KI-Einsatz für Effizienzgewinne bei Benchmarkstudien im Bereich Transfer Pricing

Mittels Web Crawling und Natural Language Understanding

Mittels KI zur Teilautomatisierung des qualitativen Screenings bei der Erstellung von TP Benchmarkstudien entwickeln die Fachexperten der WTS und das DFKI gemeinsam einen Software-Prototyp – auch um so dem kompetitiven Umfeld mit einer digitalen Lösung entgegenzutreten.

ALEXANDER BEUTHER, PROF. DR. PETER FETTKE, DR. VANESSA JUST UND ANDREAS RIEDL

1. Einleitung

„Transfer Pricing – it's an art, not a science!" besagt ein bekanntes Zitat. Die angesprochene Kunst stößt jedoch manchmal an äußerst unsanfte Restriktionen. In der Praxis der Verrechnungspreisarbeit gibt es einige Herausforderungen bei der Bestimmung von Fremdvergleichsdaten. Im Rahmen der Ermittlung angemessener Transferpreise ist der international anerkannte Maßstab des Fremdvergleichsgrundsatzes anzuwenden. Typischerweise erfordert der Grundsatz des Fremdvergleichs basierend auf § 1 Abs. 1 AStG, dass Transaktionen zwischen verbundenen Unternehmen mit denen von unabhängigen Unternehmen vergleichbar sind. Die Analysen, die in diesem Zusammenhang angefertigt werden, stoßen schnell an ihre Grenzen.[1]

Ganz konkret geht es um sog. „Benchmarkstudien". In den Studien fällt regelmäßig der größte Teil der Arbeit für das manuelle Screening der Vergleichsunternehmen an. In der Praxis werden Webseiten angesteuert, beurteilt und Screenshots gemacht; nicht selten von mehreren hundert Unternehmen. Dabei ist die Beurteilung der Webseiten subjektiv und – durch den Zeitaufwand für das Screening – die Suchmöglichkeiten beschränkt. Insgesamt ist dem Prozess so inhärent, dass die Ergebnisse durch eine Verbesserung des manuellen Screening-Prozesses verbessert werden können.[2] Gerade die weitere Automatisierung des Prozesses und das hohe Potential von Künstlicher Intelligenz (KI) in diesem Zusammenhang sorgt hier für euphorische Aussichten.[3]

Digitale Technologien haben bereits in vielen Bereichen Revolutionen ausgelöst. Digitale Fotografie beispielsweise lässt analoge Fotografie antiquarisch wirken. Online-Handel hängt den stationären Handel seit Jahren hinsichtlich des Umsatzwachstums ab. Diese Veränderungen haben die jeweiligen Strukturen auf den Märkten nachhaltig verändert. Auch Steuerabteilungen und Steuerberatungen müssen sich ebenso verändern, um wettbewerbsfähig zu bleiben.[4] Vor allem bei Verrechnungspreisen trifft dies besonders zu, da das Volumen interner Transaktionen laut Schätzungen des BMF und der OECD etwa 60 % des weltweiten Handels ausmacht. Es handelt sich um eine signifikant hohe Anzahl an internen Transaktionen, hohe Volumina und erhebliche Prozesskosten.[5] Aufgrund des enormen technischen Potentials in diesem Zusammenhang erforschen die Autoren innovative Konzepte zur Digitalisierung aller Aufgaben, die im Zusammenhang mit Verrechnungspreisen stehen. Hierzu wurde insbesondere auch ein Prototyp entwickelt, der basierend auf jahrzehntelanger Erfahrung bei der Erstellung von Benchmarkstudien den Arbeitsschritt des manuellen Screenings unterstützt.

---

1 Zu den Grenzen von Benchmarkstudien ua *Krüger/Nientimp/Schwarz* Ubg 2016, 221.

2 Zu den weiteren Suchkriterien im Rahmen der Erstellung von Benchmarkstudien ua *Vögele/Borstell/Bernhardt*, Verrechnungspreise, Kap. H: Berechnung, Benchmarking, Profit Split für Gewinnmethoden, Rn. 123–125.

3 Zur Verbesserung der Qualität hat die OECD ansonsten ein Toolkit veröffentlicht (A Toolkit for Addressing Difficulties in Accessing Comparables Data for Transfer Pricing Analyses in https://www.oecd.org/tax/toolkit-on-comparability-and-mineral-pricing.pdf). Die grundsätzliche Problematik im Rahmen des manuellen Screenings wird hierdurch jedoch nicht beseitigt.

4 *Fettke/Herzog/Lahann/Maus/Neumann*, Künstliche Intelligenz im Steuerbereich: Innovationsstudie zur Digitalisierung und den Potentialen Künstlicher Intelligenz im Bereich Steuer, WTS Group AG Steuerberatungsgesellschaft, Deutsches Forschungszentrum für Künstliche Intelligenz, 2017.

5 *Kleinbietpaß/Hanken*, Verrechnungspreise – inkl. eBook: im Spannungsfeld von Controlling und Steuern, 2014, Vol. 1207.

