# JURISTISCHE AUSBILDUNG IN ZEITEN DER DIGITALISIERUNG

**DIRK HARTUNG**
Executive Director Legal Technology, Bucerius Law School, Hamburg

Stichworte: Legal Technology, Juristenausbildung, Digitalisierung, IBM Watson

Der Beitrag beschreibt, wie in einem Kurs an der Bucerius Law School in Hamburg Studierende der Rechtswissenschaft und Informatik gemeinsam eine webbasierte Anwendung zur Auswertung von Gerichtsurteilen entwickelt haben. Er zeigt, wie schnell Jurastudierende technisch sprachfähig werden, juristische Analysen mit IBM Watson skalieren, und gibt ein gutes Gefühl dafür, was derzeit technisch machbar ist. Ziel ist es, die Diskussion um Legal Technology um konkrete Erfahrungen zu bereichern und zu einem optimistischen Blick auf das Thema anzuregen.

## I. Digitalisierung des Rechtsdienstleistungsmarktes

Durch die Digitalisierung wird sich für zukünftige Juristinnen und Juristen vieles verändern.[1] Dabei stellen die gegenwärtigen Entwicklungen nur eine Vorstufe zu den gesamtgesellschaftlichen Produktivitätssteigerungen dar. Dies liegt im Wesentlichen im Fortschritt bei der Erfassung und Analyse grosser Datenmengen *(Big Data)* und der immensen Kapazitätssteigerung bei der Komplexitätsbewältigung von Computern *(Machine Learning)* begründet. Dahinter wiederum stehen bemerkenswerte technische Fortschritte bei Prozessoren, insbesondere Grafikprozessoren,[2] die zu deutlich höherer Rechenleistung bei gleichzeitig niedrigeren Kosten geführt haben. Schliesslich werden die Fortschritte im *Machine Learning* dadurch begünstigt, dass grosse Mengen an Daten zur Verfügung stehen. Vieles, was bislang nicht erfass- oder messbar war, kann nun ausgewertet und zum Training von *Machine Learning*-Algorithmen verwendet werden.[3]

Diese Veränderungen betreffen zum ersten Mal auch Tätigkeiten, die bisher aufgrund der grossen Abhängigkeit von der menschlichen Sprache als für Computer nicht zu bewältigen galten. Dazu gehören juristische Tätigkeiten und Rechtsdienstleistungen aller Art. Dabei verändert sich weniger, was Juristen im Kern tun: Sie managen rechtliche Risiken und helfen bei der Bewältigung daraus entstehender Komplexität.[4]

In Zukunft wird sich allerdings durchaus verändern, wie sie diese Aufgabe bewältigen und mit wem sie dafür zusammenarbeiten. Das Spektrum juristischer Tätigkeit reicht dabei von der Haftung bei Unternehmenskäufen über die Rückfälligkeitswahrscheinlichkeit zu verurteilender Straftäter bis zur Gestaltung von Regeln für neue Technologien. All diese Aufgaben werden durch die Digitalisierung sowohl qualitativ als auch quantitativ anspruchsvoller, da die Menge der Daten und Interaktionen deutlich zunimmt, wenn der digitale Raum erschlossen wird. Gleichzeitig erweitert sie die Handlungsmöglichkeiten, ermöglicht neue Formen der Arbeitsteilung und Zusammenarbeit und führt zu ganz neuen Produkten juristischer Tätigkeit.[5]

Wenn wir junge Menschen, die sich für einen juristischen Beruf entscheiden, angemessen auf die Zukunft vorbereiten und ihnen ermöglichen wollen, an den digitalen Wertschöpfungsprozessen teilzuhaben, müssen sich technologische Inhalte auch in der juristischen Ausbildung wiederfinden. Dies gilt umso mehr, wenn man bedenkt, dass heutige Studienanfänger – in der Schweiz wie in Deutschland – erst in frühestens sechs bis sieben Jahren ihre Ausbildung beenden und auf den Arbeitsmarkt treten werden. Angesichts der rasanten Entwicklungsgeschwin-

