etwas im Rechtssinne anonym ist

Die rechtlichen Anforderungen an Anonymität sind dann erfüllt, wenn eine Zuordnung der direkten und indirekten Identifikatoren zu einer betroffenen oder beteiligten natürlichen oder juristischen Person nur mit einem unvernünftig großen Aufwand an Zeit, Kosten und Arbeitskraft möglich ist.

Datengrundlage und Stand des Forschungsprojektes

Geschätzte und tatsächliche Datengrundlage

Geschätztes Datenmaterial nach dem Forschungsvertrag:

- ca. 300 amtsgerichtliche Gerichtsakten im Bereich des Wohnraummietrechts aus den Jahren 2015–2019
- geschätzter Umfang pro Akte von 5–10 Seiten (Urteil) und 15–30 Seiten (Klageschrift, Klageerwiderung, Replik, Duplik)
- geschätzter Durchschnittswert von 200 Wörtern pro Textseite
- Geschätzter Gesamtumfang der Daten (ca. 1,2 Millionen – 2,4 Millionen Wörter)
    - ca. 300.000–600.000 Wörter (Urteile) und
    - ca. 900.000–1.800.000 Wörter (restliche Schriftsätze)

Für erfolgreiche Projektdurchführung sollten 500.000 Wörter Urteile + 1.500.000 Wörter restliche Schriftsätze = ca. **2 Millionen Wörter Text** angestrebt werden.

Tatsächlich verfügbares Datenmaterial:

Tatsächlich haben wir nun Daten im Umfang von 5,8 Millionen Token, was ca. **5,3 Millionen Wörtern** Text entspricht
(die Anzahl Token ist erfahrungsgemäß meist ca. 10% größer als die Anzahl Wörter).

Tatsächliche Datengrundlage im Einzelnen

Wohnraummietrecht:

- 281 Urteile mit ca. 750.000 Token, davon 32 Nahduplikate (Diakonie-Urteile) mit ca. 350.000 Token (fast die Hälfte!)
- 1758 Schriftsätze mit ca. 2,9 Millionen Token, davon 378 Nahduplikate (Diakonie-Urteile) mit ca. 1,6 Millionen Token (über die Hälfte!)

Verkehrsrecht:

- 323 Urteile mit ca. 575.000 Token
- 1353 Schriftsätze mit ca. 1,5 Millionen Token

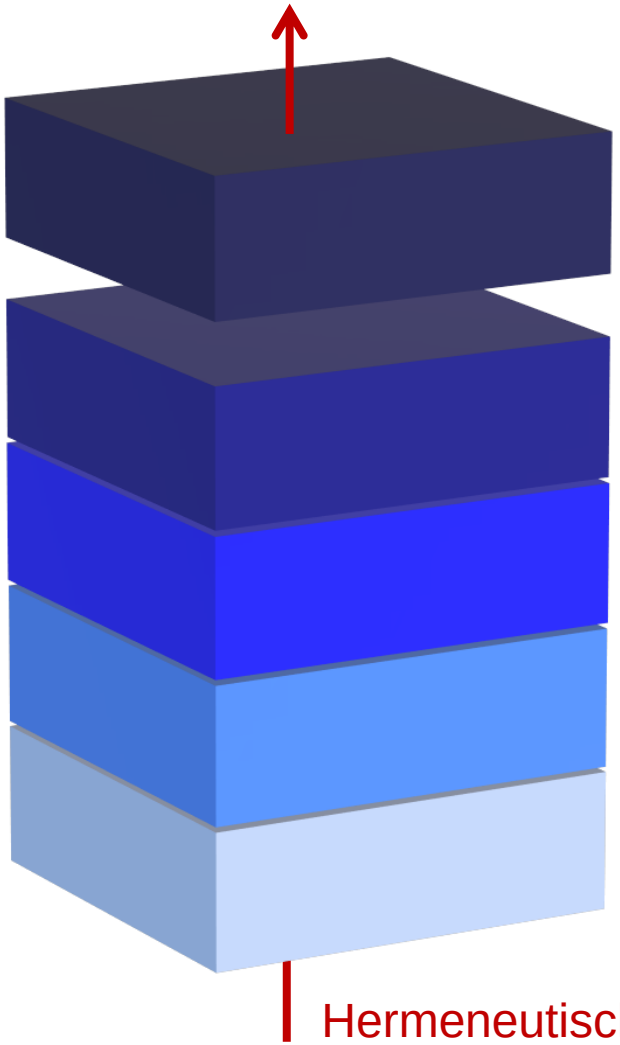
Gesamtmenge (Token aller Urteile und Schriftsätze):

- „brutto“: (1,3 Millionen in Urteilen + ca. 4,5 Millionen in Schriftsätzen): 5,8 Millionen Token
- „netto“ (d.h. ohne Nahduplikate und Rechtsbehelfsbelehrung): **3,6 Millionen Token**
  (650.000 Token in Urteilen und 3 Millionen Token in Schriftsätzen)

Exkurs- Durchschnittliche Zahl Token pro Urteil:

- Rubrum: ca. 55 Token
- Urteil: ca. 1028 Token
- Rechtsbehelfsbelehrung: ca. 176 Token

Besonderheit des Projektes: Analyse auch der Schriftsätze



- Urteil
- Duplik
- Replik
- Klageerwiderung
- Klageschrift

Hermeneutische Entwicklung hin zum Urteil

Schnelligkeit der manuellen Bearbeitung der Datengrundlage

Pro Woche schafft geschätzt ein Mitarbeiter/eine Mitarbeiterin mit einer Wochenarbeitszeit von 7 Stunden im Schnitt:

Annotation:

- ca. 30.000 – 50.000 Token

Adjudikation:

- ca. 80.000 – 100.000 Token

Pseudonymisierung:

- ca. 40.000 Token

Stand des Forschungsprojektes

Urteile sind vollständig bearbeitet und Experimente sind durchgeführt:
es wurden alle 572 Urteile vollständig, also alle 1,3 Millionen Token
- 4-fach annotiert und
- 2-fach adjudiziert und
- pseudonymisiert => Ergebnisse werden gleich von Stephanie Evert präsentiert…

Derzeit
arbeiten wir mit Hochdruck an der Bearbeitung der
- Schriftsätze: ca. 4,5 Millionen Token (brutto) bzw. 3 Millionen Token (netto), um möglichst viel Text
- 4-fach zu annotieren und
- 2-fach zu adjudizieren und zu
- Pseudonymisieren, sowie

an der Konzeptionierung und Durchführung der Experimente mit Urteilen und Schriftsätzen, was eine sehr komplexe Aufgabe ist und sehr viel Kreativität erfordert!

