LEGAL REVOLUTIONary
Recht in der digitalen Wirtschaft

Smart Sentencing

Grundriss einer teilautomatisierten Strafzumessungsdatenbank

Prof. Dr. Dr. Frauke Rostalski | Inhaberin des Lehrstuhls für Strafrecht u.a., Universität Köln | Mitglied des deutschen Ethikrates

Timothée Schmude | Lehrbeauftrager Universität Köln

Malte Völkening | Wiss. Mitarbeiter | Universität Köln

Jin Ye | Wiss. Hilfskraft | Universität Köln

12. August 2021

LRZ 2021, Seiten 166 bis 178 (insgesamt 13 Seiten)

Der Beitrag befasst sich mit der Rationalisierung der Strafzumessungspraxis in Deutschland. Vergangene Untersuchungen haben Hinweise auf normativ nicht gerechtfertigte Unterschiede insbesondere zwischen den Strafzumessungsentscheidungen verschiedener Gerichte festgestellt. Smart Sentencing ist ein Datenbanksystem, das zur Lösung dieses Problems beitragen soll. Richter*innen bietet es die Möglichkeit, sich über die Entscheidungen ihrer Kolleg*innen in sachlich vergleichbaren Fällen zu informieren. Anwält*innen und Staatsanwält*innen profitieren von der besseren Transparenz und Vorhersehbarkeit der richterlichen Entscheidungen. Um die Arbeitsbelastung der Gerichte nicht noch weiter zu erhöhen, wird eine technologiegestützte Analyse der Urteile erprobt. Erste Ergebnisse sind ermutigend. Zur weiteren Verbesserung sind aber zusätzliche Trainingsdaten erforderlich.

I. Einleitung

Das Strafrecht als schärfstes Schwert des Staates greift erheblich in die Rechte des Einzelnen ein.[1] Trotz der Bedeutsamkeit dieser strafrechtlichen Eingriffe für die Freiheit des Einzelnen scheint es, als würden im Bundesgebiet unterschiedliche Maßstäbe für den Einsatz dieses Schwertes gelten.

[1] *Rostalski*, Der Tatbegriff im Strafrecht, 2019, S. 16, 78; *Lagodny*, Zwei Strafrechtswelten, 2021, S. 170.

Mehr als 2.400 Strafverfahren werden täglich in deutschen Gerichtssälen abgeurteilt und rund 80% dieser Verfahren münden in der Verurteilung der Angeklagten.[2] Grundlage hierfür ist mit dem Strafgesetzbuch ein Bundesgesetz. Für alle Verurteilungen im Bundesgebiet gilt also dasselbe Strafrecht mit denselben Zumessungsregeln – für ein und dieselbe Tat wäre demnach bei allen Richter*innen grundsätzlich die gleiche Strafe zu erwarten. Dennoch variiert die Strafzumessung für vergleichbare Taten erheblich in Abhängigkeit vom entscheidenden Gericht.[3]

In Anbetracht der vielfältigen Straftatbestände und der Individualität eines jeden Einzelfalles sind divergierende Strafzumessungsentscheidungen zwar – auch für das gleiche Delikt – zu erwarten, diese müssen jedoch auf normativ relevanten Umständen beruhen. Für die Ortsabhängigkeit einer Strafzumessungsentscheidung lässt sich aber keine rechtliche Grundlage finden.[4]

Um den gravierenden Grundrechtseingriff durch Schuldspruch und Strafe zu rechtfertigen, genügt es nicht, bei der Ermittlung des Sachverhaltes sowie beim Subsumieren des Sachverhaltes unter den gesetzlich normierten Straftatbestand einen strengen Maßstab anzulegen. Vielmehr muss die Strafzumessung, welche neben dem Schuldspruch über das Ausmaß des Grundrechtseingriffs „Strafe" bestimmt, ebenfalls dem Gleichheitssatz genügen.[5] Wenn ein- und derselbe Fall von unterschiedlichen Richter*innen mit signifikant voneinander abweichenden Strafmaßen belegt würde, ist dies vor dem Hintergrund des Art. 3 Abs. 1 GG problematisch.[6]

Die Aburteilung einer Tat dient der Wiederherstellung des Rechtsfriedens innerhalb der Gesellschaft. Hierzu ist erforderlich, dass sich strafrichterliche Entscheidungen gegenüber Täter*in und Gesellschaft als angemessene Reaktion *der Rechtsgemeinschaft* auf die Tat darstellen.[7] Das kann nur gelingen, wenn die Richter*innen als Vertreter*innen der Gesellschaft auftreten und nicht als Privatpersonen.[8] Sie müssen also die rechtlichen, nicht ihre eigenen Maßstäbe anlegen. Das erfordert, dass die Strafzumessungsgründe objektiv nachvollziehbar und transparent sind. Außerdem muss nach relativer