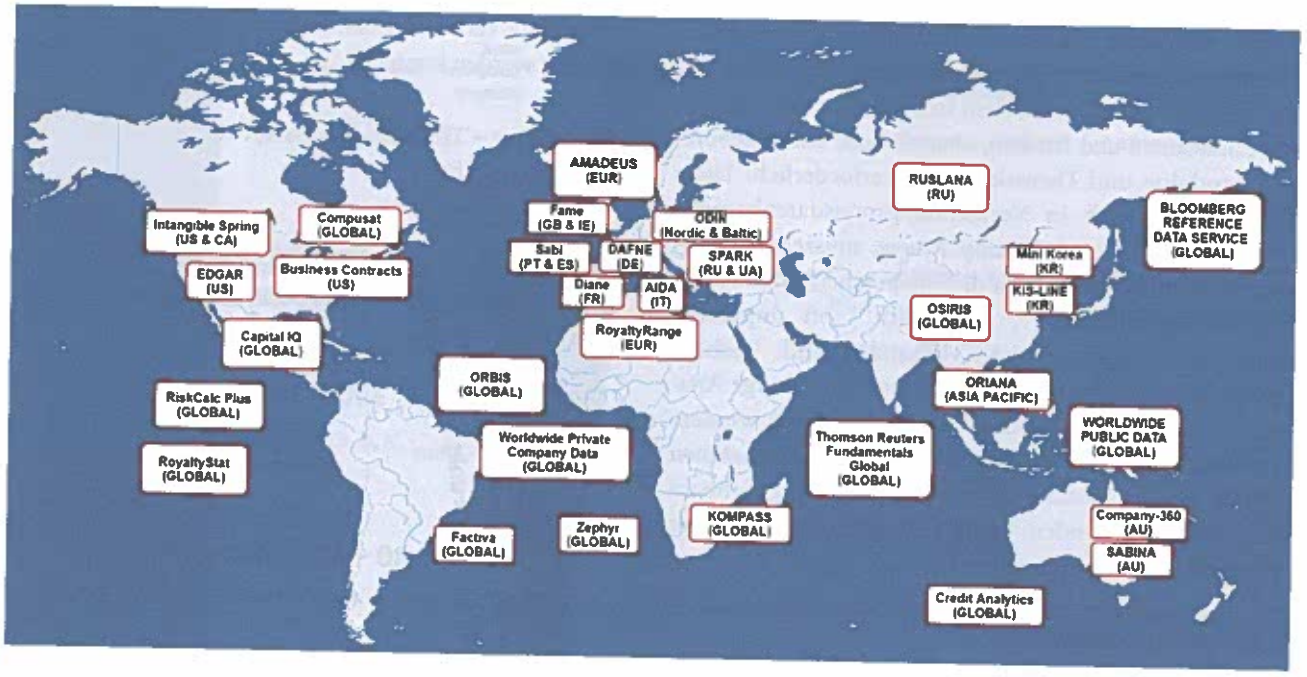


Abb. 1: Überblick global verwendeter Datenbanken

Dieser von DFKI und WTS Service Line Transfer Pricing entwickelte Prototyp soll die Benchmarking-Experten der WTS bei der Selektion von vergleichbaren Unternehmen aus Transfer Pricing-Datenbanken unterstützen, indem der manuelle Schritt der Bestimmung der Unternehmensfunktion durch eine automatische Textanalyse ergänzt wird. Dadurch wird eine Steigerung der Geschwindigkeit sowie eine Reduktion von manuellem Aufwand und damit einhergehend eine Kostenreduktion bei der Erstellung von Benchmarkstudien erwartet.

Im folgenden Beitrag wird das Vorhaben zuerst in einen größeren wissenschaftlichen Kontext eingeordnet und die Problemstellung beschrieben, anschließend wird ein konkreter Anwendungsfall dargestellt, bevor der technische Prototyp im Allgemeinen und die Technik dahinter vorgestellt wird. Der Vorstellung schließt sich eine Einordnung des Umgangs in einem Praxisbeispiel sowie die Evaluation der Methodik an. Anschließend werden besondere neue Nutzungsmöglichkeiten und Herausforderungen dargestellt sowie diskutiert.

2. Herausforderungen und Problemstellung

Verrechnungspreise sind in steuerlicher Hinsicht „Preise und Konditionen für grenzüberschreitende Geschäftsbeziehungen zwischen verbundenen Unternehmen sowie zwischen Stammhaus und Betriebsstätte“[6]; es geht also speziell um die wertschöpfungsadäquate Allokation des Steuersubstrats auf die Staaten, die an einer Transaktion beteiligt sind. Allerdings umfassen die damit zusammenhängenden Aufgaben, das sog. Verrechnungspreis-Management, weitaus mehr als die Festlegung der Methodik oder die Erstellung der Dokumentation. Dies bedeutet, dass bei der Digitalisierung der Verrechnungspreise eine Fülle von Aufgaben und Daten zu berücksichtigen ist.

Besonders Verrechnungspreise unterliegen höchsten Anforderungen an die Steuer-Compliance, da diese Preise von mehreren Parteien, den zwei Ländern, die von der Transaktion betroffen sind, akzeptiert werden müssen. Zusätzliche Auflagen durch das BEPS-Projekt, das Country-by-Country-Reporting und ein zunehmender Fokus auf Vergleichbarkeit bedeuten, dass Steuerzahler sicherstellen müssen, dass die internationalen Transaktionen verteidigt werden können. Ansonsten drohen Doppelbesteuerung, Strafzuschläge oder gar strafrechtliche Verfahren.[7] Gleichzeitig existieren aber ein größer werdender Kostendruck und die Erwartung von mehr Leistungen der Steuerfunktion sowie eine zunehmende Deregulierung des Marktes.[8]

Eine Möglichkeit zum Nachweis und zur Dokumentation der Angemessenheit konzerninterner Verrechnungspreise sind Benchmarkstudien. Die Prüfung und Dokumentation der Angemessenheit erfolgt dabei ua anhand von GuV-Daten öffentlicher Jahresabschlüsse einer Peergroup, denn als Maßstab für den Fremdvergleich (*arm's*

[6] *Kleinhietpaß/Hanken* (Fn. 5).

[7] *Kleinhietpaß/Hanken* (Fn. 5).