Bereits jetzt können wir mit dem Verfahren:
- „in house“- mit messbarer Qualität – automatisch anonymisieren und halbautomatisch pseudonymsieren
- aber noch kein Tool für Richerinnen und Richter vor Ort anbieten

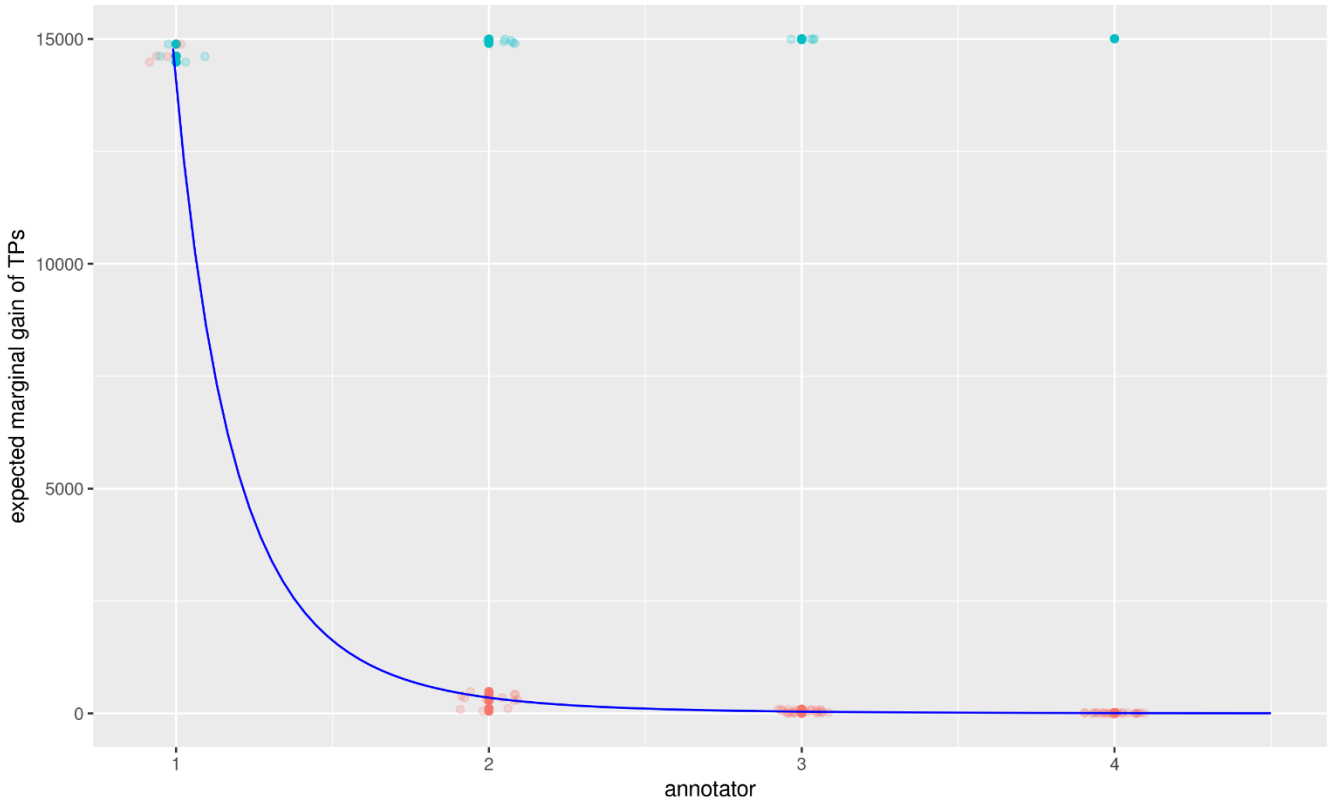
Ausblick – Verlängerung des Projektes bis Herbst 2023:
- Wir haben noch weitere 1.600 OLG-Urteile bekommen und wollen evaluieren, wie unser Prototyp auf andere Rechtsgebiete und andere Instanzen „reagiert“.
- Des Weiteren wollen wir eine Evaluation der Usability eines möglichen Frontends für Richterinnen und Richter durchführen, etc…

Ein echter Goldstandard

Abschätzung der Sensitivität

Wie viele Personen benötigt man, bis man alle relevanten Textstellen in den Urteilen gefunden hat?

Die durch jede weitere Annotatorin zusätzlich erkannten Textstellen können statistisch modelliert werden. Der marginale Gewinn fällt exponentiell ab.