[2] *Statistisches Bundesamt (Destatis)*, Strafverfolgung – Fachserie 10, Reihe 3, 2019, S. 16.
[3] *Kaspar*, Sentencing Guidelines versus freies tatrichterliches Ermessen, 2018, C 18 ff.
[4] *Rostalski/Völkening*, KriPoZ 2019, 265 (265 f.).
[5] Kaspar, Zur aktuellen Diskussion über die Strafzumessung im deutschen Recht, The Art of Crime 11/2019, ein Online-Journal in griechischer Sprache, eine deutsche Fassung ist abrufbar unter https://theartofcrime.gr/wp-content/uploads/2020/01/J. -Kaspar_Zur-aktuellen-Diskussion-%C3%BCber-die-Strafzumessung-im-deutschen-Recht.pdf, S.3.
[6] Albrecht, Strafzumessung bei schwerer Kriminalität, 1994, S. 119; Streng, StV 2018, 593 (594); Kaspar, Zur aktuellen Diskussion über die Strafzumessung im deutschen Recht, The Art of Crime 11/2019, deutsche Fassung, 3.
[7] *Rostalski/Völkening*, KriPoZ 2019, 265 (266); vgl. auch *Streng*, StV 2018, 593 (596).
[8] *Rostalski*, GA 2021, 198 (209).

Strafzumessungsgerechtigkeit gestrebt werden, sodass das einzelne Strafmaß weder auf den oder die Verurteilte*n noch auf die Gesellschaft im Vergleich zu anderen Strafmaßen willkürlich erscheint.[9]

Ein Ansatz hierfür besteht darin, möglichst viele Urteile zu sichten, die darin aufgeführten Strafzumessungserwägungen miteinander zu vergleichen und zu kategorisieren.[10] Bedenkt man allerdings, welche Menge an Urteilen es bereits gibt und dass jährlich rund 690.000 weitere Strafverfahren mit einer Verurteilung enden,[11] so wird schnell klar, dass eine repräsentative händische Auswertung faktisch unmöglich ist.

Hier können technologische Analysewerkzeuge Abhilfe schaffen: Machine Learning ermöglicht es, Textdokumente nach ausreichendem Training auf einem speziell aufbereiteten Datensatz automatisch zu analysieren und dabei die relevanten Faktoren aus dem Gesamttext herauszufiltern. Welche Anwendungsszenarien es für eine derart trainierte Maschine gibt, wie sie technisch umgesetzt werden könnte und welche Herausforderungen dieses Vorhaben birgt, soll im Folgenden anhand des Projekts „Smart Sentencing" erläutert werden, das am Lehrstuhl von *Frauke Rostalski* in Kooperation mit dem Fraunhofer Institut für Intelligente Analyse- und Informationssysteme durchgeführt wird.[12]

II. Anwendungsszenarien

Die Anwender*innen, auf die das Projekt Smart Sentencing zielt, sind zunächst die unmittelbar am Verfahren Beteiligten, d.h. die Richter*innen, Staatsanwält*innen und Anwält*innen. Daneben sind die gewonnenen Erkenntnisse aber auch für die Wissenschaft und die Gesellschaft im Ganzen von Interesse.

1. Aus Sicht der Richter*innen

a) Informationssuche

Auf den ersten Blick ähnelt das System den bekannten juristischen Fachdatenbanken, etwa *beck-online* und *juris*. Über eine Eingabemaske können bestimmte Suchparameter

[9] *Rostalski/Völkening*, KriPoZ 2019, 265 (267).
[10] Zur komparativen Strafzumessung mithilfe interner Strafmaßtabellen etwa Kudlich/Koch, NJW 2018, 2762 (2763).
[11] *Statistisches Bundesamt (Destatis)*, Strafverfolgung – Fachserie 10, Reihe 3, 2019, S. 16 Abb. 1.1.
[12] Weitere Informationen unter https://rostalski.jura.uni-koeln.de/forschungsprojekte/smartsentencingdb.

angegeben werden. Ausgegeben werden alle gerichtlichen Entscheidungen, die den Parametern entsprechen. Bei Smart Sentencing liegt der Fokus jedoch auf den Urteilen der Instanzgerichte, weil die Strafzumessungspraxis im Ganzen abgebildet werden soll. Um die große Datenmenge zu bewältigen, weicht die Art der Suchkriterien von den herkömmlichen Datenbanksystemen ab. Statt den Volltext nach bestimmten Begriffen zu durchsuchen, ist Smart Sentencing auf eine inhaltliche Analyse der Entscheidungen ausgelegt. Gesucht wird nach standardisierten Strafzumessungsfaktoren. Dazu gehören beispielsweise die Höhe des entstandenen Schadens, ein Geständnis des oder der Angeklagten und ähnliches. Die Ausgabe enthält Entscheidungen, deren Strafzumessungserwägungen den angegebenen Faktoren hinreichend ähnlich sind.