[8] *Fettke*, Künstliche Intelligenz für die Digitalisierung der Steuerfunktion, REthinking: Tax 01/2019, 12–22.

length principle) dienen Finanzkennzahlen vergleichbarer Unternehmen. Für die Zugehörigkeit zu einer Peergroup ist eine Vergleichbarkeit in Bezug auf die ausgeübten Funktionen und Risiken, aber zB auch der angebotenen Produkte und Dienstleistungen erforderlich. Diese Daten finden sich in Verrechnungspreisdatenbanken, wie *AMADEUS* und *Royalty Range*, müssen aber zusätzlich geprüft werden, da die entsprechenden Einträge („Handelsbeschreibung", „Überblick") oft unpräzise, veraltet oder schlicht nicht vorhanden sind. Deshalb müssen diese Informationen durch eine aufwendige Analyse externer Quellen beschafft und ausgewertet werden. Insbesondere, wenn so mehrere hundert Unternehmen geprüft werden müssen, um einige wenige vergleichbare Unternehmen zu finden, ergibt sich ein hoher manueller Aufwand mit entsprechenden Kosten.

3. Ein Anwendungsfall

Der Prozess zur Erstellung von Benchmarkstudien besteht aus drei Schritten: Dem quantitativen Screening, dem qualitativen Screening und der Anfertigung eines Reports. Beim quantitativen Screening werden Datensätze aus externen Datenbanken (zB *TP Catalyst* von *Bureau van Dijk*) extrahiert. Gefiltert werden diese Datensätze nach individualisierten Suchparametern, die an das zu untersuchende Unternehmen des Auftraggebers angepasst werden. Die Datensätze enthalten neben Finanzkennzahlen ua auch Unabhängigkeitskriterien und Unternehmensbeschreibungen. Alleinstehend sind diese Datensätze noch nicht aussagekräftig für eine Benchmarkstudie. Die Vergleichbarkeitsanalyse muss erst noch durchgeführt werden, und dafür ist ein weiterer Schritt erforderlich.

Im zweiten Schritt erfolgt ein manuelles Screening der gefilterten Datensätze. Der Schwerpunkt liegt dabei auf der Untersuchung der Webseiten der potentiellen Vergleichsunternehmen. Das Ziel ist es, Unternehmen zu finden, deren Spektrum ausgeübter Funktionen und Tätigkeitsschwerpunkte mit der zu testenden Einheit der Unternehmensgruppe (möglichst weit) übereinstimmen. Diese manuelle Suche nach einer passenden Peergroup ist derzeit mit einem hohen personellen Aufwand verbunden. Sowohl das manuelle Aufrufen der Webseiten als auch die Klassifizierung der Unternehmen ist enorm zeitintensiv. Darüber hinaus muss der gesamte Prozess prüfungssicher (zB über das Anfertigen von Screenshots) dokumentiert werden.

Der dritte Schritt besteht aus der Dokumentation des Prozessablaufs und der gewonnenen Ergebnisse. Für die Erstellung des Reports werden standardisierte Vorlagen verwendet, so dass der Report automatisch generiert werden kann.

4. Prototyp – Technische Unterstützung des Anwendungsfalls

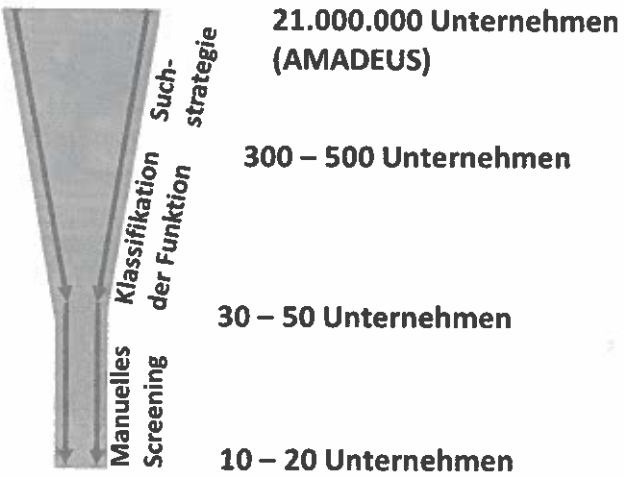
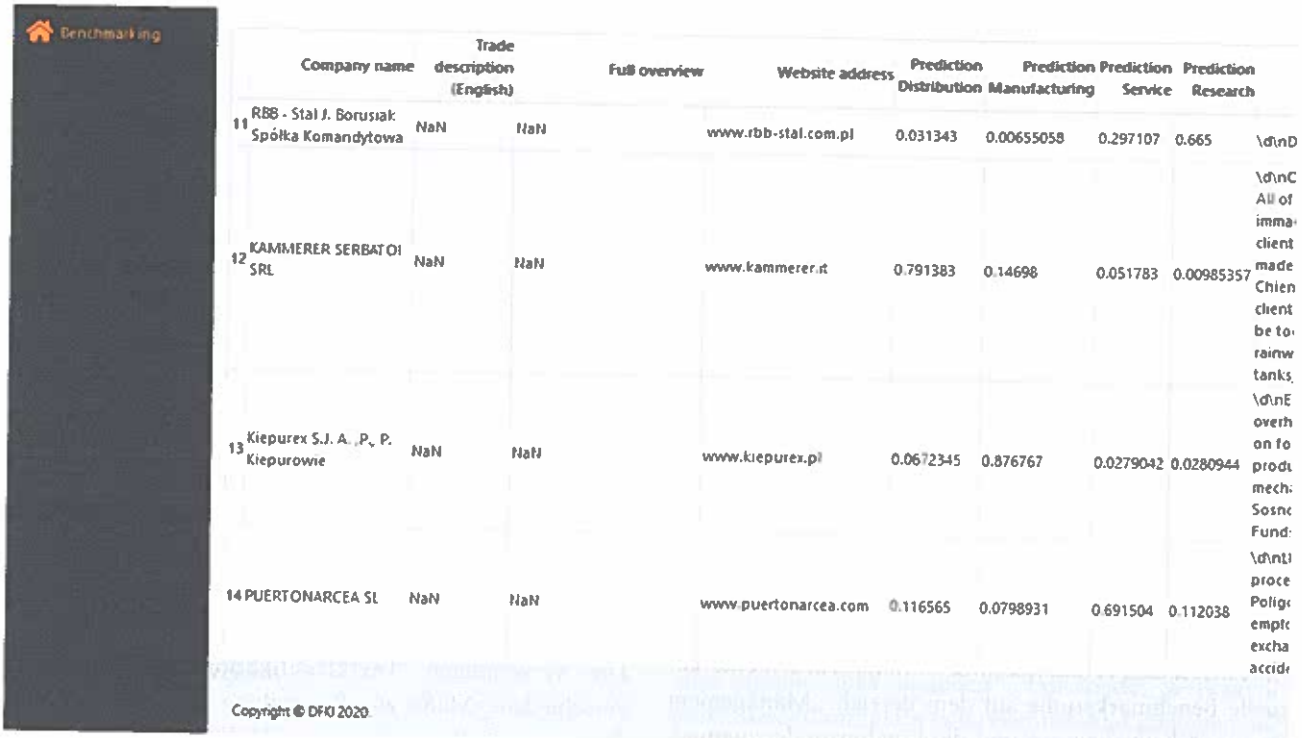