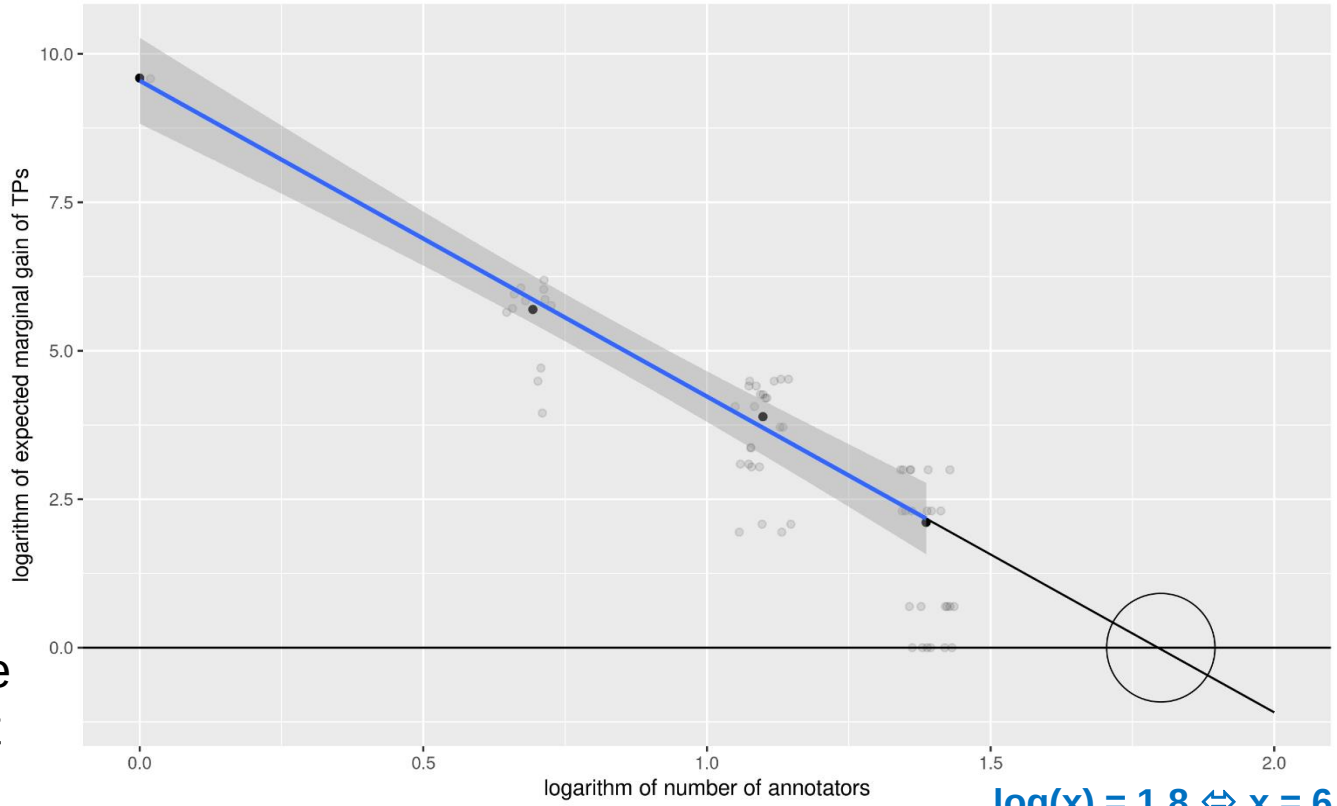


Abschätzung der Sensitivität

Man benötigt bis zu 6 Annotator:innen!

Im logarithmierten Modell kann eine Ausgleichsgerade angepasst werden.

Der Schnittpunkt mit der y-Achse zeig, dass die **6**. Annotatorin im Schnitt weniger als eine weitere Textstelle finden wird ➞ vollständige Anonymisierung ist erreicht.



log(x) = 1.8 ⇔ x = 6
log(y) = 0 ⇔ y = 1

Ein echter Goldstandard

Goldstandard bedeutet, dass jeder Urteilstext mindestens

- **4-fach annotiert** und
- **2-fach adjudiziert** wurde

und damit fast sicher alle relevanten Textstellen gefunden wurden.

Alle annotierten Textstellen wurden durch realistische **Pseudonyme** ersetzt, um den Goldstandard auch für Experimente mit Deep-Learning-Verfahren auf HPC-Rechnern außerhalb des geschützten Raumes nutzen zu können.

Goldstandard: Annotiertes und pseudonymisiertes Urteil

---------- formal:court AMTSGERICHT KEMPTEN ----------

Az. : formal-ref-number 2 C 820/18

[image8.png]

IM NAMEN DES VOLKES

In dem Rechtsstreit

1) nat-name NAUMANN Josefin , address-name Dachsberg 19 , 94032 Passau - Klägerin -

2) nat-name NAUMANN Ulf , address-name Dachsberg 19 , 94032 Passau - Kläger -

Prozessbevollmächtigte zu 1 und 2 : Rechtsanwältin nat-jur UHLIG Debora , address-name Thaler Straße 40 , 97424 Schweinfurt , Gz . : formal-ref-number 94 - 6503 / z40

gegen

1) nat-name REITER Elena Annett , address-name Schwarzer Weg 16 , 87437 Kempten (Allgäu) - Beklagte -

2) nat-name DR. REITER Günther Emil , address-name Schwarzer Weg 16 , 87437 Kempten (Allgäu) - Beklagter -

Prozessbevollmächtigte zu 1 und 2 : nat-jur Rechtsanwälte DR. GERRIT & KOLLEGEN , address-name Feldweg 19 , 97424 Schweinfurt , Gz . : formal-ref-number 903/03

wegen Räumung

erlässt das formal:court Amtsgericht Kempten durch den Richter am Amtsgericht nat-jur Schulze am date-process 06. 06. 2018 aufgrund der mündlichen Verhandlung vom date-process 08. 05. 2018 folgendes