Es gibt zwei Möglichkeiten, mit diesen Informationen zu arbeiten. Zum einen können (wie bei *juris*) die Volltexte der Entscheidungen angezeigt werden, um die Genauigkeit der maschinellen Auswertung und etwaige nicht standardisierte Faktoren (wie die kriminelle Energie des oder der Angeklagten) zu beurteilen. Zum anderen analysiert das System die Strafhöhe, die aufgrund der angegebenen Zumessungsfaktoren tendenziell verhängt wird. Diese Tendenz kann statistisch auf die einzelnen Faktoren heruntergebrochen werden. So ließe sich beispielsweise errechnen, welchen Einfluss ein Geständnis oder eine bestimmte Schadenshöhe unter sonst gleichen Umständen üblicherweise auf die Strafhöhe hat.[13]

Um einen Ankereffekt, d.h. die Verzerrung des Urteils in Richtung eines vorgegebenen Wertes,[14] zu vermeiden, werden keine Durchschnittswerte angegeben. Stattdessen ist eine Visualisierung über Konfidenzintervalle oder eine Heat Map vorzugswürdig, um die Varianz in der Datengrundlage abzubilden. Bei ersteren wird die Spannweite angegeben, in der sich der Großteil der bisherigen Strafzumessungsentscheidungen bewegt, bei letzterer ein Farbverlauf, der die Verteilung der Entscheidungen wiedergibt.

Richter*innen können diese auf ihren konkreten Fall zugeschnittenen Tendenzen als Einstiegspunkt für die Strafzumessungsentscheidung verwenden.[15] Diese Aufgabe wird derzeit durch die persönliche Erfahrung (das „geheime[ ] Metermaß“[16]), den Flurfunk und

[13] Dazu bereits Rostalski/Völkening, KriPoZ 2019, 265 (270, 272). Händisch können solche Analysen auch jetzt schon durchgeführt werden, s. etwa Albrecht, ZStW 102 (1990), 596 (612 ff.). Wegen des großen Codieraufwands ist die Datenbasis aber stark eingeschränkt. Der genannten Untersuchung etwa lagen Urteile aus fünf deutschen Landgerichtsbezirken zugrunde (Albrecht, ZStW 102 (1990), 596 (608)).
[14] Dazu Nickolaus, Ankereffekte im Strafprozess, 2018, S. 23 ff., speziell zum Einfluss auf das Strafmaß s. S. 77 ff. m.w.N. Vgl. zu diesem Problem auch Kaspar/Höffler/Harrendorf, NK 32 (2020), 35 (46 Fn. 54).
[15] Vgl. Rostalski/Völkening, KriPoZ 2019, 265 (268 ff.). Zur Bedeutung des Einstiegspunktes s. Meier, Strafrechtliche Sanktionen, 5. Aufl. (2019), 168, 241, 245.
[16] Dreher, MDR 1961, 343 (344).

teilweise auch durch informelle interne Tabellen erfüllt.[17] Im Vergleich dazu bietet Smart Sentencing eine Rationalisierung, weil ein objektiv nachprüfbarer und transparenter[18], überregionaler und von individuellen Einstellungen und Präferenzen[19] abstrahierender Anknüpfungspunkt verwendet werden kann. Das ist ein rechtsstaatlicher Fortschritt, aber auch eine Erleichterung für junge Richter*innen, denen die nötige (eigene) Erfahrung fehlt.[20]

Dabei darf Smart Sentencing nicht als erster Schritt hin zum „Robo-Judge" missverstanden werden.[21] Das System nimmt nur eine statistische Auswertung vor. Ob sich Richter*innen an dieser Auswertung orientieren, müssen sie selbst entscheiden.[22] Die richterliche Unabhängigkeit erlaubt ihnen, die Bedeutung der in ihrem Fall relevanten Strafzumessungsfaktoren (begründet) anders einzuschätzen als ihre Kolleg*innen.[23] Außerdem können nur solche Faktoren automatisiert ausgewertet werden, die hinreichend standardisierbar sind. Sie müssen also häufig vorkommen und dürfen nicht zu wertungsoffen sein.[24] Das bedeutet, dass die von Smart Sentencing errechneten Rahmen nicht einfach übernommen werden dürfen. Stets ist zu prüfen, ob der konkrete Fall relevante Besonderheiten und Abweichungen gegenüber der früheren Rechtsprechung enthält. Dann muss das Strafmaß entsprechend angepasst werden.[25]

---

[17] Vgl. Meier, Sanktionen (Fn. 15), S. 244 f.; zu internen Tabellen auch Kudlich/Koch, NJW 2018, 2762 (2763). Ein Beispiel findet sich bei Schäfer/Sander/van Gemmeren, Praxis der Strafzumessung, 6. Aufl. (2017), Rn. 1719 ff.

[18] Kritisch in Bezug auf interne Tabellen der StA Kinzig, in: Schönke/Schröder, StGB, 30. Aufl. (2019), § 46 Rn. 72.

[19] Zur Bedeutung individueller Präferenzen der Richter*innen für die Strafzumessung s. Streng, in: Kindhäuser/Neumann/Paeffgen, NK-StGB, 5. Aufl. (2017), § 46 Rn. 203 m.w.N. Laut Meier, Sanktionen (Fn. 15), S. 261 spielt die Person des oder der Richter*in dagegen nur eine „geringe Rolle".

[20] Vgl. zu letzteren Meier, Sanktionen (Fn. 15), S. 244.

[21] Einem solchen Irrtum scheint Greco, RW 2020, 29 (31 Fn. 10) zu unterliegen, dagegen bereits Rostalski/Völkening, ZfDR 2021, 27 (28 Fn. 4). Allgemein zum „Robo-Judge" etwa Gless/Wohlers, in: FS-Kindhäuser, 2019, S. 147 (147 ff.).