1 Sofern nachfolgend die männliche Form verwendet wird, umfasst diese Bezeichnung selbstverständlich Frauen und Männer.
2 Diese auf den ersten Blick ungewöhnliche Art der Nutzung von Grafikchips wird als General Purpose Computation on Graphics Processing Unit (GPGPU) bezeichnet.
3 Zum starken Anstieg der Datenmengen im Internet: STEPHANIE PAPPAS, How Big Is The Internet Really?, 2016, http://www.livescience.com/54094-how-big-is-the-internet.html (besucht am 2. 6. 2017).
4 Etwas traditioneller könnte man das auch als Erteilung von Rechtsrat bezeichnen.
5 Für eine ausführliche Untersuchung s. VEITH/BANDLOW/HARNISCH/WENZLER/HARTUNG/HARTUNG, How Legal Technology Will Change the Business of Law, 2016, <http://buceri.us/legaltechstudy2016> (besucht am 2. 6. 2017).

digkeit der oben genannten Technologien wird die juristische Arbeitswelt 2024 noch viel stärker von Technologie beeinflusst sein.

## II. Die Integration digitaler Inhalte in die juristische Ausbildung

Bei der Bewältigung dieser Veränderungen geht es juristischen Ausbildungseinrichtungen ähnlich wie der Anwaltschaft: Wir befinden uns auf weitgehend unbekanntem Terrain und müssen unter unsicheren Rahmenbedingungen dennoch handeln. Es versteht sich von selbst, dass die materiell- und formellrechtliche Ausbildung nicht vernachlässigt werden darf – Juristen sind zu allererst Domänenexperten für rechtliche Fragestellungen. Daneben müssen sie jedoch sowohl mit der Digitalisierung ihres eigenen Berufsstandes als auch mit einer Digitalisierung des gesamten wirtschaftlichen Umfelds, das sich auf die von ihnen zu bearbeitenden Sachverhalte auswirkt, umgehen können. Dies gilt einheitlich für Anwälte, Richter, Verwaltungs- und Unternehmensjuristen.

Statt einer abstrakten Beschreibung, wie sich solche Inhalte in die Ausbildungssysteme verschiedener Länder einfügen lassen und wie das inneruniversitäre Veränderungsmanagement gelingt,[6] geht dieser Beitrag einen anderen Weg. Anhand eines konkreten Erfahrungsberichts mit einem alternativen Kursformat, das sich mit technologischen Inhalten beschäftigt, möchte der Verfasser zum Ausprobieren inspirieren. Nur so können wir Erfahrungen in dieser häufig noch unbekannten Materie sammeln und Stück für Stück ein modernes Studium zusammenstellen, das der juristischen Lebenswirklichkeit im 21. Jahrhundert Rechnung trägt.

## III. Wie man IBM Watson Jura beibringt – eine Fallstudie

### *1. Einführung*

In einer Kooperation des Autors mit Prof. Dr. Chris Biemann, Leiter der Sprachtechnologiegruppe der Fakultät für Informatik an der Universität Hamburg, wurde im Frühjahr 2017 ein gemeinsamer, dreiwöchiger Kurs mit dem Titel *«Hands on Machine Learning in Law»* als Pilotprojekt angeboten.

Der Kurs war für die Informatiker in ein verpflichtendes Softwarepraktikum und für die Juristen in ein freiwilliges Projektstudium im Rahmen des Studiums Generale der Bucerius Law School eingebunden. Die Kooperation wurde durch IBM, insbesondere die Abteilung für Wissenschaftsbeziehungen und die *Cognitive Solutions Unit* in der DACH-Region personell vermittelt und technologisch unterstützt. Insgesamt fanden sich jeweils sechs Studierende beider Disziplinen, aus dem zweiten, dritten und vierten Studienjahr, die überwiegend keine Vorkenntnisse im jeweils fremden Fach hatten.

Die Aufgabe bestand darin, unter Verwendung von Werkzeugen aus den Watson bzw. IBM Bluemix Services eine Anwendung zu schaffen, die aus einem Corpus von ca. 40 000 frei verfügbaren Entscheidungen des Bundesgerichtshofs in Karlsruhe einen Mehrwert schafft.[7] Die weitere Ausgestaltung war nicht vorgegeben und das Projektmanagement bewusst den Studierenden überlassen. Die studentischen Arbeitsgruppen wurden von wissenschaftlichen Mitarbeitern nach der SCRUM-Methode durch regelmässige Sprint-Meetings und jeweils nach der Hälfte der Zeit und am Ende bei einer Produktpräsentation unterstützt. Abschliessend verfassten die Studierenden Projektberichte, die neben der programmierten Anwendung und den Präsentationen Grundlage der Bewertung waren.