Abb. 2: Konsolidierung zum Fremdvergleich

Um den Aufwand bei der Erstellung von Benchmarkstudien zu reduzieren, liegt es nahe, den manuellen Screening-Prozess vom Umfang der zu testenden Einheiten aus der Datenbanksuche her möglichst klein zu halten. Dazu muss die Anzahl der zu prüfenden Unternehmen weiter reduziert werden. Konkret soll hier eine Vorauswahl getroffen werden, welche Unternehmen eine von der Suche abweichende Unternehmensfunktion besitzen und diese ausgeschlossen werden. Dadurch schrumpft die Anzahl möglicher Kandidaten, die manuell geprüft werden müssen, und damit die Arbeitslast der Experten.

Diese Reduzierung der möglichen Kandidaten erfolgt durch einen Zwischenschritt zwischen Datenbanksuche und manuellem Screening. Dabei werden die gefundenen Kandidaten hinsichtlich ihrer Unternehmensfunktion klassifiziert. Dies erfolgt analog zur manuellen Bearbeitung, die das Aufrufen der Unternehmenswebseite, das Übersetzen der Inhalte und das Lesen und Verstehen der Webseitentexte, inkl. der Bewertung umfasst. Dazu wird

Abb. 3: Upload der Input-Daten in den Demonstrator



Benchmarking

	Company name	Trade description (English)	Full overview	Website address	Prediction Distribution	Prediction Manufacturing	Prediction Service	Prediction Research	
11	RBB - Stal J. Borusiak Spółka Komandytowa	NaN	NaN	www.rbb-stal.com.pl	0.031343	0.00655058	0.297107	0.665	\d\nD
12	KAMMERER SERBATOI SRL	NaN	NaN	www.kammerer.it	0.791383	0.14698	0.051783	0.00985357	\d\nC All of imma client made Chien client be to rainw tanks,
13	Kiepurex S.J. A., P., P. Kiepurowie	NaN	NaN	www.kiepurex.pl	0.0672345	0.876767	0.0279042	0.0280944	\d\nE overh on fo produ mech: Sosnc Fund:
14	PUERTONARCEA SL	NaN	NaN	www.puertonarcea.com	0.116565	0.0798931	0.691504	0.112038	\d\nU proce Polig empl excha accid

Copyright © DFKI 2020

Abb. 4: Tabelle mit beispielhaften Ergebnissen

erst der Datenbankauszug als Excel-Datei in den Prototypen geladen (s. Abb. 3 auf S. 318). Anschließend werden die drei Module für jeden Eintrag in der Excel-Datei durchlaufen.

Das Modul „Webseiten Aufrufen" oder auch „Web Crawling" steuert automatisiert eine Instanz eines definierbaren Browsers, wie Google Chrome, Microsoft Edge oder Mozilla Firefox, und imitiert das menschliche Verhalten, indem die URL aus dem Datenbankauszug kopiert und aufgerufen wird sowie der angezeigte Text in den Prototypen kopiert wird. Links zu Unterseiten werden identifiziert, aufgerufen und ebenfalls kopiert.[9] Das zweite Modul übersetzt maschinell Texte[10] und entfernt irrelevante Passagen, die zuvor definierte Schlagworte enthalten oder weiteren Selektionskriterien (Zeichen- und Wortanzahl, Vorkommen von Zahlen, Sonderzeichen etc) entsprechen.[11] Das dritte Modul, das Lesen und Verstehen der Webseitentexte, ist technologisch nicht mit dem menschlichen Verstehen von Texten gleichzusetzen. Es können jedoch Modelle trainiert werden, die bestehende Texte inklusive einer Annotation der Unternehmensfunktion enthalten. Dieser Textkorpus dient dann als Referenz für eine Unternehmensfunktion. Neue, unbekannte Texte können dann anhand ihrer Ähnlichkeit zu den Texten einer Unternehmensfunktion ebendieser zugeordnet werden.