[22] Vgl. Rostalski/Völkening, ZfDR 2021, 27 (28 Fn. 4).

[23] Vgl. Streng, StV 2018, 593 (599) und die Kritik von Blankenburg, in: FS-Rottleuthner, 2011, S. 262 (265) an von Richter*innen aufgestellten „Leitlinien", die zu einer „Stromlinien-Organisation von Gerichten" führe. Blankenburg sieht die richterliche Unabhängigkeit aber bereits durch die Notwendigkeit, Abweichungen zu begründen, eingeschränkt. Vgl. außerdem Kaspar/Höffler/Harrendorf, NK 32 (2020), 35 (46); Streng, in: Kindhäuser/Neumann/Paeffgen, § 46 Rn. 202.

[24] Zu diesem Problem bereits Rostalski/Völkening, KriPoZ 2019, 265 (272).

[25] Vgl. Streng, in: Kindhäuser/Neumann/Paeffgen, § 46 Rn. 120, 202. Zur Methode der vergleichenden Strafzumessung s. Rostalski/Völkening, KriPoZ 2019, 265 (268 ff.); Meier, Sanktionen (Fn. 15), S. 241 ff.; Maurer, Komparative Strafzumessung, 2005, 190 ff., 209 ff.

b) Informationsbereitstellung

Um ein möglichst vollständiges und aktuelles Bild von der Strafzumessungspraxis zu erhalten, müssen Richter*innen die von ihnen gefällten Urteile kontinuierlich in die Datenbank einpflegen. Die wesentliche Neuerung gegenüber den bisherigen Vorschlägen[26] besteht darin, dass die Erfassung des Inhalts der Entscheidungen nicht von Hand, sondern automatisiert erfolgt. Richter*innen müssen die (ggf. automatisiert[27]) anonymisierten Urteile also nur hochladen. Smart Sentencing markiert die Textstellen, die für die Strafzumessung von Bedeutung sind, einschließlich der entsprechenden Klassifikationen. Um technische Fehler auszuschließen, sollten die Ergebnisse noch einmal kontrolliert werden. Der Arbeitsaufwand ist vergleichbar zur Rechtschreibprüfung in Word.

2. Aus Sicht der (Staats-)Anwält*innen

Die übrigen Verfahrensbeteiligten können das System einerseits wie die Richter*innen als Orientierungshilfe für das (zu erwartende) Strafmaß verwenden, um ihre Anträge und die Beratung der Mandant*innen entsprechend anzupassen. Andererseits gibt die Analyse des Einflusses einzelner Strafzumessungsfaktoren Anhaltspunkte für kritische Aspekte, denen bei der Beweisführung besondere Aufmerksamkeit zukommen sollte. Die Daten können sich auch auf die Prozesstaktik auswirken, wenn etwa bekannt ist, in welchem Maße und unter welchen Umständen sich ein Geständnis für den oder die Mandant*in strafmildernd auswirken würde. Zurzeit sind (Staats)Anwält*innen für diese Informationen auf Erfahrungswissen angewiesen, das junge oder mit den örtlichen Modalitäten nicht vertraute Jurist*innen bestenfalls aus zweiter Hand haben können.

Schließlich kann Smart Sentencing Auswirkungen auf die Revision haben. Der BGH hat wiederholt entschieden, dass eine erhebliche Abweichung vom in vergleichbaren Fällen üblicherweise vergebenen Strafmaß einen Revisionsgrund darstellen *kann*.[28] Was das übliche Strafmaß ist, richtet sich bisher vor allem nach der persönlichen Erfahrung der obersten Bundesrichter*innen.[29] Smart Sentencing ermöglicht es, die Vorgaben des BGH

[26] S. Kaspar, NJW-Beil 2018, 37 (40); Streng, in: Kindhäuser/Neumann/Paeffgen, § 46 Rn. 201 f.; Streng, StV 2018, 593 (599).
[27] S. z.B. Raj/D'Souza, IJMET 12 (2021), 25 (26 ff.) zu verschiedenen Ansätzen zur automatisierten Anonymisierung von Texten in natürlicher Sprache.
[28] BGH, StV 1986, 57 (57); vgl. StV 1995, 173 (173 f.), ebenso BayObLG, JR 2002, 166 (167). Vgl. auch Meier, Sanktionen (Fn. 15), S. 244, außerdem Streng, in: Kindhäuser/Neumann/Paeffgen, § 46 Rn. 120, 201; Maier, in: MüKo-StGB, 4. Aufl. (2020), § 46 Rn. 103; Mellinghoff, in: FS-Hassemer, 2010, S. 503 (518). Ausführlich Maurer, Strafzumessung (Fn. 25), S. 139 ff. Vgl. zum Ganzen auch Rostalski/Völkening, KriPoZ 2019, 265 (268).
[29] Maurer, Strafzumessung (Fn. 25), S. 140 m.w.N; Kaspar/Höffler/Harrendorf, NK 32 (2020), 35 (47). Eine bei Maurer genannte Ausnahme stellt BayObLG, JR 2002, 166 (167) dar, wo statistische

in objektiv nachprüfbarer Weise zu erfüllen und so für mehr Strafgerechtigkeit zu sorgen.[30]