Ziel des Pilotprojektes war es, herauszufinden, wie gut interdisziplinäre Teams aus Softwareentwicklern und Juristen gemeinsam Projekte bewältigen können. Im Vordergrund stand daher nicht die Qualität des entwickelten Softwareprodukts, sondern die Interaktion der Studierenden untereinander und der Erwerb interdisziplinärer Sprachfähigkeit.

### *2. Phase 1: Vorbereitung der juristischen Szenarien*

In Vorbereitung auf das Projekt wurden die juristischen Teilnehmer zunächst mit den ca. 40 000 Entscheidungen konfrontiert und erhielten die Aufgabe, eine Präsentation zur Vorstellung möglicher Ansätze zur Analyse der Fälle zu entwerfen, um das Projekt damit bei den Informatikern vorzustellen. Da sich diese den Inhalt ihrer Aufgabenstellung für das Softwarepraktikum unter mehreren Szenarien aussuchen sollten, musste ausserdem dargelegt werden, wieso sich eine Auseinandersetzung mit der juristischen Materie lohnt.

Dazu wurden die verfügbaren Fälle grob systematisiert und verschiedene mögliche Nutzungsszenarien diskutiert. Im ersten Schritt wurden – völlig unabhängig von der technischen Umsetzung – Vorschläge dazu gesammelt, welche Informationen aus den Urteilen extrahiert werden könnten. Diese reichten von einfachen Metadaten wie dem Aktenzeichen, dem entscheidenden Gericht, dem Spruchkörper und dessen Angehörigen über die zitierten Normen, Referenzentscheidungen und zitierte Literatur bis zur Dauer des Verfahrens, zum materiellen Ausgang und zu weiteren Entscheidungen wie der Strafzumessung und der Verteilung der Verfahrenskosten.

Auf Grundlage dieser Extraktionen stellten die Studierenden mögliche Forschungsansätze bzw. Anwendungsszenarien zusammen. Dazu gehörten Suchfunktionen und Datenvisualisierung, Vorhersage bestimmter Faktoren wie der Verfahrensdauer oder Vorhersage der Wahrschein-

6 Für die USA s. KATZ, The MIT School of Law? A Perspective on Legal Education in the 21st Century, University of Illinois Law Review, 2014, S. 1431 ff.; für den deutschsprachigen Raum s. DIRK HARTUNG, Judex Calculat – Neue Berufsbilder und Technologie in der juristischen Ausbildung, in HARTUNG/BUES/HALBLEIB, Legal Tech 2017 (im Erscheinen).

7 Die Urteile stammen von der Website des Bundesgerichtshofs und machen den Grossteil aller kostenfrei verfügbaren Entscheidungen in Deutschland aus.

lichkeit des Obsiegens,[8] Darstellung und Analyse von Entscheidungsnetzwerken, die Überprüfung der offiziellen Statistik des Bundesgerichtshofs[9] und ein Recherchetool in natürlicher Sprache.

Schliesslich machten sich die Juristen mit dem *Watson Knowledge Studio,* einer Anwendung, welche die Annotation von Texten ohne Programmierkenntnisse ermöglicht, vertraut.[10]

### *3. Gemeinsame Entwicklung der Anwendungsideen und Vorbereitung der Texte*

Bei einer gemeinsamen Auftaktveranstaltung präsentierten die Juristen ihre Überlegungen zu möglichen Nutzungsszenarien und führten in die grundlegenden juristischen Begriffe ein. Damit gelang es, insgesamt sechs Studierende der Informatik in zwei Teams für juristische Anwendungen zu gewinnen. Die Studierenden reizte nach ihren Angaben vor allem die Möglichkeit, mit den späteren Nutzern direkt zusammenzuarbeiten, und darüber hinaus die Möglichkeit zur Arbeit mit echten Texten. In der folgenden Diskussion ergaben sich zwei Anwendungsszenarien, die technisch machbar und juristisch inhaltlich zielführend erschienen:

