

DER EINSATZ VON TECHNOLOGY ASSISTED REVIEW (TAR) IN DER PRAXIS

Gernot Fritz

Principal Associate | Rechtsanwalt, Freshfields Bruckhaus Deringer LLP
Seilergasse 16, 1010 Wien, AT
gernot.fritz@freshfields.com

Schlagworte: *Legal Tech, Artificial Intelligence, Technology Assisted Review, TAR, Predictive Coding, Simple Active Learning, Simple Passive Learning, Continuous Active Learning, Zivilverfahren, Document Production, Untersuchungen, Investigations*

Abstract: *Sollen im Zuge von (internen) Untersuchungen oder im Rahmen von Zivilprozessen die für eine zu untersuchende Frage relevanten Dokumente herausgefiltert werden, können durch den überlegten Einsatz von TAR – also maschinenlernende Systeme, die über bloße Stichwortsuchen weit hinausgehen – sowohl Aufwand als auch Kosten drastisch reduziert und gleichzeitig die Trefferquote erhöht werden. Der Beitrag bespricht die für Juristen relevanten Konzepte hinter TAR und die Tücken des Einsatzes von TAR in der juristischen Praxis.*

1. Einleitung¹

Dieser Beitrag beschreibt die Fallstricke des Einsatzes von Technology Assisted Review (TAR) anhand eines konkreten Mandats («Project X»), welches aufgrund seiner Anforderungen bestens für den Einsatz von TAR geeignet war, aber auch gerade wegen seiner Anforderungen an den TAR-Prozess sowohl die beteiligten Juristen als auch die beteiligten Data Scientists stark forderte.

2. Generell: Einsatzgebiete von TAR

TAR kommt regelmäßig in zwei Konstellationen zur Anwendung:

- Bei unternehmensinternen Untersuchungen, bei denen ein bestimmter Sachverhalt / eine bestimmte Verdachtslage aufgeklärt werden soll. Dazu werden die auf der IT-Infrastruktur des Unternehmens gespeicherten Dokumente (im technischen Sinn) durchsucht – E-Mails, Kalendereinträge, Word-Dokumente etc.
- Bei der sogenannten «pre-trial discovery». Damit ist eine vorprozessuale Phase gemeint, welche vor allem im anglo-amerikanischen Zivilprozess verbreitet ist. Im Rahmen der «pre-trial discovery» kann eine Partei von der Gegenpartei die Übermittlung von Dokumenten verlangen, die sich auf einen bestimmten Sachverhaltskomplex beziehen (dies wird auch als «document production» bezeichnet). Die Gegenpartei ist gut beraten, dem Verlangen bestmöglich und umfassend zu entsprechen, da eine Nichtentsprechung, eine Schlechtentsprechung oder das (unbeabsichtigte oder absichtliche) Zurückhalten von Dokumenten gravierende negative Konsequenzen (bis zum Prozessverlust) haben kann.²

Beim Einsatz von TAR zur Auswahl der relevanten Dokumente im Rahmen einer «document production» sind neben dem TAR einsetzenden Unternehmen auch noch die die Übermittlung verlangende Gegenpartei und das Gericht beteiligt. In vielen Jurisdiktionen muss daher vor dem Einsatz von TAR zur Beantwortung der «document production»-Anfrage der Gegenpartei mit dem Gericht noch ein sogenanntes «TAR-Protocol»

¹ Dieser Beitrag gibt ausschließlich die persönliche Meinung des Autors wieder.

² KLAUSEGGER/FRITZ, U.S. Discovery: The Austrian Perspective, DAJV Newsletter 1/2012, S. 12 (S. 13).

abgeschlossen werden, welches unter anderem regelt, (i) in welcher Dokumentenmenge («search universe») gesucht wird, (ii) wie das eingesetzte TAR System beschaffen ist, (iii) wie der Trainingsprozess aufgesetzt wird und (iv) wie validiert wird, dass das System seine Aufgabe³ erledigt hat.

