www.jusletter-it.eu

Janice Deborah Kowalski

Die Vorhersage von Gerichtsurteilen des EGMR durch Machine Learning – eine Fallstudie

Machine Learning Methoden werden heutzutage in immer mehr Bereichen eingesetzt. Auch in der Jurisprudenz gewinnen diese Methoden an Bedeutung. Die vorliegende Arbeit befasst sich mit der Vorhersage von Gerichtsurteilen des Europäischen Gerichtshofs für Menschenrechte anhand der Anwendung von Machine Learning, dabei wird die Methode des Natural Language Processing in Kombination mit Klassifikationsmethoden auf ihre Prädiktionsfähigkeit untersucht. Dieser Beitrag zeigt schlussendlich, wie Machine Learning im Schweizerischen Rechtssystem eingesetzt werden könnte, um Gerichtsentscheide vorauszusagen.

Beitragsart: Next Generation
Region: Schweiz; EU
Rechtsgebiete: Artificial Intelligence & Recht; Rechtsinformatik; LegalTech

Zitiervorschlag: Janice Deborah Kowalski, Die Vorhersage von Gerichtsurteilen des EGMR durch Machine Learning – eine Fallstudie, in: Jusletter IT 30. September 2021

Inhaltsübersicht

1. Einleitung

[1] Die Zunahme der Digitalisierung in der Jurisprudenz widerspiegelt sich sinnbildlich an der steigenden Anzahl von digital erhältlichen rechtlichen Dokumenten.[1] Dies gilt Batsakis et al. zur Folge insbesondere dort, wo es um juristische Argumentationen geht: «*[…] legal reasoning was one of the first application domains of artificial intelligence since its emergence, and more so after the proliferation of expert systems during the 1980s*».[2] Mit dieser Entwicklung, Gerichtsentscheide öffentlich zugänglich zu machen, wurde die Grundlage für «*Big Data*» Analysen bzw. Machine Learning Modelle geschaffen.[3] Machine Learning gilt als Unterkategorie von LegalTech, die «*Verbindung von Recht und Technologie*»[4], wobei Machine Learning wiederum der Technologie 2.0 angehört.[5] Bei Letzterem werden technische Modelle oder Lösungen als Ersatz menschlicher Arbeit verwendet. So werden heutzutage Machine Learning Tools als Hilfestellung für oben genannte Problematik in der Jurisprudenz eingesetzt, wie bei der Vorhersage der Rechtsprechung. Dementsprechend analysiert die vorliegende Arbeit die Studie von Medvedeva, Vols und Wieling zur Vorhersage von Gerichtsurteilen des Europäischen Gerichtshofs für Menschenrechte (EGMR) anhand von Machine Learning. Es soll aufgezeigt werden, inwiefern sich das *Natural Language Processing* in Kombination mit *Klassifikationsmethoden* in den Rechtswissenschaften als zukunftsfähig erweist.

2. Anwendungshintergrund

[2] Während sich in den Vereinigten Staaten solche quantitativen Analysen schon länger etabliert haben, sind anderenorts vorwiegend sogenannte *doctrinal research methods* anzutreffen. Obwohl verhältnismässig wenig rechtliche Studien auf der Anwendung von Machine Learning basieren, nimmt diese Entwicklung in den letzten Jahren international allmählich zu. So veröffentlichten Aletras et al. im Jahr 2016 eine wissenschaftliche Arbeit zur gleichen Thematik wie die der vorliegenden Arbeit.[6] Das oben erwähnte Paper von Medvedeva, Vols und Wieling baut auf Letzterem auf, verwendet aber eine grössere Datenmenge bzgl. der Anzahl analysierten Gesetzesartikeln und den dazugehörigen Gerichtsentscheiden der Europäischen Menschenrechtskonvention

1 Glaser/Scepenkova/Matthes, S. 61.

2 Batsakis et al., S. 31.

3 Medvedeva/Vols/Wieling, S. 237.

4 Kummer/pfäffli, S. 134.

5 Kummer/pfäffli, S. 135.

6 Medvedeva/Vols/Wieling, S. 239 ff., vgl. Aletras et al.

(EMRK).[7] Die vorliegende Studie basiert nur auf bereits entschiedenen und nicht künftigen Fällen des EGMR.