Ein Team sollte eine Anwendung entwickeln, welche die Beziehungen zwischen einzelnen Urteilen, insbesondere gemeinsame Zitate, Urteile und Normen, visualisieren konnte. Dies sollte den Umgang mit der grossen Datenmenge erleichtern und die Grundlage für eine weitere wissenschaftliche Auswertung schaffen. Gleichzeitig sollte die Möglichkeit vorbereitet werden, die Urteile mit Analysemethoden der empirischen Sozialforschung wie der Netzwerkanalyse zu untersuchen.[11]

Das andere Team entschied sich für eine Anwendung, die umfangreiche statistische Auswertungen des Urteilscorpus bieten sollte. Neben der automatischen Extraktion der dafür erforderlichen Informationen, nahmen sich die Studierenden die ambitionierte Aufgabe vor, Auffälligkeiten und Muster bei den Revisionen zu erkennen. So sollte die Frage beantwortet werden, wie sich bestimmte Faktoren wie die Ausgangsgerichte, der zeitliche Ablauf und die rechtlichen Fragen auf den Erfolg der Revisionen auswirken. Dem lag die Überlegung zugrunde, dass dadurch (prozess)taktische Erwägungen, beispielsweise bei der Wahl des Gerichts in Fällen eines sog. fliegenden Gerichtsstands oder bei Gerichtsstandvereinbarungen, ermöglicht würden.

Für beide Anwendungen mussten also bestimmte Inhalte wie das Aktenzeichen, Normen, Verweise auf andere Urteile und Literatur und der Verfahrensausgang in den einzelnen Urteilen automatisch erkannt und klassifiziert werden. Dazu sollte vor allem *AlchemyLanguage*[12] verwendet werden. Dieses Werkzeug aus den *Watson Services* ermöglicht die Extraktion bestimmter Informationen aus Texten. Um diese zu identifizieren, werden – technisch vereinfacht ausgedrückt – die Texte nach bestimmten Mustern durchsucht. Diese Muster müssen den Algorithmen jedoch erst beigebracht werden und hängen stark von der verwendeten Sprache ab. Dazu werden beim hier verwendeten *Supervised Machine Learning* Datensätze mit Trainingsdaten benutzt, in denen die relevanten Informationen bereits markiert sind. Auf dieser Grundlage identifiziert der *Machine Learning Algorithmus* die Muster in den Trainingsdaten. Dieses Ergebnis – das sog. *Machine Learning Modell* – wird verwendet, um auch bei einer anderen Datengrundlage die gewünschten Informationen zu finden.

Da es für die beschriebenen Anwendungsfälle in juristischer, deutscher Sprache keine bereits trainierten Modelle gab, mussten diese von den Studierenden selbst trainiert werden.

### *4. Vorbereitung der Texte und Training der Modelle*

Beide Anwendungen benötigten also ein Modell, mit dessen Hilfe Aktenzeichen, Entscheidungsart und -datum, am Verfahren beteiligte Gerichte, zeitlicher Ablauf des Verfahrens sowie die zitierten Normen, Urteile und Literaturfundstellen in den einzelnen Urteilen bzw. Beschlüssen gefunden und extrahiert werden konnten. Das zweite Team musste zusätzlich noch sprachliche Indikatoren dafür finden, ob die Revision erfolgreich bzw. nicht erfolgreich war. Dazu mussten die Juristen die entsprechenden Textbestandteile in den Trainingdatensätzen markieren, also einzelne Urteile lesen und die oben genannten Bestandteile bzw. Entitäten annotieren.