Da bei unternehmensinternen Untersuchungen nur das konkrete Unternehmen beteiligt ist, ist beim Einsatz von TAR zwar formal nichts weiter zu beachten. Soll aber die Verwertbarkeit der Ergebnisse in allfällig nachfolgenden Behördenverfahren sichergestellt sein, so ist zu empfehlen, ähnliche Qualitätsanforderungen an den TAR-Prozess, die auch in einem TAR-Protocol vereinbart werden müssten, rechtssicher und nachweisbar zu dokumentieren.

3. Konkret: Das «Set-Up» von Project X

Beim Project X handelte es sich um eine unternehmensinterne Untersuchung. In seiner Größenordnung sicherlich einzigartig, bestand das «search universe» aus über 150 Millionen Dokumenten, die auf der Unternehmensinfrastruktur zu den relevanten «Custodians»⁴ vorhanden waren. Das «search universe» bestand zum Großteil aus E-Mails, Word-Dokumenten, PowerPoint-Präsentationen, XLS-Listen und Kalendereinträgen.

Die Menge an Dokumenten war viel zu groß, um dieses Projekt allein mit manueller Durchsichtung und menschlichen «Reviewern»⁵ abarbeiten zu können.⁶ Im klassischen Prozess (ohne TAR) wäre naheliegend gewesen, mittels bloßer Stichwortsuche den Suchradius einzuschränken – also die Menge an zu untersuchenden Dokumenten für die menschlichen Reviewer künstlich einzuschränken. Eine reine Stichwortsuche birgt allerdings immer das Problem, dass sich die tatsächlich relevanten Stichworte zumeist erst im Laufe der Untersuchung herauskristallisieren und die Auswahl der Stichworte, mit denen die vermeintlich relevanten Dokumente gefiltert werden sollen, zu Beginn einer Untersuchung nicht treffsicher genug sein kann. Daneben haben Stichwortsuchen auch einen großen Mangel: Mit ihnen können Tippfehler (die in der Praxis doch häufiger als erwartet vorkommen) nicht oder nur mit sehr großem Aufwand abgedeckt werden.

Mit diesen Parametern war der Einsatz von Legal Tech (in concreto: TAR) vorgezeichnet.

4. Was ist TAR? Das Finden der Nadel im Heuhaufen

Verglichen mit der Gesamtanzahl der Dokumente sind meistens nur wenige Dokumente für eine bestimmte Fragestellung tatsächlich relevant. Der Einsatz gleicht dann einer Suche nach der Nadel im Heuhaufen. Aus einem rezenten Hollywood-Film stammt das Zitat «*If you want to find a needle in a haystack, bring a magnet*». Im geplanten und überlegten Einsatz kann TAR als derartiger Magnet dienen. Beim Project X war aufgrund des Untersuchungsgegenstandes sehr wahrscheinlich, dass die Anzahl der für eine bestimmte Fragestellung relevanten Dokumente (diese werden auch als «responsive documents» bezeichnet) im Verhältnis zur gesamten Dokumentenanzahl extrem gering sein würde – die zu suchende Nadel also sehr klein sein wird.⁷

TAR beschreibt nun Systeme, die eine große Menge an Dokumenten auf bestimmte Fragestellungen untersuchen können und weit über bloße Stichwortsuchen hinausgehen. Ein gut trainiertes TAR-System kann relevante Dokumente sogar dann identifizieren, wenn diese kein einziges der bekannten Schlüsselwörter aufweist.

Technisch handelt es sich bei TAR-Systemen um «supervised machine learning». Wie jedes «machine learning» System muss dieses zunächst trainiert werden: Der technische Aspekt des Trainings wird in der Praxis zumeist entweder mittels Support Vector Machines oder logistischer Regression realisiert.⁸ Wie der Begriff «supervi-

³ Und zwar, einen bestimmten Prozentsatz der tatsächlich relevanten Dokumente im festgelegten «search universe» zu finden.