3. Experimente

[3] In der Studie von Medvedeva, Vols und Wieling wurden drei verschiedene Experimente mit der Anwendung von *Natural Language Processing* und der *Klassifikation* durchgeführt, welche nachfolgend analysiert werden. Bei allen Experimenten wurde das Modell jeweils trainiert und getestet, beim dritten Experiment geschah dies jedoch mit unvollständigen Informationen.[8]

3.1. Erstes Experiment

[4] Im ersten durchgeführten Experiment der Studie wurden Wörter und Sätze der verwendeten Fälle als Input genutzt, um die Gerichtsentscheide vorherzusagen. Als Grundlage dafür dienten alle englischen und auf der Webseite HUDOC des EGMR verfügbaren Entscheide bis zum Stand vom 11. September 2017. Das Ziel dessen stellte das Einteilen des jeweiligen Falls in zwei Kategorien dar, das heisst, es sollte jeder Fall entweder als *Verstoss* oder als *kein Verstoss* des vorliegenden Gesetzesartikels der EMRK eingestuft werden. Aufgrund dessen sollten die Entscheide der Richter dazu vorhergesagt werden.[9] Als Methode wurde vom Machine Learning Bereich *Natural Language Processing* Gebrauch gemacht. Die *Klassifikationsmethode* kann grundsätzlich einen Text per se nicht einordnen, daher wurden die Texte in für die Maschine lesbare *n-grams* umgewandelt. Diese *uni-*, *bi- oder trigrams* wurden je nach Bedeutung für den Fall gewichtet und als Zahlen dargestellt. Die hier verwendete *Methode der Klassifikation* wird als *supervised* Machine Learning bezeichnet.[10]

[5] Die *Klassifikation* des Outputs in separate Kategorien ist als ein Teilbereich des Machine Learnings einzustufen.[11] In der vorliegenden Studie wurde dafür der Ansatz der *Support Vector Machine (SVM)* eingesetzt. Es handelt sich dabei um ein binäres Modell. Für jeden Gesetzesartikel wurde das Modell einzeln trainiert, damit dieses die zugehörigen Gerichtsfälle als *Verstoss* oder als *kein Verstoss* der Gesetzesartikel der EMRK klassifiziert.[12] Um das Modell für beide Klassen gleichmässig zu trainieren, wurden die Datensätze ausgewogen. So wurden von beiden Kategorien eine identische Anzahl Fälle verwendet.[13] Allerdings liessen Medvedeva, Vols und Wieling die rechtliche Argumentation im Modell explizit aussen vor, sodass die Vorhersagen nur auf Fakten und nicht auf für den Verfahrensausgang richtungsweisenden Argumentationen und Gesetzesartikeln basierten. Die Aufteilung der Daten wurde dann zufällig in einen *training*

7 Vgl. Medvedeva/Vols/Wieling, S. 242: Analyse von neun Gesetzesartikeln der EMRK; Aletras et al., S. 6: Analyse von drei Gesetzesartikeln der EMRK.

8 Vgl. Medvedeva/Vols/Wieling, S. 248, 259.

9 Medvedeva/Vols/Wieling, S. 242 f., 246.

10 Vgl. Schrider/Kern, S. 301 ff.: *Supervised* Machine Learning beruht auf bereits vorhandenem Wissen eines Musterdatensatzes, was für Vorhersagen für neue Daten verwendet wird im Unterschied zu *unsupervised* Machine Learning.