Diese Funktion bietet das *Watson Knowledge Studio.* Allerdings lagen die Urteile ausschliesslich als PDF-Dateien vor und mussten zunächst in reine Textdokumente mit der richtigen Kodierung[13] umgewandelt werden.[14] Da dieses Projekt eines der ersten öffentlichen Projekte in deutscher Sprache war, mussten einige Anlaufschwierigkeiten überwunden werden. Für die Darstellung der Texte zur anschliessenden Annotation ging das Watson Knowledge Studio beispielsweise davon aus, dass Punkte (also das Zeichen «.») exklusiv zwei Sätze von einander trennt. Während für das amerikanische Datumsformat eine Ausnahme bestand, mussten alle Daten in den deutschen Urteilen umformatiert werden, sodass aus dem 10. 05. 2011 der 10-05-2011 wurde. Ein ähnliches Phänomen

**8** Die Vorhersage aufgrund statistischer Auswertungen bezeichnet man als Predictive Analytics, s. <https://en.wikipedia.org/wiki/Predictive_analytics> (besucht am 2. 6. 2017).

**9** Zur offiziellen Statistik s. <http://www.bundesgerichtshof.de> unter Service/Statistik für die Statistik nach Senaten (besucht am 2. 6. 2017).

**10** Weitere Details s. <https://www.ibm.com/de-de/marketplace/supervised-machine-learning> (besucht am 2. 6. 2017).

**11** Ein Beispiel für diesen vielversprechenden Ansatz zur Rechtsprechungsanalyse bietet RYAN WHALEN, Legal Networks: The Promises and Challenges of Legal Network Analysis, Michigan State Law Review 2016, S. 539 ff.

**12** Seit Kurzem heisst der Dienst Watson Natural Language Understanding Service, s. <https://www.ibm.com/watson/developercloud/alchemy-language.html> (besucht am 2. 6. 2017).

**13** Das Watson Knowledge Studio benötigt Texte in UTF-8 mit maximal 40 000 Zeichen inkl. Leerzeichen.

**14** PDF-Dokumente enthalten häufig weitaus mehr Informationen als den reinen Text, um die gleiche Anzeige unabhängig vom verwendeten Programm zu ermöglichen. Für die Verarbeitung mit dem Watson Knowledge Studio müssen sie allerdings gelöscht werden.

trat bei Abkürzungen mit einem Punkt auf, wobei insbesondere juristische Abkürzungen (f., ff., a. a. O. usw.) einzeln eingepflegt werden mussten.

Im nächsten Schritt mussten die Studierenden eine ausreichende Menge an Entscheidungen annotieren. Was zunächst nach einer eher einfachen Aufgabe klingt, erwies sich als arbeits- und zeitintensivste Projektkomponente für die Juristen.

Um ein Gefühl für den Zeitaufwand zu bekommen, ist es sinnvoll, sich folgende Werte zu verdeutlichen: Gute Annotatoren schafften am Ende der zwei Wochen durchschnittlich 25 bis 30, langsamere zwischen 15 und 20 Entscheidungen pro Stunde. Da für ein minimal zuverlässiges Modell einige Hundert Entscheidungen annotiert werden mussten, bedeutete dies viele Stunden Arbeit. Nach unseren Erfahrungen sinkt die Qualität der Annotationen nach etwa zwei Stunden so signifikant, dass die Annotationen die Qualität des Modells verschlechtern. Spätestens nach diesem Zeitraum sollte daher eine Pause eingelegt werden. Ausserdem stellten die Studierenden nach den ersten Annotationseinheiten schnell fest, dass unterschiedliche Annotatoren sehr unterschiedliche Textbestandteile markierten. Da für die Qualität des Modells möglichst eindeutige Muster und damit konsistente Trainingsdaten erforderlich sind, musste diesem Problem begegnet werden. Gemeinsam mit den Informatikern wurden dazu zwei im *Natural Language Processing* übliche Ansätze gewählt:

Zunächst wurde in den sog. Annotationsrichtlinien möglichst detailliert festgelegt, was jeweils innerhalb einer Entitätenkategorie zu markieren war. Die Richtlinien wurden dabei zwischen den einzelnen Annotationseinheiten zu einem einheitlichen Zeitpunkt basierend auf den bisherigen Erfahrungen und den Qualität-Scores des Modells angepasst. Hier zeigt die Erfahrung, dass es bei den meisten Entscheidungsbestandteilen – wie beispielsweise der Kostenentscheidung – selbst bei einem Corpus von nur einigen Hundert Entscheidungen mehr sprachliche Variationen gibt, als sich die Studierenden vorher vorstellen konnten. Die vermeintlich formalisierte juristische Fachsprache erweist sich jedenfalls im Detail als deutlich individueller als gedacht.