⁴ Der Begriff «Custodian» bezeichnet bei Untersuchungen eine relevante Person, die im Fokus der Untersuchung steht und deren Daten untersucht werden sollen.

⁵ Der Begriff «Reviewer» bezeichnet die Personen, die die Daten im Rahmen einer Untersuchung sichten und bewerten.

⁶ Unter Annahme der durchschnittlichen Review-Kapazität eines menschlichen Reviewers wären mehr als 10'000 Reviewer notwendig gewesen, um den Review innerhalb eines Jahres (!) abgeschlossen zu haben.

⁷ Dies hat sich bei Project X im fortschreitenden Prozessverlauf auch bestätigt.

⁸ GROSSMAN/CORMACK, Continuous Active Learning for TAR, Practical Law Journal, April/May 2016, S. 32 (S. 33 f.).

sed» machine learning nahelegt, erfolgt das Training unter menschlicher Beteiligung. Für die Lernphase bei TAR existieren verschiedene Protokolle:⁹

- Simple Passive Learning (SPL): Die Dokumente für das Training werden von Menschen ausgewählt; die Maschine lernt alleine an diesen Beispielen und kennzeichnet die zur Verfügung gestellten Dokumente als relevant oder nicht relevant. Diese werden dann von menschlichen Reviewern gegengeprüft. Sind die Resultate nach Ansicht der Reviewer unbefriedigend, werden dem TAR-System weitere Dokumente zum Lernen zur Verfügung gestellt. Erst wenn die Resultate als ausreichend angesehen werden, wird das Training beendet.
- Simple Active Learning (SAL): Auch hier werden die Dokumente für das initiale Training von Menschen ausgewählt. Nach dem ersten Trainingsschritt wählt allerdings das TAR-System aus dem gesamten «search universe» die von Menschen zu gegenprüfenden Dokumente aus – und zwar eine vorher festgelegte Anzahl der nach Ansicht des TAR-Systems relevantesten Dokumente, welche von Menschen auf ihre tatsächliche Relevanz überprüft werden sollen. Hierbei handelt es sich stets um «neue» Dokumente, also solche, die von menschlichen Reviewern bislang noch nicht durchgesehen wurden. Die Ergebnisse dieser Gegenprüfung werden an das TAR-System zurückgegeben und das TAR-System verarbeitet diesen Input für das weitere Training. Das System lernt also aus dem Feedback, welche der für das TAR-System unbekanntesten Dokumente es richtig und welche falsch klassifiziert hat. Eine derartige Abfolge (Verbesserung des Algorithmus – Versehen der Dokumente mit einem Relevanzwert – Gegenprüfung der nach Ansicht des TAR-Systems relevantesten Dokumente durch menschliche Reviewer) wird als «Iteration» bezeichnet.
- Continuous Active Learning (CAL): Nach menschlicher Auswahl initialer relevanter Trainingsdokumente oder Anfertigung eines hypothetischen relevanten Dokuments (sogenanntes «synthetic document») schlägt das TAR-System immer das nächste am wahrscheinlichsten relevante Dokument zur Begutachtung durch einen menschlichen Reviewer vor. Nach unmittelbarem Feedback des menschlichen Reviewers findet ein Update des Trainings statt und das System schlägt unter sofortiger Berücksichtigung des Feedbacks des menschlichen Reviewers zum unmittelbar vorigen Dokument das nächste am wahrscheinlichsten relevante Dokument vor.