11 Hastie/Tibshirani/Friedman, S. 101 ff.

12 Medvedeva/Vols/Wieling, S. 243 f.

13 Medvedeva/Vols/Wieling, S. 247 f.

und *testing* Teil gegliedert. In der Phase des *trainings* liess man das Modell den Zusammenhang zwischen den Texten und den Gerichtsentscheiden lernen. Hingegen in der Phase des *testings* schätzte das Modell eigenständig die Entscheide anhand von den Texten alleine. Die Evaluierung wurde schliesslich anhand vom Anteil der korrekt vorhergesagten Entscheide vorgenommen. Hierfür wurden die Vorhersagen des Modells mit den tatsächlichen Entscheiden aus dem *testing* Teil abgeglichen.[14] Um die Stabilität der Ergebnisse zu prüfen, wurde die Aufteilung in den *training* und *testing* Teil, als auch die Schätzung des Modells, mehrmals wiederholt. Dies nennt sich *cross-validation*.[15]

[6] Es bleibt bzgl. den Resultaten vom Experiment eins festzuhalten, dass durchschnittlich Vorhersagen von Gerichtsentscheiden des EGMR mit einer Präzision von ungefähr 75 % bei beiden Evaluationskriterien erreicht wurden. Würde man die Resultate zufällig schätzen, läge die Wahrscheinlichkeit bei 50 %. Dementsprechend kann auf die Prädiktionsfähigkeit der Methode geschlossen werden, obwohl der Input lediglich auf Textbausteinen der jeweiligen Gerichtsfälle beruht. Dies wurde auch mittels zweier weiterer statistischer Kriterien bestätigt.[16] Jedoch können Unterschiede in der Genauigkeit der Vorhersage je nach Gesetzesartikel festgestellt werden. Dabei gilt generell, je mehr Daten zur Verfügung stehen, desto grösser ist die Präzision bzgl. ähnlicher Fälle zu einem Gesetzesartikel.

3.2. Zweites Experiment

[7] Beim zweiten Experiment wurde im Vergleich zu Experiment eins nichts abgeändert, ausser der Wahl der zeitlich chronologischen Reihenfolge der Daten für das *training* und *testing* des Modells, um effektiv zukünftige Gerichtsfälle vorherzusagen.[17] Bei Experiment eins wurde eine solche Abfolge nicht eingehalten. Experiment zwei hat den Vorteil, dass diese Vorgehensweise logischer und realitätsnäher ist. Gleichwohl mussten die Gerichtsfälle neu balanciert werden, um die zeitliche Chronologie einhalten zu können. Die Konsequenz davon waren Datenverluste, wobei Medvedeva, Vols und Wieling sich aufgrund dessen nur noch auf drei der ursprünglich neun Gesetzesartikel der EMRK konzentrierten (Artikel 3, 6 und 8 der EMRK). Bei diesen verbleibenden Gesetzesartikeln blieben genügend Daten übrig, damit die Aussagekraft der Vorhersage nicht verloren geht.[18] Schliesslich konnte hierbei festgestellt werden, dass das Vorhersagen von künftigen Gerichtsentscheiden des EGMR basierend auf vergangenen Fällen schwieriger war und die Präzision dessen sank. Grundsätzlich kann gefolgert werden, je grösser der zeitliche Abstand zwischen bereits ergangenen und zukünftigen Gerichtsurteilen, desto unpräziser ist die Vorhersage.

14 Medvedeva/Vols/Wieling, S. 243, 248 f.

15 Medvedeva/Vols/Wieling, S. 244.

16 Vgl. Medvedeva/Vols/Wieling, S. 253 f.

17 Medvedeva/Vols/Wieling, S. 257 f.

18 Medvedeva/Vols/Wieling, S. 257 f.

3.3. Drittes Experiment

[8] Experiment drei weicht von den ersten beiden Experimenten ab, so wurden die Nachnamen der jeweils beteiligten Richter als einziger Input verwendet. Die eigentlichen Textelemente der Urteile und somit auch deren Gewichtung waren kein Bestandteil dieser Vorhersage. Als ausschlaggebend galt lediglich, ob ein Richter einer Kammer in einem bestimmten Entscheid urteilte oder nicht.[19]

[9] Grundsätzlich wiesen die Resultate dieser Schätzung eine niedrigere Genauigkeit auf (durchschnittlich 65 %), wie wenn die Textteile von Entscheiden als Input genutzt wurden (durchschnittlich 75 %). Dies erscheint nachvollziehbar, zumal der Informationsgehalt bei einem Nachnamen kleiner war, wie bei einem Text. Allerdings zeigte die Präzision je nach Gesetzesartikel der EMRK eine grosse Heterogenität auf. So schätzte das Modell bei gewissen Artikeln ebenso genau wie bei einem Text. Bei anderen Gesetzesartikeln wurde nur leicht besser vorhergesagt, wie wenn zufällig geraten würde.[20]