Darüber hinaus wurden die Annotationsdatensätze so zusammengestellt, dass ein bestimmter Anteil der Entscheidungen (nach einigen Experimenten brachten etwa 20% Überschneidungen zwischen zwei Annotationssätzen gute Ergebnisse) von jeweils mindestens einem anderen Annotator bearbeitet wurde. Die Annotationen dieser Entscheidungen wurden dann verglichen und gemeinsam eine Variante für verbindlich erklärt. Damit wurden einerseits Widersprüche in den Trainingsdaten beseitigt und andererseits unterschiedliche Interpretationen der Annotationsrichtlinien identifiziert, sodass diese anschliessend verfeinert werden konnten. Selbst während des kurzen Projektzeitraums konnten die Messwerte[15] für Übereinstimmungen zwischen den einzelnen Annotatoren dadurch soweit verbessert werden, dass bei Berücksichtigung der meisten Kategorien grundsätzliche Übereinstimmung[16] erreicht wurde.

Aufgrund des kurzen Projektzeitraums und dank umfassender entsprechender Funktionen des *Watson Knowledge Studios* erschien es nach ersten Tests sinnvoll, die Texte nicht alleine von Annotatoren markieren zu lassen, sondern nach bestimmten Regeln vorzuannotieren. Dies geschah durch zwei Mechanismen:

Sofern bestimmte Kategorien wie Gerichtsbezeichnungen und -orte (nahezu) abschliessend bekannt waren, konnten Listen (sog. *Dictionaries*) erstellt und die darin enthaltenen Informationen automatisch in den Texten markiert werden. Dafür mussten beispielsweise alle 1561 deutschen Gerichte recherchiert und zusammengestellt werden.

Daneben konnten bestimmte Informationen wie Aktenzeichen und bestimmte Verweise durch reguläre Ausdrücke[17] beschrieben werden. Mit diesen automatischen Vorannotationen konnten die einzelnen Datensätze vom Watson Knowledge Studio markiert werden, bevor sie das erste Mal von menschlichen Annotatoren bearbeitet wurden. In den entsprechenden Kategorien mussten diese dann nur noch überprüfen, ob die Annotationen vollständig waren und wirklich die richtigen Wörter markierten. Dies erleichterte die Arbeit erheblich und führte zur deutlichen Erhöhung der Arbeitsgeschwindigkeit, sodass die oben genannten Entscheidungen pro Stunde erreicht werden konnten.

### *5. Qualität des trainierten Modells*

Unter diesen Voraussetzungen wurden angesichts des kurzen Projektzeitraums zuletzt in den meisten Kategorien qualitativ zufriedenstellende Ergebnisse erreicht, was durch sog. F1-Masse von mehr als 0,8 ausgedrückt wird.[18] Um diesen Wert zu verstehen, ist eine kurze Auseinandersetzung damit erforderlich, wie die Qualität solcher Modelle – also der durch den Algorithmus erkannten Muster im Text – bestimmt wird. Will man bestimmte Informationen in Texten klassifizieren, sind vier Arten von Ergebnissen denkbar:

- richtig positive: Der Textabschnitt gehört zu einer bestimmten Kategorie und wurde als dieser Kategorie zugehörig klassifiziert (eine Zeichenfolge ist tatsächlich ein Aktenzeichen und wurde als solches erkannt).
- Falsch negative: Der Textabschnitt gehört zu einer bestimmten Kategorie, wurde aber nicht als dieser zugehörig klassifiziert (eine Zeichenfolge ist ein Aktenzeichen, wurde aber nicht als solches erkannt).

**15** Die Kennzahlen für generelle Übereinstimmung sind Fleiss' Kappa Scores, s. <https://en.wikipedia.org/wiki/Fleiss%27_kappa> (besucht am 2. 6. 2017).

**16** Als grundsätzliche Übereinstimmung wird ein Fleiss' Kappa Score zwischen 0,61 und 0,80 bezeichnet, im Anschluss an LANDIS/KOCH, The Measurement of Observer Agreement for Categorical Data, Biometrics 1977, S. 159 ff.