Endurteil

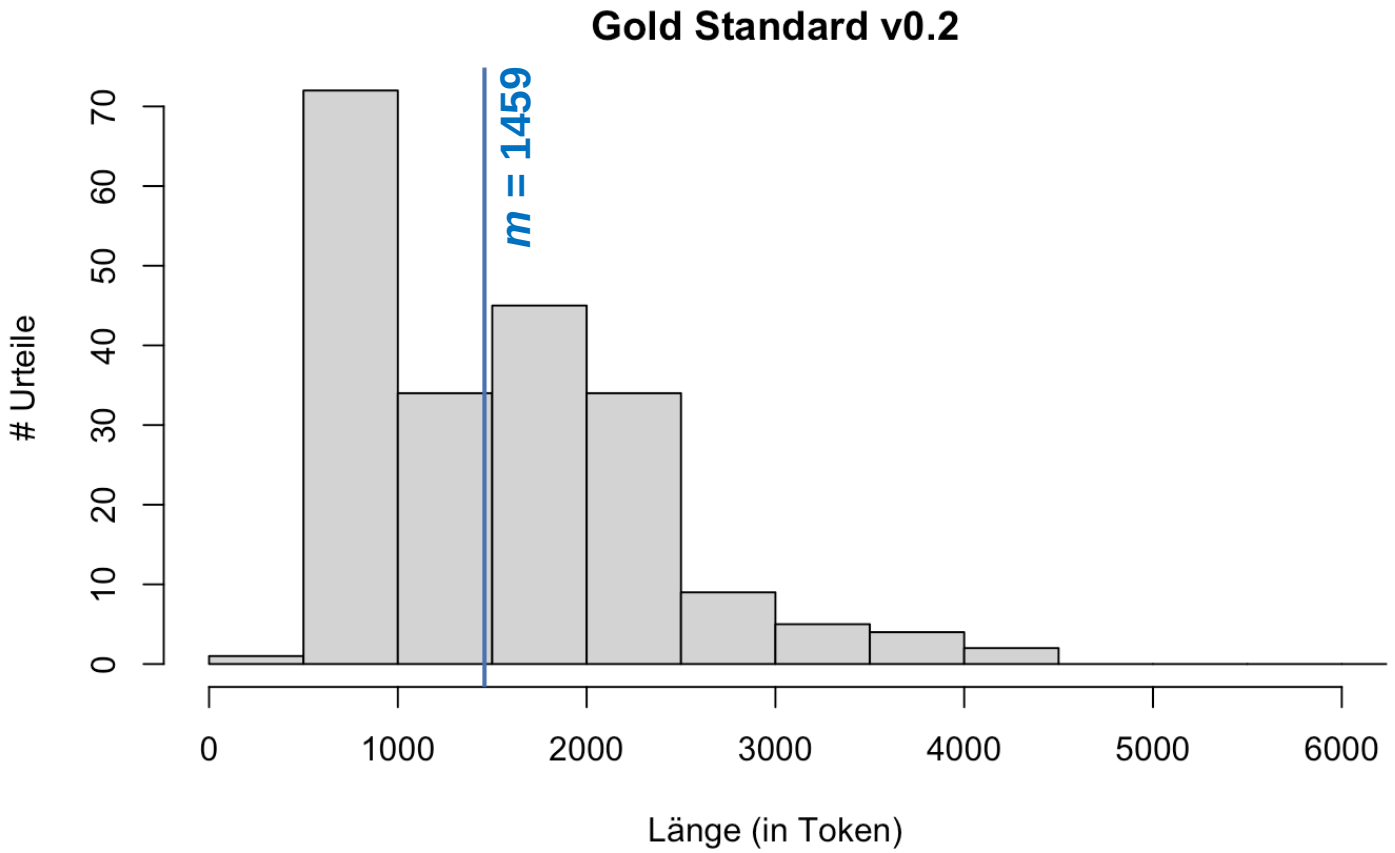
1. Die Beklagten werden verurteilt , die Wohnung im address-idx 3. Obergeschoss address-idx links des Anwesens address-name Schwarzer Weg 16 , 87437 Kempten , bestehend aus address-idx vier Zimmern , einer Abstellkammer , Küche , Flur , Bad , Toilette und das Kellerabteil Nr. address-idx 44 sowie den address-idx Stellplatz Nr. address-idx 90 , zu räumen und geräumt an die Kläger herauszugeben .

2. Den Beklagten wird eine Räumungsfrist bis date-process 31. 10. 2018 gewährt .

Quantitative Auswertung

Datenbasis: vorläufiger Goldstandard zum **Mietrecht**

- **209 Urteile** mit insgesamt
- **352.118 Token**



→ im folgenden stets auf „Normurteil“ von 1.500 Token Länge umgerechnet

Quantitative Auswertung

im Mittel 45,52 zu anonymisierende Textstellen in einem fiktiven Urteil von 1.500 Token (Mietrecht)

12,35 Textstellen mit hohem Risiko
2,16 Textstellen mit mittlerem Risiko
31,01 Textstellen mit niedrigem Risiko

	hoch	*mittel*	*niedrig*
Name: natürliche Person	4.95	0.00	0.00
Name: juristische Person	0.13	0.77	0.02
Name: Justiz	2.47	0.00	0.79
Adresse	4.80	0.64	1.70
Datum: Fakten	0.00	0.02	10.99
Datum: Prozessablauf	0.00	0.00	7.15
Aktenzeichen usw.	0.00	0.03	2.79
Gerichtsort	0.00	0.00	3.63
Merkmal: natürliche Person	0.00	0.06	0.35
Merkmal: Ortsangabe	0.00	0.64	3.58

Tabelle 1: Erwartete Anzahl zu anonymisierender Textstellen in einem Urteil durchschnittlicher Länge (1500 Token), aufgeteilt nach Informationskategorie und Risikoniveau.

Quantitative Auswertung

Im Rubrum:
Ein Großteil der Adressangaben mit hohem Risiko steht im Rubrum (4,10 von 4,80 Textstellen), sowie mehr als die Hälfte der Namen natürlicher Personen (2,60 von 4,95 Textstellen).