Dazu wurden in früheren Benchmarkstudien akzeptierte Unternehmen als Referenz für die Unternehmen einer Geschäftsfunktion herangezogen. Die Texte wurden aus den zugehörigen Webseiten kopiert und dem jeweiligen Cluster hinzugefügt. Damit wurde ein Wort-Vektor-Modell erstellt, das Merkmale der eingegebenen Texte einer Kategorie enthält und mit dessen Hilfe Unternehmen gleicher Funktion klassifiziert werden können. Grundlage dazu sind die Häufigkeiten von Wörtern und das inverse der Dokumentenhäufigkeit.[12] Mit diesem Maß von Häufigkeiten von Wörtern und Dokumenten lässt sich der Abstand von Dokumenten messbar machen. Das nutzen die Modelle, um neue, bisher nicht gesehene Texte analysieren und Texte zu der am besten passenden Kategorie zuordnen zu können.

Diese Einschätzung der Unternehmensfunktion soll den TP-Spezialisten bei der Selektion von passenden Unternehmen unterstützen. Dies geschieht mit dem Ergebnis

[9] Vgl. *Liu*, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2007.

[10] Hierfür können sowohl kommerziell vorhandene Dienste wie *DeepL* oder *Google Translate* als auch speziell für den Steuerbereich von der WTS und dem DFKI entwickelte maschinelle Übersetzer zum Einsatz kommen.

[11] Vgl. *Vijayarani/Ilamathi/Nithya*, Preprocessing Techniques for Text Mining – an Overview, International Journal of Computer Science & Communication Networks 2015, 5(1), 7–16.

[12] Vgl. *Qaiser/Ali*, Text mining: Use of TF-IDF to Examine the Relevance of Words to Documents, International Journal of Computer Applications 2018, 181(1), 25–29.

der Prototypen in Form einer tabellarischen Auflistung, die die Datenbankeinträge Unternehmensname, Handelsbeschreibung und Überblick sowie die vorhergesagten prozentualen Werte für die jeweilige Unternehmensfunktion und den aus dem Internet kopierten Text ausgibt.

Dieser Prototyp soll zwei grundlegende Charakteristika abdecken: Erstens hat er einen explorativen Charakter und soll dazu dienen, nachzuweisen, dass die verwendeten Techniken und zugrunde liegenden Ideen zur Lösung des (Teil-)Problems tauglich sind. Zweitens sollen auch erste praktische Erfahrungen mit dem Prototyp gesammelt werden, welche dann in umfangreiche Problemanalysen und Systemspezifikationen für die Weiterentwicklung eingehen.[13]

5. Praxisbeispiel

Im Einsatz des Prototyps in der Praxis wurden eine aktuelle Benchmarkstudie auf dem Bereich „Management Services" herangezogen und die Ergebnisse der manuellen Suche mit den Ergebnissen des Prototyps verglichen. Dazu wurde die Benchmarkstudie um eine erste Auswertung der Ergebnisse erweitert, da in der klassischen Analyse nur das Ergebnis Unternehmensfunktion „stimmt überein" oder „weicht ab" eingetragen wird. Dazu wurde die gefundene Unternehmensfunktion für jedes zu testende Unternehmen in das Kommentarfeld eingetragen. Führen Unternehmen mehrere Unternehmensfunktionen aus, wurden sie im manuellen Screening der vermeintlich am stärksten ausgeprägten Funktion zugeordnet. Als Vergleichswert wurde die von der Textklassifikation am wahrscheinlichsten angesehene Unternehmensfunktion verwendet. Die Ergebnisse dieses ersten Tests wurden zur Validierung des Prototyps verwendet (siehe Evaluation).

Mit dem ersten Einsatz des Prototyps konnten die Fachexperten der *WTS-Service Line Transfer Pricing* folgende drei Möglichkeiten zum Einsatz im Anwendungsfall identifizieren:

1. Die Klassifizierungsergebnisse werden in einer Art Cockpit angezeigt und bieten einen weiteren Indikator zur Bewertung des Sachverhalts. Der Economist der Service Line führt weiterhin die manuellen Analysen durch. Der Anwendungsfall ändert sich in diesem Fall nicht.
2. Die Klassifizierungsergebnisse werden durch einen Fachexperten final genehmigt bzw. eingegrenzt. Hier führt der Prototyp die Bewertung und ein Fachexperte einen Review der Klassifikation durch. Der Anwendungsfall verändert sich, aber die Zahl der zu testenden Unternehmen bleibt gleich.
3. Der Prototyp schließt selbständig Unternehmen mit eindeutig nicht passender Unternehmensfunktion aus. Der Anwendungsfall ändert sich, indem ein Teil des manuellen Screenings ersetzt wird. Die Zahl der manuell zu testenden Unternehmen verringert sich.

Dadurch wird der außerordentlich zeitintensive manuelle Screening-Prozess zur Auswahl der geeigneten Vergleichsunternehmen zu einem Großteil durch KI ersetzt. Nach Schätzungen der Mitarbeiter der Service Line wird eine Zeitersparnis von bis zu 80 % erwartet, denn der Spezialist für Transferpricing hat am Ende lediglich die Aufgabe, das manuelle Screening von ca. 30 Unternehmen zu verproben, anstatt mehrere hundert Unternehmen selbst zu screenen.