3. Aus Sicht der Wissenschaft und der Gesellschaft

Das System ist aber nicht nur für die Verfahrensbeteiligten, sondern auch für die Wissenschaft und die Gesellschaft im Ganzen von Nutzen. Der Zugang soll uneingeschränkt möglich sein.[31] Die einzigartige, bei verbreiteter Nutzung repräsentative Datensammlung erlaubt es beispielsweise, die vielfach diskutierte[32] Variation in der Strafzumessungspraxis besser zu beurteilen. In früheren Untersuchungen von Gerichtsentscheidungen konnte nicht immer hinreichend berücksichtigt werden, ob abweichende Strafzumessungsentscheidungen durch sachliche Unterschiede gerechtfertigt waren.[33] Smart Sentencing ermöglicht dank der inhaltlichen Analyse der in den Urteilen angegebenen Strafzumessungsfaktoren, zwischen nicht gerechtfertigten Ungleichbehandlungen und sachlich notwendigen Differenzierungen zu unterscheiden.

Neu ist auch die Möglichkeit, den Einfluss einzelner Strafzumessungsfaktoren zu isolieren. Dadurch kann erstmals nicht nur das *System* der Strafzumessung, sondern auch die Bedeutung einzelner *Faktoren* umfassend beurteilt werden. Das ermöglicht einen juristischen und gesellschaftlichen Diskurs über die Angemessenheit ihres jeweiligen Einflusses und ggf. notwendige Korrekturen, seien sie legislativer oder judikativer Art.

III. Technische Umsetzung des Projekts Smart Sentencing

Das Projekt Smart Sentencing fußt auf Methoden des Machine Learnings und im Besonderen auf modernen Architekturen neuronaler Netze. Die konkret genutzte Technologie im Bereich neuronaler Netze heißt „Transformer“, das dafür verwendete

---

Auswertungen zurate gezogen werden. Mangels hinreichender Datengrundlage werden dort aber keine tatsächlichen Unterschiede zwischen den Entscheidungen berücksichtigt.

[30] Allgemein dazu Rostalski/Völkening, KriPoZ 2019, 265 (270 ff.). Vgl. außerdem Kaspar/Höffler/Harrendorf, NK 32 (2020), 35 (47); Streng, in: Kindhäuser/Neumann/Paeffgen, § 46 Rn. 202.

[31] Denkbar ist allerdings ein für die Allgemeinheit offenes Bezahlmodell nach dem Vorbild von juris, beck-online etc.

[32] Etwa beim Deutschen Juristentag 2018, dazu z.B. Kaspar/Höffler/Harrendorf, NK 32 (2020), 35 (47) m.w.N.

[33] Insb. bei Grundies, in: Hermann/Pöge, Kriminalsoziologie, 2018, S. 295 (299 f.). Andernorts (Albrecht, Strafzumessung bei schwerer Kriminalität, 1994, S. 349 ff.) wurden inhaltliche Unterschiede zwar berücksichtigt. Die dazu notwendige händische Auswertung führte jedoch zu einer erheblichen Beschränkung der Datenbasis, sodass die Untersuchung zumindest nicht für das gesamte Strafrecht repräsentativ ist (s. Albrecht, Strafzumessung (Fn. 33), S. 237 ff. zur Beschränkung auf Fälle schwerer Kriminalität und zur Repräsentativität in diesem Bereich).

Modell heißt *German BERT*[34]. Der Vorteil von Machine-Learning-Methoden ist die große Effektivität im Lösen von Problemstellungen, die als Klassifizierungs- oder Regressionsaufgaben formuliert werden.

Ein für Smart Sentencing geeignetes Programm muss in der Lage sein, bestimmte vordefinierte Informationen innerhalb der Strafurteile zu erkennen und die entsprechenden Passagen (etwa die Schadenshöhe oder das Strafmaß) zu markieren, damit diese weiterverarbeitet oder visuell aufbereitet werden können. Die Hürden dieses Ansatzes liegen in der Erkennung der Faktoren: woher weiß das Programm, welche Passagen extrahiert werden sollen? Wie verhindert man Verwechselungen? Was geschieht bei einer falschen Erkennung?

Die Umsetzung des Projekts erfolgt in drei Phasen:

- Datenbeschaffung und -aufbereitung

  Zunächst wird eine Datenbasis aus Strafurteilen angelegt. Die verwendeten Urteile müssen gewisse Anforderungen erfüllen, die teils durch die Methode der Informationsextraktion, teils durch äußere Umstände (Anonymisierung) vorgegeben sind.