**17** Gemeint sind *Regular Expressions*, mit denen sich bestimmte Zeichenfolgen syntaktisch beschreiben lassen, s. <https://de.wikipedia.org/wiki/Regulärer_Ausdruck>.

**18** Vergleichswerte für die F1-Masse finden sich am Ende dieses Abschnitts.

– Falsch positive: Der Textabschnitt gehört nicht zu einer bestimmten Kategorie, wurde aber als dieser zugehörig klassifiziert (eine Zeichenfolge ist kein Aktenzeichen, wurde aber als ein solches vermeintlich erkannt).
– Richtig negative: Der Textabschnitt gehört nicht zu einer bestimmten Kategorie und wurde auch nicht als dieser Kategorie zugehörig klassifiziert (eine Zeichenfolge ist kein Aktenzeichen und wurde auch nicht als solches erkannt.)

Die Qualität eines Modells zur Klassifikation wird üblicherweise an zwei Kriterien gemessen: Genauigkeit *(precision)* und Trefferquote *(recall),* die jeweils zwischen 0 und 1 liegen können. Dabei gibt die Genauigkeit den Anteil relevanter Inhalte an allen gefundenen Inhalten an (also wie viele der als Aktenzeichen identifizierten Worte tatsächlich Aktenzeichen sind) oder die Anzahl richtig positiver Ergebnisse geteilt durch die Summe richtig positiver und falsch positiver Ergebnisse. Die Trefferquote wiederum gibt an, wie viele aller relevanten Inhalte (also wie viele Aktenzeichen aller vorkommenden Aktenzeichen) gefunden wurden oder die Anzahl richtig positiver Ergebnisse geteilt durch die Summe richtig positiver Ergebnisse und falsch negativer Ergebnisse.

Dabei beeinflussen sich Genauigkeit und Trefferquote gegenseitig. Man kann beispielsweise eine sehr hohe Trefferquote erreichen, indem einfach alle Worte als einer bestimmten Kategorie zugehörig klassifiziert werden. Allerdings sinkt dadurch die Genauigkeit dramatisch. Andersrum könnte man nur sehr wenige Worte, bei denen die Indikatoren sehr stark sind, als einer bestimmten Kategorie zugehörig klassifizieren. Da damit die Anzahl falsch positiver Ergebnisse sehr gering ist, erreicht man eine hohe Genauigkeit, findet jedoch nur sehr wenige aller richtig positiven Ergebnisse, was zu einer geringen Trefferquote führt. Um dem zu begegnen gibt man die Qualität von Modellen üblicherweise als mathematische Kombinationen von Genauigkeit und Trefferquote an: die sog. F-Masse. Dabei können die einzelnen Kriterien unterschiedlich gewichtet werden. Bei Gewichtung mittels des harmonischen Mittels[19] ergibt sich das sog. F1-Mass, was als allgemeines Qualitätskriterium gilt.

Lediglich bei den Verweisen auf andere Urteile und Literaturfundstellen (F1-Mass ca. 0,6 je nach Datensatz) und beim inhaltlichen Erfolg der Revision (F1-Mass ca. 0,45) konnten keine überwiegend zuverlässigen Resultate erreicht werden. Während dies für die Studierenden durchaus überraschend war, verwundern die Ergebnisse angesichts der hohen Schwierigkeit der dahinterliegenden Aufgaben nicht. Während sich der Wert für Verweise durch präzisere Formulierung der regulären Ausdrücke und umfangreicheres Training deutlich erhöhen lassen müsste, liegt der Fall beim Revisionserfolg anders:

Die angewandten Methoden zur Extraktion und Klassifikation von bestimmten Entitäten innerhalb der Texte beruhen vor allem auf einer Analyse des unmittelbaren Wortumfelds und der Struktur der Entitäten. Diese ist insbesondere bei den Indikatoren für Erfolg bzw. Misserfolg der Revision deutlich komplizierter als bei einzelnen Metadaten. Wollte man diese Aufgabe für die Zukunft angehen, müssten sehr viel genauere Annotationen in mehreren Arbeitsschritten mit jeweiliger Verfeinerung vorgenommen und andere technische Methoden angewendet werden. Während dies *prima facie* technisch machbar erscheint, stand den Studierenden dafür schlicht zu wenig Zeit zur Verfügung.