In Urteilsbegründung usw.:
Immer noch erhebliche Anzahl kritischer Textstellen zu erwarten (z.B. 2,35 Namen natürlicher Personen). Sonstige identifizierende Merkmale sind ausschließlich außerhalb des Rubrums zu finden!

	hoch	*mittel*	*niedrig*
Name: natürliche Person	2.60	0.00	0.00
Name: juristische Person	0.04	0.31	0.00
Name: Justiz	2.22	0.00	0.44
Adresse	4.10	0.22	0.00
Datum: Fakten	0.00	0.00	0.00
Datum: Prozessablauf	0.00	0.00	1.47
Aktenzeichen usw.	0.00	0.03	2.90
Gerichtsort	0.00	0.00	2.00
Merkmal: natürliche Person	0.00	0.00	0.00
Merkmal: Ortsangabe	0.00	0.00	0.00

Tabelle 2: Durchschnittliche Anzahl zu anonymisierender Textstellen im Rubrum eines Urteils, aufgeteilt nach Informationskategorie und Risikoniveau.

Automatische Anonymisierung

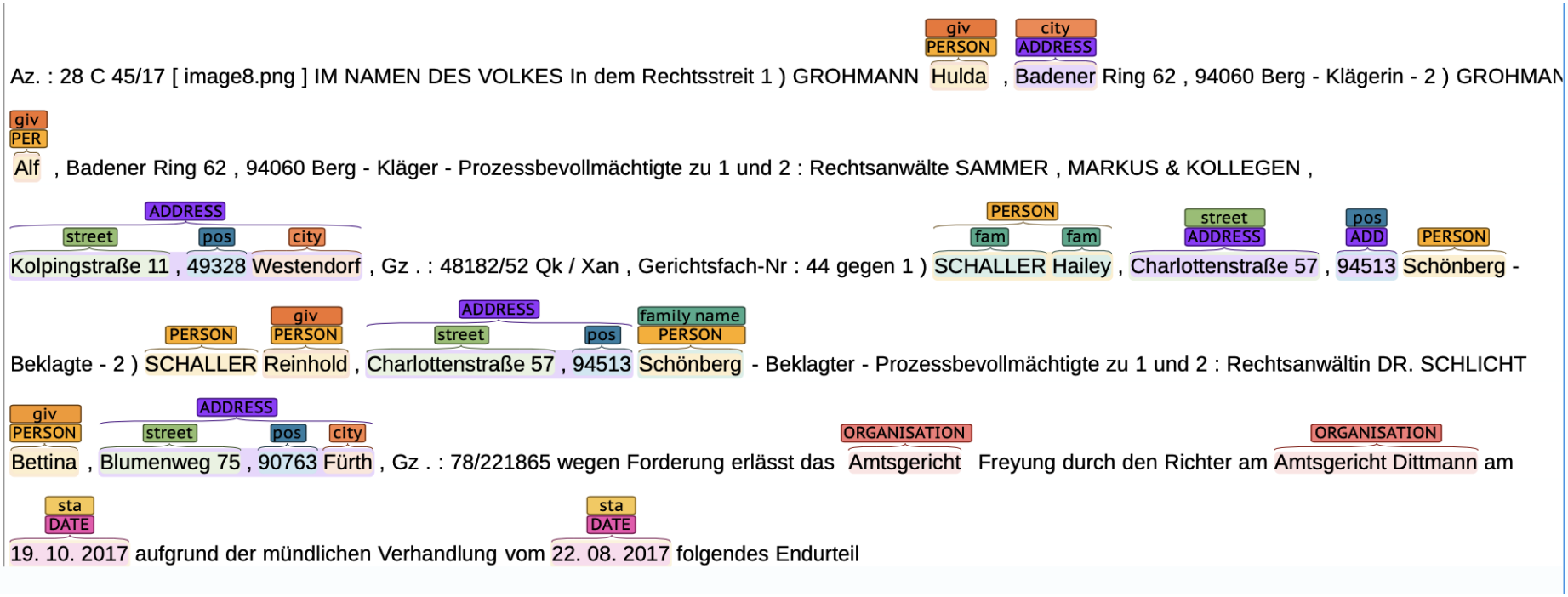
Forschungsprojekte und Softwaretools

- A-Tool von BALO.AI
  - Anonymisierungslösungen für Gerichte der Schweiz
- EU-Projekt MAPA
  - *Multilingual Anonymisation for Public Administrations*
  - Erkennung und Maskierung von Textstellen in 7 Sprachen
- OpenRedact (BMBF Prototype Fund)
  - Open-Source-Tool zur Anonymisierung von deutschsprachigen Texten v.a. durch Behörden und Gerichte
- Text Anonymization Benchmark (TAB, Pilán et al. 2022)
  - Goldstandard: EGMR-Urteile mit semi-automatischer Annotation
  - sehr gute KI-Ergebnisse bei zweifelhafter Qualität des Goldstandards
- HILANO (BMBF kmu-Innovativ)
  - *Human-in-the-Loop Lernverfahren für verteilte inkrementelle Anonymisierung* (Bucerius Law School, U Hamburg, CIB, Glanos)

Ansätze zur computergestützten Anonymisierung

- Manuelle Anonymisierung
  - unterstützt durch Word-Plugin („Office Tech“)
- Semiautomatische Verfahren
  - manuelle Überprüfung der vom Computer generierten Vorschläge
  - Beschleunigung des Verfahrens durch selbstlernende KI
- **Vollautomatische Anonymisierung**
  - nur dieser Ansatz skaliert auf ca. 1,5 Millionen Urteile / Jahr
  - Grundidee: es geht v.a. um Namen, Adressen, Datumsangaben
    → named entity recognition (NER)
  - sonstige identifizierende Merkmale bisher weitgehend ignoriert!