- Informationsextraktion

  Die automatisierte Extraktion von Informationen aus semi- oder nicht strukturierten Textquellen ist ein Problem, für das noch keine universelle Lösung entwickelt wurde. Das Problem besteht darin, dass menschliche Sprache, selbst formalisiert wie in der Form eines Strafurteils, ein hohes Maß an Komplexität und Varianz besitzt. Die Erkennung von Paraphrasen etwa stellt eine Maschine vor große Schwierigkeiten, so wie auch die Einordnung von Referenzen und sehr lange Textpassagen. Bis zum jetzigen Zeitpunkt gibt es kein System, das eine echte Konversation führen oder diese auch nur fehlerfrei transkribieren kann. Diese Schwierigkeiten gelten auch für geschriebene Sprache, doch Durchbrüche in den Bereichen von Machine Learning und neuronalen Netzen haben das Repertoire der Analysewerkzeug für menschliche Sprache in den letzten Jahren maßgeblich erweitert.

- Auswertung der Ergebnisse

  Teil des Extraktionsvorgangs von Informationen ist auch die Extrapolation und Darstellung der dabei gewonnen Einsichten. Entsprechend ist die Auswertung ein

[34] Siehe: https://huggingface.co/bert-base-german-cased, zuletzt abgerufen am 07.07.2021.

wichtiger Aspekt des Vorgangs, besonders wenn die Ergebnisse von Anwender*innen für die Recherche verwendet werden sollen.

Die drei genannten Phasen werden folgend anhand des Beispielprojektes beschrieben, um die Umsetzung einerseits anschaulich zu gestalten und andererseits Ansätze für weitere Forschung in diesem Bereich aufzuzeigen.

1. Daten

Für eine Textanalyse und die daran angeschlossene Informationsextraktion wird eine Basis an verlässlichen, bereinigten Daten benötigt, der sogenannte Goldstandard. Auf einem Teil dieses Standards, dem Trainingssatz, werden die Analysemethoden entwickelt und anschließend auf einem kleineren Teil der Daten, dem Testsatz, validiert und getestet.[35] Im Machine Learning ist es dabei wichtig, Trainings- und Testsatz streng zu trennen, um zu verhindern, dass das Modell die Beispiele auswendig lernt, ohne dabei echte Schlussfolgerungen zu ziehen (genannt *Overfitting*).[36]

Die Datenbasis besteht im Beispielprojekt aus einer Sammlung von etwa 140 Strafgerichtsurteilen, die beim Amtsgericht Leipzig angefragt und als PDF-Dateien zur Verfügung gestellt wurden. Es wurden dabei für den Prototypen des Projektes ausschließlich Strafurteile berücksichtigt, die den Tatbestand des Diebstahls erfüllen und nur eine*n Angeklagte*n betreffen. Diese Einschränkungen wurden vorgenommen, um eine konsistente Ausgangssituation für die Analyse zu schaffen. Sowohl eine Generalisierung über mehrere Tatbestände als auch eine Korrelierung der Daten verschiedener Angeklagter in demselben Strafurteil hätte zusätzliche Komplexität zur Folge gehabt, die die erste Implementierung deutlich erschwert hätte. Für die weitere Entwicklung sollten diese Einschränkungen schrittweise aufgehoben werden.

Die Strafurteile wurden anschließend von PDF-Dateien in das CAS XML-Datenformat[37] umgewandelt, das eine Anreicherung der Volltexte um Metainformationen ermöglicht. Unter anderem ist es mit diesem Datenformat möglich, die genaue Position eines Satzes oder eines Wortes innerhalb des Textes auszumachen und Textspannen mit Annotationen zu versehen. Das Einfügen dieser Annotationen ist ein wichtiger Aspekt, da die Anwendung von Machine-Learning-Verfahren auf Textdaten meist voraussetzt, dass

[35] Typischerweise werden Trainings- und Testsatz um einen Validierungssatz ergänzt, der verwendet wird, um die Hyperparameter (die ”Einstellungen”) eines Modells feineinzustellen. Der Kürze halber wird an dieser Stelle nicht weiter darauf eingegangen.
[36] Shalev-Shwartz/Ben-David, Understanding Machine Learning: From Theory to Algorithms, 2014, S. 35.
[37] CAS: Common Analysis System, ein vom UIMA Framework bereitgestelltes Datenformat. XML: Extensible Markup Language, verbreitetes Datenformat zur Encodierung von Textdokumenten.

die zu erkennenden Informationen für das Trainingsverfahren entsprechend annotiert sind.[38] Im Beispielsatz

```
Strafrechtlich ist er in Deutschland bislang nicht belangt worden.
```

wird die Information, dass der Angeklagte nicht vorbestraft ist, gekennzeichnet, indem die relevante Textpassage mit einer Annotation versehen wird. Eine Annotationsform unter Verwendung der Wortposition im Satz könnte dann folgendermaßen aussehen:

[5, 7, Vorstrafe_nein][39].

Die Strafurteile wurden manuell und auf der Grundlage einer Auswahl bestimmter Faktoren annotiert. Die Auswahl dieser Faktoren sollte dabei die maßgeblichen Aspekte des Strafurteils abbilden (etwa Schadenshöhe, Vorbestrafung und Geständnisse). Diese Auswahl stellt gleichzeitig eine Hypothese über den Einfluss dieser Faktoren auf das Strafmaß dar und könnte im weiteren Verlauf angepasst werden, falls sich diese Darstellung als fehlerhaft oder unvollständig erweist.