[10] Die Ergebnisse in Bezug auf die Bedeutung der massgebenden Wörter lassen sich in Experiment drei eindeutiger interpretieren als in Experiment eins. Es lässt sich also feststellen, welche Richter häufiger in einer Kammer dabei waren, welche für einen *Verstoss* oder für *keinen Verstoss* eines Gesetzesartikel der EMRK gestimmt hatten.[21] Allerdings kann von keinem kausalen Zusammenhang ausgegangen werden. So besteht keine Annahme, dass Richter des EGMR immer pauschal für oder gegen einen Verstoss der EMRK urteilen würden.

4. Diskussion

[11] In den vorhergehenden Kapiteln konnte gezeigt werden, dass die vorliegende Studie von Medvedeva, Vols und Wieling die Vorstudie von Aletras et al. weiterentwickelte. Medvedeva, Vols und Wieling basierten ihre Ergebnisse zum einen auf einer grösseren Anzahl Gesetzesartikel sowie auf mehr Gerichtsurteilen pro Gesetzesartikel. Zum anderen liessen sie rechtliche Argumentationen bewusst aussen vor, um dem Modell nicht im Vorfeld Hinweise auf den endgültigen Entscheid zu liefern. Letzteres wurde von der Vorstudie nicht beachtet, was unter Umständen deren Ergebnisse verfälscht hatte. Der Chronologie der Reihenfolge in Bezug auf das *training* und *testing* hatte die Vorstudie keine Beachtung geschenkt.[22] Dies erklärt auch, weshalb die Genauigkeit der Vorhersage bei Aletras et al. bei ungefähr 80 % und diese bei der vorliegenden Studie lediglich bei 75 % lag.[23]

[12] Grundsätzlich können einige Vorteile der Studie von Medvedeva, Vols und Wieling für Vorhersagen im Allgemeinen genannt werden. So handelt es sich bei der hierfür verwendeten Methode der *Support Vector Machine* um keine sogenannte Black Box Methode, das heisst, der Algorithmus kann intuitiv interpretiert werden. Damit kann einerseits erklärt werden, welche Begriffe für die Vorhersagen von Bedeutung waren. Andererseits deutet dies auf keinen kausalen

19 Medvedeva/Vols/Wieling, S. 259 f.

20 Medvedeva/Vols/Wieling, S. 259 f.

21 Medvedeva/Vols/Wieling, S. 261.

22 Medvedeva/Vols/Wieling, S. 259.

23 Vgl. Aletras et al., S. 2; Medvedeva/Vols/Wieling, S. 237.

Korrelationszusammenhang hin. Ferner ist die Komplexität des *Natural Language Processing* verhältnismässig, sodass sie auf einfache Art und Weise repliziert und auf andere Fälle angewandt werden kann. Jedoch fehlt es teils an inhaltlicher Bedeutung der *n-grams*, welche die Textbausteine der Urteile repräsentierten.[24] Diese Problematik könnte z.B. mit semantischen Analysen verhindert werden.[25]

[13] Gleichermassen sind gewisse Nachteile der Methode von Medvedeva, Vols und Wieling zu erwähnen. Aufgrund des Balancierens der beiden Kategorien *Verstoss* oder *kein Verstoss* der Gesetzesartikel der EMRK wurden Daten ausgelassen, was zu einer schlechteren Performance der Methode führte. Es lässt sich folgern, dass die Präzision der Methode trotz dem hohen Innovationsgehalt noch ausbaufähig ist. Mit *under-* oder *oversampling* hätte dem entgegengewirkt werden können, da mit diesen Methoden keine Datenmengen aufgegeben werden müssen.[26] Als weiterer Kritikpunkt gilt, dass technische Aspekte im Vergleich zu den rechtlichen in der Studie mehr Gewicht erhielten. So wurde nicht direkt beleuchtet, was solche Vorhersagen für die Jurisprudenz bedeuten würden, ob sie überhaupt rechtlich legitimiert sind und wo der EGMR oder natürliche Personen konkret von der Methode Gebrauch machen könnten. Dies könnte an der noch verbesserungsfähigen Genauigkeit der Vorhersagen oder an der Simplizität der Methode liegen. Mit anderen komplexeren Machine Learning Methoden, wie z.B. das Deep Learning, könnte dem entgegengewirkt werden.[27]