Um die Relevanz eines Dokuments messbar machen zu können, weist das TAR-System jedem Dokument einen Relevanzwert zu. Beim beschriebenen Project X erfolgte dies etwa auf einer Skala von 0,00 bis 1,00, wobei ein Wert von 1,00 bedeutet, dass das TAR-System das Dokument in jedem Falle als relevant ansieht. Ein Wert von 0,00 hingegen kennzeichnet nach Ansicht des TAR-Systems in jedem Falle ein nicht relevantes Dokument. Relevanz bzw. «responsiveness» ist immer anhand einer konkreten Fragestellung zu verstehen¹⁰ – ein Dokument kann daher für eine bestimmte Unterfrage des Untersuchungsgegenstandes relevant sein, für eine andere aber nicht. Wie breit in Hinblick auf den TAR-Einsatz gute (Unter-)Fragen gefasst werden sollen, hängt vom Untersuchungsgegenstand, dem «search universe» und ggf. auch den Resultaten des Trainings ab. Zu beachten ist, dass TAR-Systeme nur Text verarbeiten können. In einem ersten Schritt müssen daher alle Dokumente in maschinenlesbaren Text umgewandelt werden – aus Bildern/Präsentationen wird daher nur der darin enthaltene Text ausgelesen und in ein «Schatten-Dokument» umgewandelt. Der Algorithmus arbeitet nur auf diesen Schatten-Dokumenten, der menschliche Reviewer sieht aber die Original-Dokumente.

⁹ Dazu etwa GROSSMAN/CORMACK, Technology-Assisted Review in Electronic Discovery, 2017, S. 6 f, abrufbar unter https://judicialstudies.duke.edu/wp-content/uploads/2017/07/Panel-1_TECHNOLOGY-ASSISTED-REVIEW-IN-ELECTRONIC-DISCOVERY.pdf (zuletzt abgerufen im Januar 2018); GROSSMAN/CORMACK, Continuous Active Learning (Fn. 8), S. 32 (S. 36).

¹⁰ Im Folgenden wird auf diese Klarstellung – der einfacheren Lesart geschuldet – verzichtet, muss aber, wenn von relevanten Dokumenten gesprochen wird, selbstverständlich immer «mitgelesen» werden.

5. Erkenntnis 1: Die Auswahl des Seed Set ist nicht trivial

Leider sind TAR-Systeme noch nicht so weit, dass ihnen eine Fragestellung als bloße Frage formuliert gestellt werden könnte und sie daraufhin mit der Klassifizierung von Dokumenten als relevant oder nicht relevant beginnen. Initial muss einem TAR-System zunächst gezeigt werden, wonach man sucht bzw. nicht sucht. Um mit TAR beginnen zu können, bedarf es eines sogenannten «Seed Sets».¹¹ Ein Seed Set besteht aus zwei Teilen: Einem «positiven Seed Set», welches von menschlichen Reviewern als relevant eingestufte Dokumente beinhaltet, und einem «negativen Seed Set», welches von menschlichen Reviewern als nicht relevant eingestufte Dokumente beinhaltet.

Das Finden von relevanten Dokumenten für das positive Seed Set kann je nach Fragestellung und Vorbefassung schwieriger oder einfacher sein. Sind noch keine relevanten Dokumente bekannt, kann etwa mit einer Stichwortsuche versucht werden, zumindest eine Handvoll relevanter Dokumente zu finden, die dem TAR-System als positives Seed Set mitgegeben werden können. Je mehr relevante Dokumente von Beginn an das TAR-System weitergegeben werden, umso schneller wird das TAR-System bessere Ergebnisse liefern. Werden dem TAR-System nur wenige relevante Dokumente als positives Seed Set übergeben, so ist der menschliche Review-Aufwand in weiterer Folge regelmäßig höher, da weit mehr Iterationen stattfinden müssen, um zu denselben Ergebnissen zu kommen und vor allem in den ersten Iterationen ein großer prozentueller Anteil an tatsächlich nicht relevanten Dokumenten zur Gegenprüfung durch menschliche Reviewer von dem TAR-System präsentiert werden wird. Da sich ein kleines positives Seed Set allerdings im Laufe der Iteration «normalisiert», ist empfehlenswert, nicht unverhältnismäßig viel Aufwand zu tätigen, um auf eine fixe Zahl (etwa 100) relevanter Dokumente als positives Seed Set hinzuarbeiten.