6. Evaluation

Die verwendeten Textklassifikationsmodelle nutzen verschiedene Maße zur Beurteilung der Relevanz von Termen in Dokumenten einer Dokumentensammlung. Nach einem Vergleich der Methoden TF, TF-IDF und Word2Vec erweist sich hier die TF-IDF-Methode am effektivsten. TF steht hierbei für die Term Frequency, also die (relative) Häufigkeit von Wörtern in einem Text, welche durch Berücksichtigung der inversen Dokumentenhäufigkeit (IDF – Inverse Document Frequency), also der Spezifität eines Terms für die Gesamtmenge der betrachteten Dokumente multipliziert wird. Durch diese Verrechnung wird verhindert, dass ein vielfach vorkommender, aber irrelevanter Term nicht auch in gleichem Maße zur Relevanz der Klassifikation beiträgt.[14]

Für das Clustern selbst, also die Gruppierung von Texten zu einer Kategorie und auch die gezielte Abgrenzung zu einer anderen Kategorie, stellten sich nach einem Vergleich der Methoden SVM, KNN, MNB und labled-LDA die „support vector machines" als besonders geeignet heraus. Zur Evaluation wurde der og Trainingsdatensatz in 85 % Trainingsdaten und 15 % Testdaten aufgeteilt. Damit wurde auf den Trainingsdaten eine Treffergenauigkeit („accuracy") von 97 % und eine Trefferquote („recall") von 95 % erreicht. Auf den Testdaten wurde eine Treffergenauigkeit von 80 % und eine Trefferquote von 77 % erreicht. Die Analyse der Ergebnisse ist in der Wahrheitsmatrix (Konfusionsmatrix, „confusion matrix") in Abb. 5 (S. 321) zu sehen, in der die Anzahl richtig und falsch klassifizierter Dokumente erkennbar ist. Das zuvor beschriebene Praxisbeispiel

[13] Vgl. *Warfel*, Prototyping: A Practitioner's Guide, 2009.

[14] Vgl. *Felden/Bock/Gräning/Molotowa/Saat*, Evaluation von Algorithmen zur Textklassifikation, No. 10/2006, Freiberger Arbeitspapiere.

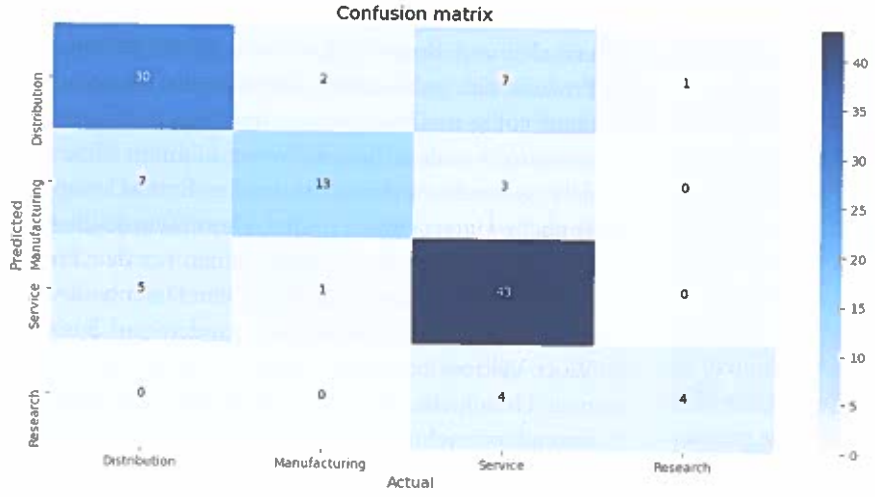


Abb. 5: Die Wahrheitsmatrix zeigt die richtig und falsch klassifizierten Texte.[15]

diente als eine Validierung, und es wurde eine Trefferquote von 72 % erreicht.

7. Herausforderungen

Der Prototyp hat einen explorativen und experimentellen Anspruch, woraus sich Herausforderungen an verschiedene Module ergeben. Für eine bessere Übersicht zeigt Abb. 6 die einzelnen Schritte des Prototyps mit den zugehörigen Herausforderungen in der Praxis an.

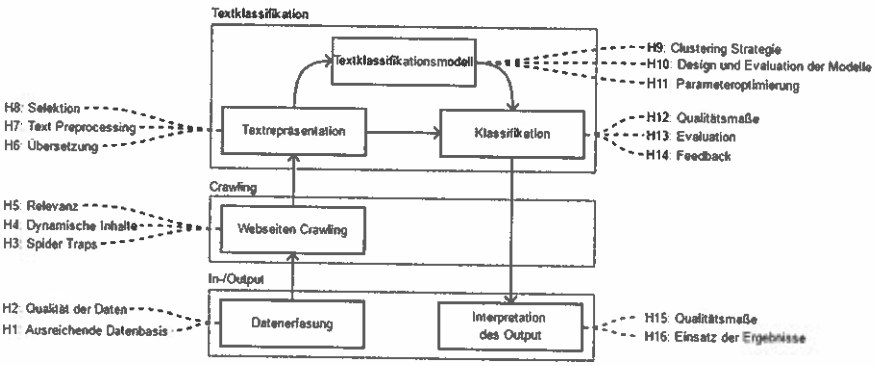


Abb. 6: Übersicht über aktuelle Herausforderungen

Die Herausforderungen H1 und H2 betreffen die Daten, die in das System eingegeben werden. Das sind zum einen die annotierten Daten zum Training der Modelle, zum anderen Daten, die einer Kategorie zugeordnet werden sollen. Es wurde ersichtlich, dass die Daten zum Training hohe Qualitätsanforderungen erfüllen müssen[16], ua Verfügbarkeit, Umfang und maschinelle Auswertbarkeit, um später für neue Daten eine zuverlässige Klassifikation zu berechnen. Weiter müssen hier die Umfänge der einzelnen Kategorien gleich häufig sein, um eine Über- bzw. Unterrepräsentation zu vermeiden. Ebenfalls müssen neue Daten zur Klassifikation diesen Ansprüchen in ähnlicher Weise gerecht werden.