Kommerzielle Anbieter juristischer Informationsdienste erreichen nach eigenen Angaben F1-Masse von über 0,95, indem sie regelbasierte Bearbeitung und zeitintensive menschliche Kontrolle kombinieren. Während also die Ergebnisse des Projekts in einigen Kategorien qualitativ eindeutig darunter liegen, ist das eigentlich Bemerkenswerte, dass sie in einer solch kurzen Zeit von Beteiligten mit wenig Erfahrung und auf freiwilliger Basis erreicht wurden.

### *6. Vom Modell zur Anwendung*

Nach dem Training der Modelle konnten damit Urteile über die *AlchemyLanguage*-Schnittschnelle ausgewertet werden. Die Ergebnisse dieser Auswertung wurden von den Informatikern in die in der Zwischenzeit programmierten Anwendungen eingebettet. Im Anschluss wurden die Anwendungen von den Juristen getestet und nach deren Nutzerfeedback gemeinsam mit den Informatikern verbessert.

Die beiden Programme LawNet[20] und LawStats[21] wurden im Rahmen einer Präsentation mit Livedemonstration vorgestellt und in einem Projektbericht dokumentiert. Sie erreichen technisch das Stadium eines Prototyps bzw. eines Machbarkeitsnachweises und zeigen, wie schnell Informatiker und Juristen selbst im Ausbildungsstadium gemeinsam nutzbare und sinnvolle Anwendungen entwickeln können. Neben umfangreichen statistischen Auswertungen ermöglichen die Anwendungen eine Darstellung der Entscheidungen inklusive Verweisen und Normen als gerichteter Graph. Beide Anwendungen können die Grundlage für intensivere Forschungen mit quantitativen Methoden bilden. Gleichzeitig können neue Entscheidungen hochgeladen und ausgewertet werden. Schliesslich liesse sich mit einigen, kleinen Veränderungen ein Werkzeug zur juristischen Recherche schaffen, das sich am – technisch natürlich deutlich weiterentwickelten – Ravel Law orientiert. Damit könnten neue Leitentscheidungen identifiziert und die Entwicklung bestimmter Entscheidungsmuster in der Rechtsprechung nachvollzogen werden.

**19** Man multipliziert den doppelten Genauigkeitswert mit dem Trefferquotenwert und teilt das Ergebnis durch die Summe von Genauigkeits- und Trefferquotenwert, zum harmonischen Mittel s. https://de.wikipedia.org/wiki/Harmonisches_Mittel (besucht am 2. 6. 2017).

**20** Zur technischen Dokumentation s. <https://github.com/BIGABIG/WEBAPP>.

**21** Zur technischen Dokumentation s. <https://github.com/5menzel/Praktikum>.

## IV. Wie es weitergeht

Das Pilotprojekt wurde von den Studierenden beider Einrichtungen überwiegend positiv evaluiert. Diese Bewertung deckt sich mit den Erfahrungen der Projektverantwortlichen, sodass die Kooperation mindestens jährlich fortgesetzt werden soll. Der Erfolg des Kurses zeigt, dass solche alternativen Formate durchführbar und für Studierende interessant sind. Die entwickelten Anwendungen könnten selbstverständlich noch an vielen Stellen verbessert, erweitert und weiterentwickelt werden, was einzelne Studierende oder die Fachcommunity möglicherweise übernimmt.

Aus akademischer Sicht viel wichtiger ist jedoch, dass der Kurs beweist, dass Informatiker und Juristen schnell auf Augenhöhe produktiv arbeiten können. Studierende beider Disziplinen haben ihr Verständnis und ihre Sprachfähigkeit gegenüber den jeweils anderen erheblich verbessert und hatten dabei auch noch Freude an den Projekten. Auf diesem gegenseitigen Verständnis kann die weitere Entwicklung von Legal Technology ausserhalb von Zeitschriftenartikeln und Konferenzvorträgen im wirklichen Leben aufbauen. Das ist den nicht unerheblichen Aufwand solcher Formate zweifellos Wert.