Anonymisierung als NER-Aufgabe



Softwareprototyp des EU-Projekts Multilingual Anonymisation for Public Administrations (MAPA, 2020–2021)

https://mapa-demo.pangeamt.com/mapa/1.0/anonymization/model_showcase

Evaluation automatischer Verfahren

	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

Tabelle 3: Evaluation der korrekten Erkennung von Textstellen (Testset: pseudonymisierte Urteile zum Mietrecht)

	Alle Textstellen			Nach Risiko		
System	Precision	Recall	F_1	*hoch*	*mittel*	*niedrig*
Standard-NER (Flair)	0.64	0.27	0.38	0.60	0.61	0.09
OpenNLP	**0.95**	0.90	0.93	0.94	0.68	0.92
Riedl & Padó	0.91	0.94	0.93	0.96	0.76	0.96
Fine-tuned GottBERT	0.90	**0.98**	**0.94**	**0.99**	0.93	0.98

Tabelle 4: Evaluation auf Tokenebene (Testset: pseudonymisierte Urteile zum Mietrecht)

- **Recall** = Wie viele der relevanten Textstellen werden vom Computer gefunden?
- **Precision** = Wie viele der vom Computer vorgeschlagenen Textstellen sind relevant?
- **F_1** als harmonisches Mittel von Precision und Recall
- Tabelle 3: Strenge Evaluation auf Ebene der Textstellen (müssen exakt gefunden werden)
- Tabelle 4: „Weiche“ Evaluation auf Tokenebene (Prozentsatz „richtiger“ Wort-Token)

Softwareprototyp „LeAK“

Urteilsart

ENDURTEIL:

Unmittelbare Enscheidung in der Streitsache

I. Die Beklagten werden verurteilt, das Anwesen Döbrabergstraße 24, 95191 Leupoldsgrün, zum 30.09.2003 zu räumen und an die Kläger herauszugeben. Im übrigen wird die Klage abgewiesen.

Kostenentscheidung

II. Die Kosten des Rechtsstreits werden gegeneinander aufgehoben.

Vorläufige Vollstreckbarkeit

III. Das Urteil ist vorläufig vollstreckbar. Die Beklagten können eine Vollstreckung der Kläger durch Sicherheitsleistung in Höhe von 10.000,00 € abwenden, wenn nicht die Kläger vor der Vollstreckung Sicherheit in gleicher Höhe leisten.

vortrainiertes NLM

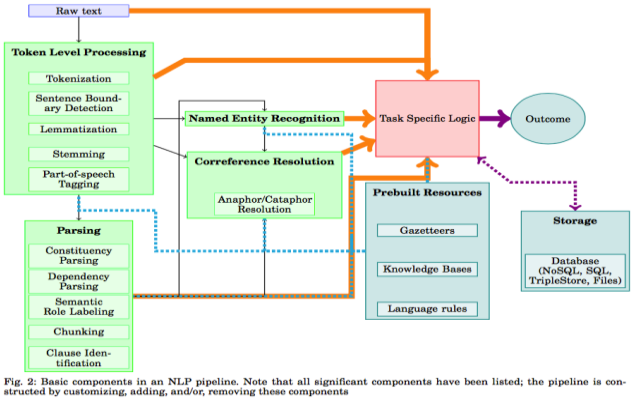
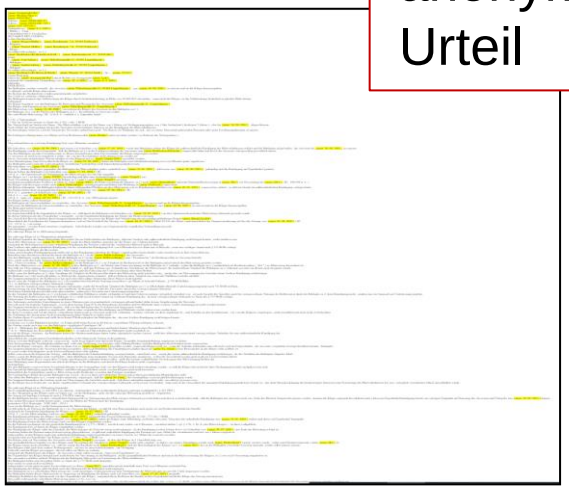


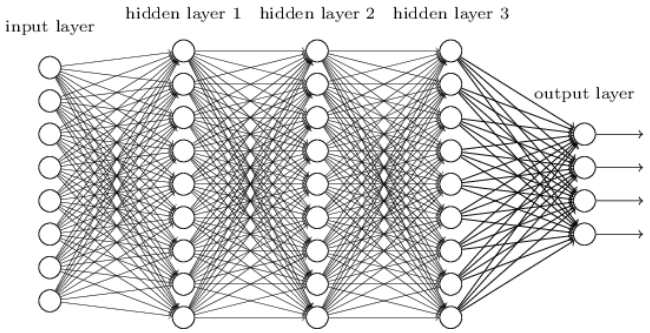
Fig. 2: Basic components in an NLP pipeline. Note that all significant components have been listed; the pipeline is constructed by customizing, adding, and/or, removing these components

https://gokulchittaranjan.files.wordpress.com/2015/09/pipeline1.png

anonymisiertes Urteil



Deep Learning



http://neuralnetworksanddeeplearning.com/images/tikz36.png

Prof. Dr. Axel Adrian
Prof. Dr. Stephanie Evert

www.str2.rw.fau.de/lehrstuhl/honorarprofessor/
www.linguistik.fau.de/team/lead

Softwareprototyp auf Basis des vortrainierten Deep-Learning-Sprachmodells GottBERT, das durch zusätzliche Layer für die Anonymisierungsaufgabe nutzbar gemacht und auf pseudonymisierten Mietrechtsurteilen trainiert wurde.

	Alle Textstellen			Nach Risiko		
System	Precision	Recall	F_1	*hoch*	*mittel*	*niedrig*
Mietrecht (Textstellen)	0.79	0.90	0.84	0.96	0.76	0.89
Mietrecht (Token)	0.90	0.97	0.93	0.99	0.83	0.98
Verkehrsrecht (Textstellen)	0.85	0.90	0.87	0.98	0.81	0.87
Verkehrsrecht (Token)	0.93	0.97	0.95	0.99	0.99	0.95

Tabelle 5: Evaluation des Prototypen auf nicht-pseudonymisierten Urteilen

Tabelle 5 zeigt eine Evaluation des Prototypen auf nicht pseudonymisierten Originalurteilen, d.h. unter völlig realistischen Bedingungen. Obwohl der Prototyp ausschließlich auf der Domäne Mietrecht trainiert wurde, erzielt er bei Urteilen aus dem Verkehrsrecht ebenfalls einen Recall von 90%!