Das Finden von für eine bestimmte Fragestellung nicht relevanten Dokumenten scheint meistens kein großes Problem zu sein. – Bei einer Suche nach der Nadel im Heuhaufen findet sich definitionsgemäß eine große Menge Heu im Haufen. Zu beachten ist aber, dass auch die negativen Dokumente gut ausgewählt werden müssen, da das negative Seed Set einen guten Kontrast zu den relevanten Dokumenten des positiven Seed Sets bieten sollte. Eine gutes negatives Seed Set beinhaltet nach Möglichkeit sowohl (i) Dokumente, die für eine bestimmte (Unter)Frage zwar nicht relevant sind, aber für eine andere zu untersuchende (Unter)Frage hingegen schon, als auch (i) Dokumente, die für gar keine der Fragestellungen des Untersuchungsgegenstandes relevant sind.

Für ein gutes Seed Set ist weiters zu beachten, dass alle darin enthaltenen Dokument-Typen in beiden und nicht nur im positiven oder im negativen Seed Set vorkommen. Finden sich im negativen Seed Set etwa keine Power-Point Präsentationen, im positiven Seed Set hingegen schon, so ist sehr wahrscheinlich, dass Power-Point Präsentationen vom System als besonders relevant angesehen werden, da sie bestimmte Gemeinsamkeiten teilen, die im negativen Seed Set vermutlich nicht vorkommen.¹² Auch das würde sich in mehreren Iterationen normalisieren, aber es spart viel Zeit und Aufwand, wenn dieser Aspekt bereits bei der Erstellung des Seed Sets berücksichtigt wird.

6. Erkenntnis 2: Gutes Verständnis der Funktionsweise des TAR-Systems aufseiten der trainierenden Juristen und enge Abstimmung mit Data Scientists ist unumgänglich

TAR-Systeme sind (noch) keine «plug and play» Lösungen. Es gibt TAR noch nicht «out of the box». Ein TAR-System muss – wie ein neues menschliches Teammitglied – trainiert werden und kann sich erst durch kontinuierliches Feedback verbessern. Um rasch und effizient zu den gewünschten Resultaten zu gelangen,

¹¹ Jedenfalls in der Ausprägung SAL oder SPL und überwiegend auch in der Ausprägung CAL, sofern man hier nicht auf ein synthetisches Dokument zurückgreift.

¹² Etwa die Wendung «Vielen Dank für Ihre Aufmerksamkeit».

ist besonders wichtig, dass die Reviewer, die das System trainieren, sowohl mit dem zu untersuchenden Sachverhalt, den rechtlichen Implikationen der gefundenen Erkenntnisse, aber auch der Funktionsweise des eingesetzten TAR-Systems vertraut sind. Dies erfordert zwar keine vertieften Informatik-Kenntnisse, aber dennoch ein gewisses technisches «Grundverständnis» des TAR-Prozesses. Aus diesem Grund ist auch eine stetige und enge Abstimmung mit den auf Seiten des Anbieters des TAR-Systems beteiligten Data Scientists erforderlich, etwa zur Bewertung von Zweifeldokumenten im Rahmen der menschlichen Gegenprüfung und der Bewertung des Trainingsfortschrittes des TAR-Systems.