[14] Es wurde von den Autoren jedoch nie beabsichtigt die Richter des EGMR zu ersetzen, sondern nur die Erfolgschancen eines Gerichtsfalls einschätzen zu können. Dies dürfte vor allem natürlichen Personen dienen. Dementsprechend wurden die Erkenntnisse der Studie mit der Webseite «Juri *says*» einer breiten Öffentlichkeit zugänglich gemacht, wo richtige und falsche Vorhersagen des Modells einsehbar sind.[28]

[15] Gegebenenfalls könnte die Methode alternativ bei der Prüfung der Zuständigkeit eines Gerichts eingesetzt werden.[29] So könnte das Modell konkret die Arbeit in den Gerichten erleichtern, da die Entscheidung der Zuständigkeitsprüfung auch nicht so weitreichend wäre, wie ein Urteil eines komplexen Falles. Es ergibt sich aus der binären Natur des Modells, dass es eher für Ja- oder Nein-Entscheide dient, zumal damit keine weitgehende Fallprüfung durchgeführt werden könnte. Allenfalls wäre es möglich das Modell im schweizerischen Recht anzuwenden. So könnte die Methode unter Umständen durch gewisse Abänderungen oder Erweiterungen als Hilfsmittel bei der Verletzung von Grundrechten der Schweizerischen Bundesverfassung dienen. Dazu könnten die vom Schweizerischen Bundesgerichts auf dessen Webseite seit 1954 veröffentlichten Fälle verwendet werden.[30] Da der von Medvedeva, Vols und Wieling geschriebene Code offengelegt wurde, könnte dieser übernommen und für die schweizerischen Gerichtsurteile repliziert werden.

24 Vgl. Medvedeva/Vols/Wieling, S. 254, Fig. 3.

25 Vgl. Mikolov et al., S. 3111 ff.

26 Vgl. Medvedeva/Vols/Wieling, S. 262 f.

27 Vgl. LeCun/Bengio/Hinton, S. 436 ff.

28 Vgl. Juri Says.

29 Vgl. Medvedeva/Vols/Wieling, S. 245.

30 Vgl. BGE Rechtsprechung.

5. Fazit

[16] In der vorliegenden Arbeit wurde anhand der Studie von Medvedeva, Vols und Wieling untersucht, inwiefern Vorhersagen von Machine Learning Methoden sinnvoll erscheinen, um Gerichtsentscheide des EGMR vorherzusagen. Dazu dienten drei verschiedene Experimente, welche mittels *Natural Language Processing* und der *Klassifikationsmethode* basierend auf vergangenen Fällen des EGMR *Verstösse* oder *keine Verstösse* von neun verschiedenen Gesetzesartikel der EMRK vorhersagten. Es konnte gezeigt werden, dass durch die Umwandlung relevanter Textelemente von Gerichtsurteilen in *n-grams* deren Verfahrensausgänge mit einer Genauigkeit von 75 % vorhergesagt werden konnten. Die korrekte Schätzung von künftigen Urteilen aufgrund von vergangenen Gerichtsentscheiden wies jedoch einige Schwierigkeiten auf, während alleine basierend auf den Nachnamen der an den Urteilen beteiligten Richtern je nach Gesetzesartikel präzise Vorhersagen erfolgten. Schliesslich wurden die Vor- und Nachteile des *Natural Language Processing* in Kombination mit *Klassifkation* gegeneinander abgewogen und die Zukunftschancen einer solchen Methode für schweizerische Gerichtsentscheide aufgezeigt.

Janice Deborah Kowalski, B.A. ist Studentin der Rechtswissenschaften (MLaw) an der Universität St. Gallen (HSG).