H3–H5 betreffen die Extraktion von Webseiteninhalten von realen Unternehmenswebseiten. Dabei sind vor allem sog. „Spider Traps" zu beachten, die diesen Vorgang bewusst verhindern sollen. Dabei werden beispielsweise versteckte Texte in den Webcrawler eingeführt, die irreführende Informationen enthalten oder falsche Links ausgeben.[17] Ein solider Webcrawler muss diese Manipulationen erkennen und umgehen oder blockieren können. Zudem ist ein dynamischer Webseiteninhalt problematisch. Hier wird Inhalt kontextabhängig erzeugt und weist mit jedem Aufrufen der Webseite Unterschiede auf. Ein Beispiel sind Online-Shops, die jedem Kunden andere Produkte auf der Startseite präsentieren. Des Weiteren gibt es Unterschiede in der Qualität und Relevanz von Textpassagen und Unterwebseiten. Hier muss ein Verfahren irrelevante und qualitativ weniger gute Passagen erkennen und ausschließen (zB AGBs).[18]

Die Herausforderungen H6–H8 betreffen die Verarbeitung des Textes in ein einheitliches, maschinell verwertbares Format. Herausfordernd ist hier eine solide Übersetzung der Texte in eine einheitliche Sprache, selbst wenn der gesamte Text zwischen unterschiedlichen Passagen in der Sprache wechselt. In der anschließenden Vorverarbeitung werden definierte Stoppwörter und Muster ausgeschlossen. Die Wahl der Worte und Muster ist hier von entscheidendem Interesse, um irrelevante Passagen auf sehr feingranularer Ebene ausschließen zu können. Zudem müssen die Texte in ein maschinell verwertbares Format gebracht werden.

[15] X-Achse = die tatsächlichen Kategorien; y-Achse = die vorhergesagten Kategorien.

[16] Vgl. *Knight/Burn*, Developing a Framework for Assessing Information Quality on the World Wide Web, Informing Science 2005, Vol. 8.

[17] Vgl. *Pant/Srinivasan/Menczer*, Crawling the Web, Web dynamics 2004, 153–177.

[18] Vgl. *Hernández/Rivero/Ruiz*, Deep Web Crawling: A Survey, World Wide Web 2019, 22(4), 1577–1610.

Hier ist zwischen verschiedenen Repräsentationen zu wählen.[19]

H9–H11 sind Herausforderungen bei der Modellerstellung. Dafür müssen relevante Qualitäts-Metriken festgelegt werden[20] und verschiedene Klassifizierungsstrategien und -modelle[21] optimiert und verglichen werden, so dass die Qualitäts-Metriken bestmögliche Werte messen. Die Herausforderung besteht dabei in der geschickten Parameteroptimierung, um die Verfahren bestmögliche Ergebnisse erzeugen zu lassen. Ebenfalls von hohem Interesse sind hier die Identifikation und Analyse von Dokumenten, die zu einer Kategorie gehören, aber missverständliche Werte liefern, beispielsweise durch eine große Distanz zum Cluster.

H12–H14 betreffen die Klassifikation von neuen, unbekannten Texten. Hier muss die Klassifikation mittels der vorher definierten Qualitäts-Metriken bestimmt und evaluiert werden. Ein Feedback bzw. eine Interpretation über die Ergebnisse vom Anwender (H15) bzgl. der Klassifikation kann zusätzlich zum Verbessern des Modells einfließen (sog. „Reinforcement Learning"). Hier müssen auch Erfahrungen darüber gewonnen werden, unter welchen Bedingungen gute Ergebnisse erzielt werden und wo genau Verbesserungsbedarf besteht. Auch die Ergebnisse, die üblicherweise aus Wahrscheinlichkeiten bestehen, müssen in geeigneter Weise interpretiert werden.

8. Diskussion

Die vorgestellten Arbeiten sind technisch innovativ und haben einen prototypischen Charakter. Erste Ergebnisse sind vielversprechend, und es wird weiter an Verbesserungen gearbeitet. Dabei zeigen sich folgende Ergebnisse:

1. Die Unternehmensfunktionen sind nicht einheitlich definiert. Die hier verwendete Definition entstammt der Benchmarking-Praxis und soll hauptsächlich zwischen groben Unternehmensfunktionen unterscheiden. Daneben existieren viele weitere Definitionen, beispielsweise die des Statistischen Bundesamtes. Inwieweit eine feingranulare Unterteilung weitere Vorteile bringen könnte, ist zukünftige Forschungsarbeit.
2. Es ist derzeit noch unklar, ob und wie der maschinelle Arbeitsschritt von den Steuerbehörden akzeptiert wird.
3. Die hier vorgeschlagenen Vektor-Modelle zur Textverarbeitung betten sich derzeit in die bestehenden Prozesse ein. Jedoch liegt hier auch das Potential, um diese völlig unabhängig von der Unternehmensfunktion anzuwenden. Beispielsweise könnten diese Modelle generell zu der zu testenden Entität möglichst ähnliche Unternehmen finden. Damit würde die Ähnlichkeit zwischen zwei Unternehmen für den Fremdvergleich nicht mehr auf Basis der Datenbanksuche und damit der Branchencodes, sondern auf Basis des Wort-Vektors bestimmt werden. Das ist eine radikal neue Herangehensweise bei der Suche nach vergleichbaren Unternehmen.
4. Der erzeugte Datensatz ist derzeit stark unterschiedlich verteilt; die Kategorie „Dienstleistungen" enthält 322 Texte, während die Kategorie „Forschung" nur 56 Texte umfasst. Das hat zur Folge, dass es zur Über- bzw. Unterrepräsentation von Kategorien kommt.