7. Erkenntnis 3: Die Frage, wann das System «gut genug» ist, ist schwierig zu beantworten

Wann ist ein TAR-System «gut genug» und ausreichend trainiert? Einfache Antwort: «Wenn das System relevante Dokumente von nicht-relevanten Dokumenten unterscheiden kann». Doch wie ist das messbar? Bei Einsatz des CAL-Protokolls ist die Stop-Bedingung sehr einfach. Findet das System nur mehr wenige bzw. gar keine tatsächlich relevanten Dokumente mehr, dann kann der Review beendet werden.¹³ Bei SPL und SAL erfolgt die Evaluierung des Fortschritts des TAR-Systems regelmäßig durch Validierung mittels «Control Set». Ein Control Set ist eine zufällig aus dem «search universe» entnommene Anzahl an Dokumenten, welche von menschlichen Reviewern durchgesehen und als relevant oder nicht relevant gekennzeichnet werden. Damit das Control Set für die Validierung des Lernfortschritts verwendet werden kann, muss eine ausreichende Anzahl von relevanten Dokumenten darin enthalten sein.¹⁴ Das Control Set darf auch nicht mit anderweitig als relevant bekannten Dokumenten «aufgefüllt» werden. Dabei würde eine wichtige Eigenschaft des Control Sets verloren gehen, nämlich eine statische Repräsentanz der Gesamtpopulation zu sein. Da im Control Set bekannt ist, welche Dokumente tatsächlich relevant sind, kann die Güte des Systems daran gemessen werden, wie viele der tatsächlich relevanten Dokumente das trainierte System findet («recall»)¹⁵ und wie viele tatsächlich nicht relevante Dokumente vom trainierten System als relevant markiert werden («precision»)¹⁶. Eine weitere hier wichtige Kennzahl ist der sogenannte «F₁»-Wert, das harmonische Mittel von «recall» und «precision».¹⁷

Oft ist es aber mit vertretbarem Aufwand nicht möglich, ein Control Set mit ausreichend relevanten Dokumenten zu erhalten, womit die Control Set Validierung nicht mehr sinnvoll möglich ist. In diesen Fällen kann der Lerneffekt etwa mittels Cross-Validation erfolgen. Statt des Control Sets als Repräsentant der Gesamtpopulation wird die Güte des Systems an den bereits von Reviewern im Trainingsprozess durchgesehenen Dokumenten (bei denen daher bekannt ist, welche tatsächlich relevant sind und welche nicht) gemessen. Da dieses Set dann aber nicht zufällig aus dem «search universe» erzeugt wurde, sind Vorhersagen der Messgrößen «recall», «precision» und «F₁» auf die Gesamtpopulation (also die Menge aller Dokumente im «search universe») aus statistischer Sicht mit Vorsicht zu genießen.

Zusätzlich kann man sich in Fällen, in denen ein Control Set mit ausreichender Anzahl an tatsächlich relevanten Dokumenten nicht zur Verfügung steht, mit weiteren statistischen Messgrößen behelfen. Diese sind allerdings auch nur mit Vorsicht zu genießen und können mangels konkreter Validierungsmethode nicht mehr bieten, als eine bereits vorhandene Vermutung über die Güte des Systems (etwa aus Cross-Validation) zu stärken oder zu schwächen. Als derartige Kennzahlen können hier etwa herangezogen werden:

¹³ GROSSMAN/CORMACK, Continuous Active Learning (Fn. 9), S. 32 (S. 36).

¹⁴ GROSSMAN/CORMACK sprechen etwa von rund 70 Dokumenten, siehe GROSSMAN/CORMACK, Continuous Active Learning (Fn. 9), S. 32 (S. 36). Die tatsächlich ausreichende Anzahl ist allerdings vom Projekt und der Ausgestaltung des «search universe» abhängig.

¹⁵ «Recall» bezeichnet den Anteil der tatsächlich relevanten Dokumente, die vom System auch als relevant gekennzeichnet werden; vgl. dazu GROSSMAN/CORMACK, The Grossman-Cormack Glossary of Technology-Assisted Review, Federal Courts Law Review, 2013, S. 1 (S. 27).

¹⁶ «Precision» bezeichnet den Anteil der vom System als relevant gekennzeichneten Dokumente, die tatsächlich relevant sind; vgl. dazu GROSSMAN/CORMACK, The Grossman-Cormack Glossary (Fn. 15), S. 1 (S. 25).

¹⁷ Vgl. dazu etwa GROSSMAN/CORMACK, Technology-Assisted Review (Fn. 11), S. 15 f.