Vom Prototyp zum Anonymisierungstool

Aktueller Prototyp erkennt relevante Textstellen

§ 2 Nr. 2 Vertragsdauer
c) Nur für Zeitmietverträge im Sinne des § 564 c Abs. 2 BGB Der Vertrag läuft auf bestimmte Dauer : Das Mietverhältnis wird auf die Dauer von 3 Jahren mit Verlängerungsoption von 1 Jahr fortlaufend (höchstens 5 Jahren) , also bis {anon: 30. 09. 2003 } , abgeschlossen .
Nach Ablauf der Vertragsdauer besitzt der Vermieter ein berechtigtes Interesse an der Beendigung des Mietverhältnisses .
Als berechtigtes Interesse wird
die Absicht des Vermieters geltend gemacht : Die Räume als Wohnung für sich , die zu seinem Hausstand gehörenden Personen oder seine Familienangehörigen zu nutzen .
Die Einliegerwohnung kann vom Mieter an Frau Rechtsanwältin {anon: Henke } untervermietet werden (im Rahmen der Vertragsdauer) .
Seite {anon: 3 } {anon: 13 C1874 / 01 }
d) ...
Abweichend hiervon wird eine Kündigungsfrist von 6 Monaten vereinbart .
Mit Schreiben vom {anon: 28. 08. 2001 } und erneut mit Schreiben vom {anon: 17. 01. 2001 } wurde
den Beklagten seitens der Kläger die außerordentliche Kündigung des Mietverhältnisses erklärt und die Beklagten aufgefordert , das Anwesen bis {anon: 30. 09. 2001 } zu räumen .
Die Kündigung wurde damit begründet , daß die Beklagte zu
2) in der Einliegerwohnung des Anwesens eine {anon: Rechtsanwaltskanzlei } eingerichtet habe und damit das Anwesen vertragswidrig gewerblich nutzte .
Von den Beklagten sei eine Zwischenmauer im Hauswirtschaftsraum ohne Einverständnis der Kläger eingezogen worden .
Im Anwesen sei ein Kamin eingebaut worden , der von dem Kaminkehrer nicht abgenommen worden sei .
Ein im Anwesen aufgetretener Wasserschadeh sei den Klägern erst im {anon: August 2001 } gemeldet worden .
Trotz
Beauftragung einer Firma durch die Kläger am {anon: 23. 08. 2001 } hätten die Beklagten eine Schadensbeseitigung erst zwei Monate später zugelassen .
Die Beklagten
seien auch der mietvertraglich vereinbarten Gartenpflege nicht hinreichend nachgekommen .
Mit Schreiben vom {anon: 29. 07. 2002 } (BI .
110 d. A.) haben die Kläger gegenüber den Beklagten das Mietverhältnis zudem ordentlich zum {anon: 31. 01. 2003 } , hilfsweise zum
{anon: 30. 09. 2003 } , gekündigt und die Kündigung auf Eigenbedarf gestützt .
Hierauf haben die Beklagten mit Schreiben vom {anon: 17. 01. 2003 } (BI .
138 d. A.) ihr Optionsrecht auf Verlängerung des Mietvertrages um ein Jahr ausgeübt .
Der Kläger zu 2) ist bei der {anon: NATO } beschäftigt und übte eine Auslandstätigkeit in {anon: Neapel } aus .
Nach Vermietung an die Beklagten sind die Kläger zu 1) und 2) nach {anon: Neapel } gezogen .
Aufgrund gesundheitlicher Probleme erfolgte im Jahre {anon: 2002 } die Rückversetzung des Klägers zu 2) nach {anon: Deutschland } und eine Dienststellenzuweisung in {anon: Hof } mit Verwendung bis {anon: 2014 } (BI . 108/109 d. A.) .
Die Kläger sind deswegen wieder nach {anon: Deutschland } zurückgezogen und haben eine Wohnung in {anon: Feilitzsch } angemietet .
Die Kläger behaupten , die Beklagten hätten die ihnen obliegenden Pflichten aus dem Mietvertrag wie im Kündigungsschreiben vom {anon: 28. 01. 2001 } vorgeworfen verletzt , so daß ein Grund zur außerordentlichen Kündigung vorlegen habe .
Die Kläger haben ihren ursprünglichen Klageantrag mit Schriftsatz vom {anon: 05. 06. 2002 } (BI .
88 d. A.) , geändert mit Schriftsatz vom {anon: 28. 08. 2002 } (BI .
106 d. A.) , vom {anon: 04. 10. 2002 } (BI .
120 d. A.) und vom {anon: 28. 01. 2003 } (BI . 134/135 d. A.) um Hilfsanträge ergänzt .
Die Kläger haben zuletzt beantragt , die Beklagten als Gesamtschuldner zu verurteilen , das Anwesen
{anon: Döbrabergstraße 24 , Leupoldsgrün } zu räumen und an die Kläger herauszugeben , hilfsweise die Beklagten als Gesamtschuldner zu verurteilen , das Anwesen {anon: Döbrabergstraße 24 , Leupoldsgrün } zum {anon: 30. 09. 2003 } zu räumen und an die Kläger herauszugeben .
Die Beklagten haben beantragt , die Klage abzuweisen .
Seite {anon: 4 13 C1874 / 01 }
Sie tragen hinsichtlich des Eigenbedarfs der Kläger vor , daß durch die Beklagten mit Schreiben vom {anon: 30. 05. 2002 } von dem Optionsrecht nach dem Mietvertrag Gebrauch gemacht wurde .
Da dieses Optionsrecht dem Eigenbedarf vorausgehe , sei der