Die Faktoren enthalten dabei eine Bandbreite unterschiedlicher Datentypen. Die Passage, in der die Vorstrafe erwähnt wird, ist beispielsweise immer ein Textabschnitt („bislang nicht belangt"), während das Strafmaß immer Zahlen beinhaltet („10 Tagessätze zu je 30 Euro"). Letzteres ermöglicht eine Quantifizierung und somit auch eine numerische Auswertung, etwa die Bildung von Summen und Durchschnittswerten.

Die Datenbasis war damit vollständig und sauber aufbereitet. Auch wenn die Anzahl der Dokumente (ca. 140) durchaus gering war für das Anlernen eines Modells, war dies eine fast optimale Ausgangslage der Daten.

2. Informationsextraktion

Der Begriff der Informationsextraktion beschreibt zunächst das Herausfiltern relevanter Informationen aus einer Sammlung von Daten. Dies können sowohl Texte als auch Bilder oder Sprachaufnahmen sein. Der präzisere Begriff der Textklassifizierung meint eine spezifische Aufgabe der Informationsextraktion im Bereich des *Natural Language Processing*: der Verarbeitung menschlicher Sprache.[40]

[38] Es wird zwischen supervised, unsupervised und self-supervised Learning unterschieden: die primäre Differenz ist die (Nicht-)Verwendung von Annotationen. Der Kürze halber wird dies hier nicht weiter erläutert.
[39] Der Index beginnt in der Programmierung konventionell bei 0, daher entsprechen die Worte "bislang nicht belangt" den Indizes 5, 6 und 7. "Vorstrafe_nein" ist die dazugehörge Klasse, das Label.
[40] Jurafsky, Speech and Language Processing, 2019, S. 56.

Generell bedeutet Klassifizierung im Kontext von Machine Learning die automatisierte Einordnung eines neuen Beispiels aufgrund der Beobachtung einer großen Zahl vorheriger Beispiele (etwa eine Bilderkennung, die durch das Anlernen mit zahlreichen Tierbildern zwischen den Abbildungen von Hunden und Katzen unterscheidet).[41] Eine weitverbreitete Methode für die Durchführung von Klassifizierungsaufgaben sind neuronale Netze:[42] Verbindungen aus hintereinandergeschalteten, gewichteten Knoten, die mit einer bestimmten Sammlung von Beispielen (etwa Katzenbildern, Videoschnipseln, Strafurteilen) initialisiert werden und die anschließend mit der numerischen Darstellung (der Encodierung) eines neuen Beispiels gespeist werden können, um eine Klassifizierung anhand der Ähnlichkeit zu den vorherigen Beispielen durchzuführen. In der Praxis bedeutet dies, dass ein Machine-Learning-Modell, das mit drei Millionen Bildern von Katzen und Hunden angelernt wurde, zwar sehr gut bei der Erkennung von Katzen und Hunden funktioniert, eine Erkennung von Papageienbildern jedoch völlig fehlschlägt. Dies unterstreicht das zentrale Problem dieses Ansatzes: die Fähigkeit von Trainingsbeispielen, über neue Beispiele generalisieren zu können.

Um diese Fähigkeit zur Generalisierung bestmöglich auszubilden, werden neuere Architekturen neuronaler Netze in der Sprachverarbeitung nicht mit einzelnen Sätzen initialisiert, sondern mit ganzen Sprachkorpora. Das Grundmodell der BERT-Architektur[43] wurde beispielsweise u.a. mit der gesamten englischen Wikipedia (2,5 Milliarden Wörter) angelernt, um eine „natürliche Sprachintuition" in den Knoten des Netzes zu erzeugen, die für die verschiedensten Aufgaben nützlich ist.[44] Das deutsche Modell der BERT-Architektur (German BERT) wurde dementsprechend auf der deutschen Wikipedia sowie einem frei zugänglichen Korpus von deutschen Gerichtsurteilen angelernt[45], womit bereits ein grundlegendes Verständnis[46] von juristischen Texten im Modell angelegt ist. Aus diesem Grund und wegen der hohen Erkennungsraten in verschiedenen Vergleichstests wurde das German BERT-Modell als Verfahren für das Smart-Sentencing-Projekt ausgewählt.

[41] Es sei darauf hingewiesen, dass das Feld des Machine Learning weit mehr als nur die Methode der neuronalen Netze umfasst.
[42] Es sei darauf hingewiesen, dass das Feld des Machine Learning weit mehr als nur die Methode der neuronalen Netze umfasst.
[43] Bidirectional Encoder Representations from Transformers, eine Abwandlung der Transformer-Architektur.
[44] S. Qiu et al., Pre-trained models for natural language processing, 2020, S. 5.
[45] Chan//Schweter//Möller, German's next language model, 2020, S. 2.
[46] Dies darf nicht als semantisches Verständnis im menschlichen Sinne verstanden werden. Es handelt sich eher um eine Familiarität mit den numerischen Ausprägungen, die der Aufbau und das Vokabular juristischer Texte in den Knoten des Netzes hinterlassen.