- Die Anzahl der bisher gefundenen tatsächlich relevanten Dokumente im Vergleich zu einer statistischen Schätzung auf Basis auch eines «schlechten» Control Sets mit wenig relevanten Dokumenten. Diese Schätzung kann allerdings nur innerhalb eines bestimmten Konfidenzintervalls gegeben werden, das immer größer (und somit unbrauchbarer) wird, je weniger tatsächlich relevante Dokumente im Control Set enthalten sind.
- Die Entwicklung von Iteration zu Iteration in einem «cumulative gains» Graphen. Ein derartiger zweidimensionaler Graph zeigt auf der x-Achse den Anteil der gegengeprüften Dokumente und auf der y-Achse den Anteil der tatsächlich relevanten Dokumente. Unter der Annahme, dass alle Dokumente sortiert nach der ihr vom TAR-System gegebenen Relevanz vorliegen und man die Gegenprüfung vom Dokument mit dem höchsten Relevanzwert zum Dokument mit dem niedrigsten Relevanzwert durchgeht, zeigt der Graph, wie viele Dokumente durchgesehen werden müssten, um statistisch gesehen einen bestimmten prozentuellen Anteil an tatsächlich relevanten Dokumenten zu erhalten. Je näher der sich für eine konkrete Iteration zeigende Graph an der linken oberen Ecke liegt, umso besser.
- Die Entwicklung von Iteration zu Iteration in einem «receiver operator characteristic» Graphen (oder «ROC Graph»). Ein derartiger zweidimensionaler Graph zeigt auf der x-Achse die «false positive rate»¹⁸ und auf der y-Achse die «true positive rate».¹⁹ Der «ROC Graph» zeigt nun das Verhältnis von «false positive rate» und «true positive rate», angenommen, alle Dokumente liegen sortiert nach der ihr vom System gegebenen Relevanz vor und die Gegenprüfung der Dokumente erfolgt beginnend mit dem Dokument mit dem höchsten Relevanzwert bis zum Dokument mit dem niedrigsten Relevanzwert. Die beste Kurve hat typischerweise die größte Fläche unter sich²⁰. Die beste Kurve ist also auch hier die Kurve, die steil nach oben und dann nach rechts geht (also in die linke obere Ecke).

Daneben sind noch viele andere Kennzahlen denkbar, die man aus statistischer Sicht allerdings vermeiden und nicht heranziehen sollte.²¹

8. Fazit

Mittels gut geplante Einsatz von TAR lassen sich große Dokumentenmengen zielgerichtet und effizient auf Relevanz für bestimmte Sachverhaltsfragen durchsuchen. Hierbei ist aber einiges zu beachten, insbesondere für die Erstellung des Seed Sets, die Festlegung der Stop-Bedingung und die Verschränkung der juristischen Reviewer mit Data Scientists. Selbst wenn ein TAR-Prozess für eine bestimmte Fragestellung abgebrochen werden muss, weil der Einsatz von TAR trotz guter Planung und Durchführung dennoch nicht die gewünschten Ergebnisse liefert, ist es durchaus möglich, dass mit TAR zumindest in viel kürzerer Zeit eine ausreichende Zahl von tatsächlich relevanten Dokumenten gefunden wird, die sich als Startpunkt für weitere (dann allenfalls lediglich manuell durchgeführte) Analysen eignen.

¹⁸ «False positive rate» bezeichnet den Anteil der Dokumente, die vom System als relevant markiert wurden, aber tatsächlich nicht relevant sind; vgl. dazu GROSSMAN/CORMACK, The Grossman-Cormack Glossary (Fn. 15), S. 1 (S. 17).

¹⁹ «True positive rate» bezeichnet den Anteil der Dokumente, die vom System als relevant markiert wurden und tatsächlich relevant sind; vgl. dazu GROSSMAN/CORMACK, The Grossman-Cormack Glossary (Fn. 15), S. 1 (S. 33).

²⁰ Deshalb werden «ROC Graphen» auch oft «area under the curve» Graphen bezeichnet.

²¹ Vgl. etwa die Beispiele bei GROSSMAN/CORMACK, Continuous Active Learning (Fn. 9), S. 32 (S. 37).