9. Resümee

Um sich der Herausforderung der Verbesserung des manuellen Screenings bei der Erstellung von Benchmarkstudien via KI zu stellen, wurde ein Webcrawler entwickelt, der die Webseiten der zu untersuchenden Unternehmen eigenständig aufruft und aus dem Text relevante Inhalte extrahiert. Zusätzlich wurden verschiedene Textklassifikationsverfahren erprobt, die bisher unbekannte Texte automatisiert einer Unternehmensfunktion zuordnen sowie deren Klassifikationsleistung miteinander vergleichen. Grundlage waren die Ergebnisse von bereits abgeschlossenen Benchmarkstudien, die jeweils nur Unternehmen einer bestimmten Unternehmensfunktion enthalten.

Mit dieser Einschätzung der Unternehmensfunktion und einem Webcrawler ist eine erste Einschätzung zur Bewertung des Sachverhalts möglich und kann direkt durch den Webseitentext bestätigt oder im Falle einer Falschklassifikation widerlegt werden. Aufgrund dessen wird eine wesentliche Reduktion des Zeitaufwands für die Erstellung von Benchmarkstudien erwartet.

Die durch den KI-Einsatz realisierten Effizienzgewinne im Rahmen der Erstellung von Benchmarkstudien mittels Webcrawler und maschinellem Natural Language Understanding im Bereich Transfer Pricing gründen sich dabei maßgeblich auf die folgenden drei Punkte:

1. erhebliche Zeit- und Kostenersparnis auf Seiten der steuerlichen Fachkollegen;
2. einheitliche, standardisierte und technologiegestützte Herangehensweise;
3. Compliance-Anforderungen bleiben gewahrt durch 4-Augen-Prinzip (Maschine-Mensch).

---

[19] Siehe *Uysal/Gunal*, The impact of Preprocessing on Text Classification, Information Processing & Management 2014, 50(1), 104–112.

[20] Vgl. *Forman*, An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research 2003, 3 (Mar), 1289–1305.

[21] Vgl. *Aggarwal/Zhai*, A Survey of Text Classification Algorithms, Mining text data 2012, 163–222.

Derzeit wird eine gute Gesamtgenauigkeit bei der maschinellen Klassifikation erreicht, wobei es sich hierbei erst um einen aktuellen Arbeitsstand handelt. In der Entwicklung sind zudem noch weitere Ansätze für State-of-the-Art-Klassifikationsverfahren, die unter Einbeziehung externer Daten eine noch höhere Gesamtgenauigkeit erwarten lassen. Da es sich hierbei um eine kontinuierliche Weiterentwicklung handelt, werden des Weiteren die gefundenen Potentiale und Herausforderungen der einzelnen Komponenten weiter ausgenutzt und optimiert. Neben der reinen Zeit- und Kostenersparnis ist somit damit zu rechnen, dass im Bereich der Erstellung von Benchmarkstudien die Qualität der Ergebnisse erhöht werden kann. Dadurch, dass durch den KI-gestützten Screeningprozess deutlich mehr Daten untersucht werden können, steigt auch die Chance auf eine bessere Vergleichbarkeit der finalen Ergebnisse. Langfristig sollte somit das Risiko steuerlicher Anpassungen im Rahmen von Verrechnungspreisprüfungen gesenkt werden können. Dies könnte auch den Umgang mit Interquartilsbandbreiten in der Zukunft verändern.[22] Wie allerdings Betriebsprüfer mit KI-gestützten Benchmarkergebnissen umgehen werden, bleibt in diesem Zusammenhang abzuwarten.

Die Reise ist noch lange nicht zu Ende. Wir werden weiter mittels der Erforschung innovativer Anwendungen der KI und der Begleitung des Innovationstransfers zur Gestaltung des Steuerarbeitsplatzes der Zukunft beitragen. Dabei sehen wir in der Zukunft das Potential, dass wir über hochwertige Massendaten verfügen, um so die Treffergenauigkeit der eingesetzten Tools weiter zu erhöhen. Auch der Austausch der interdisziplinären Fachteams wird eine zusätzliche Bereicherung darstellen. Gerade in Zeiten von weiteren regulatorischen Verschärfungen im Bereich der Verrechnungspreise sollten die vorgestellten Ansätze die Hoffnung vermitteln, dass durch zukünftige technische Lösungen der schier unüberwindliche „Compliance-Berg" bezwingbar werden könnte – „Step ahead in tax and finance".

[22] Zur Diskussion von Interquartilsbandbreiten ua OECD Guidelines, July 2017, Chapter III, Paragraph 3.57; *Heidecke/Sachs* IWB 2019, 716; *Schwarz/Stein/Flanderova/Hoffmann* Ubg 2017, 638.

ALEXANDER BEUTHER

Wissenschaftler am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrücken.

PROF. DR. PETER FETTKE

Professor an der Universität des Saarlandes und Leiter des Competence Centers Tax Technology am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrücken.

DR. VANESSA JUST

Geschäftsführerin wtsAI, Hamburg.

ANDREAS RIEDL

Partner Transfer Pricing und Head of TP Technology WTS GmbH, Frankfurt a. M..

Die Autoren danken den Kolleginnen *Dr. Birgit Friederike Makowsky* (Director WTS GmbH), *Andrea Groiß* (Senior Manager WTS GmbH), *Christina Derer* (Manager WTS GmbH) und *Lisa Ziegler* (Professional WTS GmbH) für ihre wertvollen Beiträge.

Feedback bitte an digitax@beck.de.