Die Anwendung desselben umfasste letztlich nur eine Feineinstellung der Modellparameter durch ein vergleichsweise kleines Training auf den Strafurteilen der Datenbasis. Nach wenigen Durchläufen gab das Modell auf diese Weise bereits die Ergebnisse der Extraktion der relevanten Faktoren aus.

3. Auswertung

Die Leistung der Erkennung in Klassifizierungsaufgaben wird über verschiedene Werte angegeben: *precision* (Genauigkeit, Richtigkeit der Erkennung), *recall* (Trefferquote, Vollständigkeit der Erkennung) und den *f1-score* (harmonisches Mittel beider vorherigen Werte).[47] Der f1-score ist damit eine Art Gesamtindikator. *Support* listet die Anzahl der jeweiligen Klassen, die in den Testbeispielen gefunden wurden. Die Ergebnisse entsprechen den Probeläufen des Prototypen und spiegeln noch nicht die Leistungsfähigkeit der vollständigen Anwendung wider.

	precision	recall	f1-score	support
Vorstrafe_nein	1.00	1.00	1.00	3
Geldstrafe_Höhe	0.93	0.93	0.93	14
Datum	0.93	0.87	0.90	27
Ort	0.87	0.77	0.82	26
Schadenshöhe	0.73	0.79	0.76	52
Geständnis_ja	0.70	0.78	0.74	27
Vorstrafe_ja	0.72	0.72	0.69	5
Freiheitsstrafe	0.60	0.60	0.60	15
Geldstrafe_Tagessätze	0.71	0.45	0.56	33
Tatbestand_Paragraph	0.46	0.48	0.47	23
Drogenabhängigkeit	0.46	0.42	0.47	31
Tatbestand_Beschreibung	0.39	0.33	0.36	21

Insgesamt sind diese Ergebnisse für die geringe Datenmenge und den Anwendungsfall sehr gut. Die meisten Erkennungen könnten bereits mit nur geringer Prüfung in eine Datenbank geladen werden und dort zur Recherche in den Strafurteilen dienen.

Die geringen Werte in den unteren Klassen haben verschiedene Erklärungen. Teils waren in den Trainingsbeispielen nicht ausreichend viele Passagen für das Lernen einer verlässlichen Erkennung vorhanden, teils waren die Textabschnitte (im Besonderen Beschreibungen) schlicht zu lang, um vollständig erkannt werden zu können. Doch auch, wenn die Prozentwerte für diese Klassen schwanken, sind die Ergebnisse der Erkennung selbst in den schlecht erkannten Klassen durchaus wertvoll und verwendbar. Zur Einordnung: Eine fehlerhafte Erkennung der Beschreibung eines Tatbestandes[48] könnte in der Textstelle

[47] Jurafsky, Speech and Language Processing, 2019, S. 66.
[48] Der Begriff wird hier im untechnischen Sinne verwendet und umfasst auch besonders oder minder schwere Fälle.

`Diebstahls im besonders schweren Fall`

lediglich die Passage

`Diebstahls im besonders schweren`

markieren und würde damit zwar keine vollständige Einordnung geben, zumindest aber einen starken Hinweis auf die relevante Textstelle.

Damit ist die Grundlage für eine Datenabfrage über ein visuelles Interface geschaffen, etwa in Form einer Website mit Suchfunktion. Dieser Aspekt der Umsetzung betrifft allerdings nicht mehr den Vorgang der Informationsextraktion, sondern den des Interfacedesigns, und wird deshalb an dieser Stelle nicht näher behandelt.

IV. Fazit

In Deutschland bestimmen nicht nur sachlich relevante Umstände die konkrete Strafzumessungsentscheidung, sondern auch die Person des oder der individuellen Richter*in und der Entscheidungsort. Daraus ergeben sich Differenzierungen, die unter dem Gesichtspunkt des Gleichheitssatzes Schwierigkeiten aufwerfen.

Vor diesem Hintergrund zielt das Projekt Smart Sentencing auf die Schaffung einer Datenbank, die alle in Deutschland ergangenen und noch ergehenden Strafzumessungsentscheidungen sammelt und nach den enthaltenen Strafzumessungsfaktoren kategorisiert. Damit wird den Richter*innen eine Möglichkeit gegeben, sich ein objektives Bild von der Strafzumessungspraxis zu machen. Auch den übrigen Verfahrensbeteiligten verspricht das System eine bessere Entscheidungsgrundlage dank mehr Transparenz und Rationalität. Schließlich ermöglicht die Datenbank eine gesellschaftliche und wissenschaftliche Diskussion über die bestehende Praxis.

Die bisherigen Ergebnisse zeigen, dass eine händisch nicht zu stemmende Analyse sämtlicher Strafzumessungsentscheidungen mithilfe maschinellen Lernens möglich ist. Bisher liegt zwar nur eine relativ kleine Datenbasis vor. Der hiermit trainierte Algorithmus liefert aber bereits ermutigende Ergebnisse. Wenn es gelingt, die Datengrundlage zu vergrößern, indem weitere Urteile gesammelt und in das Modell eingespeist werden, bietet Smart Sentencing eine Perspektive für mehr Strafgerechtigkeit.