Der Weg vom Prototyp zum vollautomatischen Anonymisierungstool

- Optimierung der Erkennungsqualität
  - bis zu 96% Recall bereits sehr gut, aber bei 1,5 Mio. Urteilen …
  - Ziel: Recall > 99% v.a. bei hohem Deanonymisierungsrisiko

- Nutzersteuerung
  - Anonymisierungsniveau: hohes / mittleres / niedriges Risiko
  - Auswahl verschiedener Informationskategorien für Anonymisierung (z.B. juristische Personen, Aktenzeiche, ortsidentifizierende Merkmale)

- Maskierung der Textstellen
  - einfach & sicher: Schwärzung (bzw. Löschung)
  - Pseudonymisierung: Initialen, abstrakte Bezeichner (A*, B*, C*, …), …
  - realistische Pseudonymisierung → Trainingsdaten für Legal Tech

- Qualitätssicherung
  - vollautomatische Anonymisierung mit empirisch belegter Zuverlässigkeit
  - Erweiterung auf andere Rechtsgebiete und Instanzen

Zwischenstand

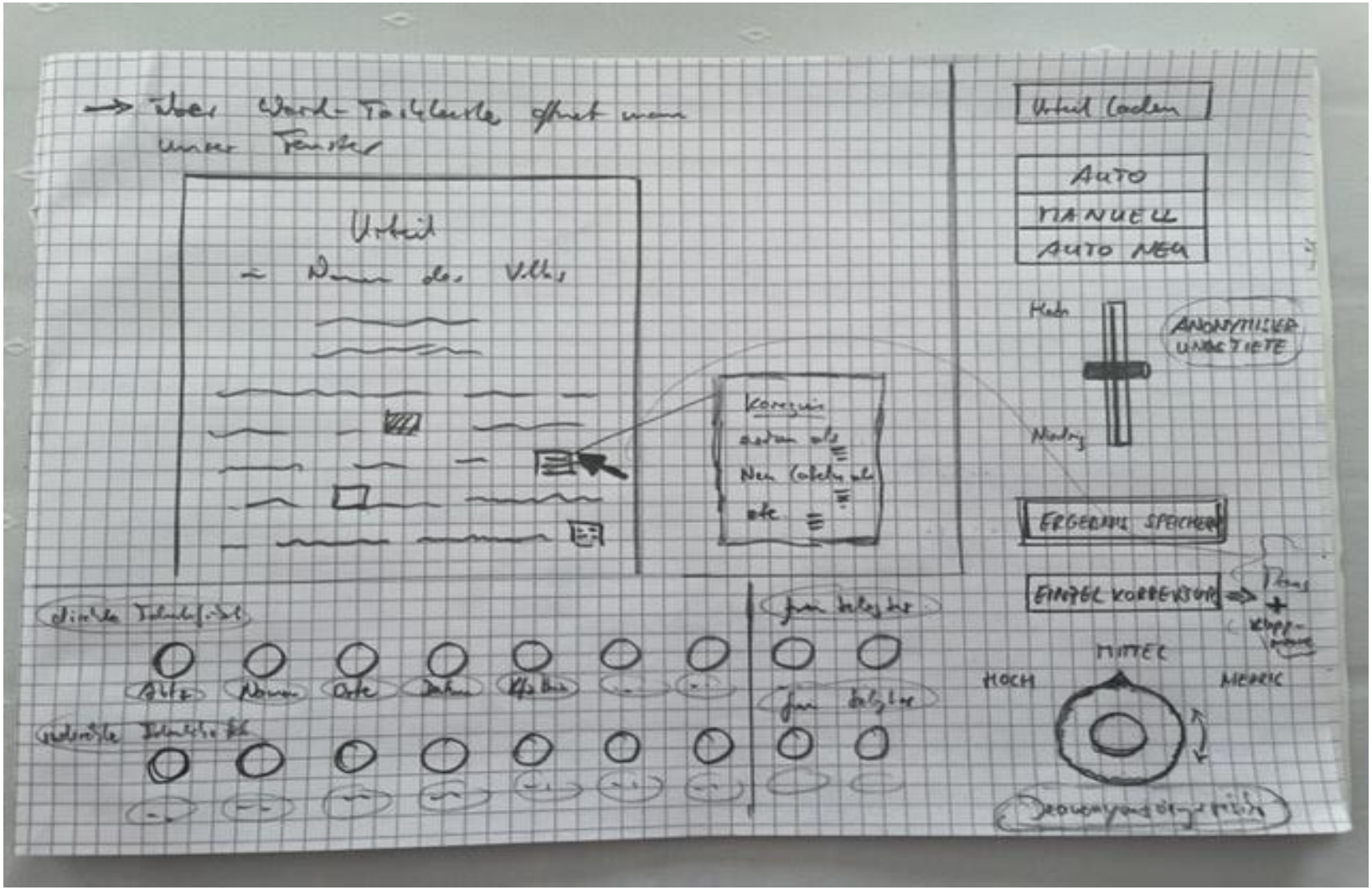
- Vollständiger Goldstandard
  - 570 Urteile (247 Miet + 323 Verkehr) mit insg. knapp 1 Mio. Wort-Token, davon ca. 244.000 Token Rechtsbehelfsbelehrung
  - 50% Trainingsdaten + 25% Development + 25% Testdaten

- KI-Optimierung
  - verschiedene vortrainierte Sprachmodelle
  - Verbesserung von Netzwerktopologie, Modellierung, Trainingsschema
  - damit **> 96% Recall** und Precision
  - darunter **98,7% aller Hochrisikostellen**

- Feinsteuerung
  - gezielte Anonymisierung von Hochrisikostellen schwierig (F_1 ~ 95%)
  - Erkennung von Informationskategorien fehlerbehaftet (insb. sonstige identifizierende Merkmale)
  - aber: > 99% Recall für Aktenzeichen + natürliche/juristische Personen

- Deanonymisierungsexperimente (2022/23)

Erste Skizze zu einem Frontend für Richterinnen und Richter



Vielen Dank

Fragen?