Weißmann, Julius; Herbst, Tim

Article

Maschinelles Lernen im Basisregister für Unternehmen: Vorstudie zum Potenzial automatischer Konsolidierung von Unternehmensstammdaten

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Weißmann, Julius; Herbst, Tim (2024) : Maschinelles Lernen im Basisregister für Unternehmen: Vorstudie zum Potenzial automatischer Konsolidierung von Unternehmensstammdaten, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 3, pp. 67-79

This Version is available at:
https://hdl.handle.net/10419/299145

MASCHINELLES LERNEN IM BASISREGISTER FÜR UNTERNEHMEN

Vorstudie zum Potenzial automatischer Konsolidierung von Unternehmensstammdaten

Julius Weißmann, Tim Herbst

Schlüsselwörter: Datensatzverknüpfung – Registerdaten – Textähnlichkeit – überwachtes Lernen – Datenanalyse

ZUSAMMENFASSUNG

Das Statistische Bundesamt hat den gesetzlichen Auftrag, ein Register über Unternehmensbasisdaten für Deutschland aufzubauen. Die Inhalte und Ziele des Basisregisters für Unternehmen werden in dem zugrunde liegenden Gesetz definiert. Diese liegen vor allem in der Konsolidierung der Basisdaten von Unternehmen aus verschiedenen Quellregistern an einer zentralen Stelle sowie die Gewährleistung von deren Aktualität und Richtigkeit. Die korrekte und effiziente Verknüpfung der Basisdaten von Unternehmen aus den verschiedenen Quellregistern hat hierbei eine herausragende Bedeutung. Dafür hat das Statistische Bundesamt intern eine explorative Studie durchgeführt, die den Nutzen von maschinellem Lernen zur Verknüpfung der Quelldaten evaluiert hat. Darüber hinaus informiert der Beitrag über einen teilautomatisierten Ansatz, welcher Verknüpfungen nur unter der Voraussetzung von hinreichend sicheren Vorhersagen vornimmt.

Keywords: *data record linkage – register data – text similarity – supervised learning – data analysis*

ABSTRACT

The Federal Statistical Office has a legal mandate to establish a register of basic enterprise data for Germany. The content and objectives of the basic register of enterprises are defined in the underlying legal act. Its main aims are to consolidate basic enterprise data from various source registers in a central location and to ensure the timeliness and accuracy of the data. Correct and efficient linkage of the basic enterprise data from various source registers is of paramount importance in this context. For this purpose, the Federal Statistical Office conducted an internal exploratory study to evaluate the benefits of machine learning for linking source data.

Julius Weißmann
ist Data Scientist und wissenschaftlicher Mitarbeiter im Referat „Künstliche Intelligenz, Big Data“ des Statistischen Bundesamtes. Er befasst sich mit dem Einsatz von maschinellem Lernen in der amtlichen Statistik sowie der Automatisierung von Statistik- und Nichtstatistikprozessen.

Tim Herbst
ist Data Analyst und im Referat „Basisregister für Unternehmen – Fachverfahren“ des Statistischen Bundesamtes tätig. Er befasst sich mit der fachlichen Integration und Zusammenführung der Quellregisterdaten für das Basisregister.

1

Einleitung

Der Aufbau einer modernen Registerlandschaft durch übergreifende Nutzbarmachung von in Registern gespeicherten Daten ist nicht erst seit dem aktuellen Koalitionsvertrag (SPD, Bündnis 90/Die Grünen und FDP, 2021) politischer Wille in Deutschland. Ziel ist, mit einer effizienten, digitalen Verwaltung sowohl Mehrwert zu generieren als auch die digitale Handlungsfähigkeit des Staates sicherzustellen. Dies stellt eine große Herausforderung, aber auch eine große Chance für die öffentliche Verwaltung dar. Das gilt gleichermaßen für Aufbau und Inbetriebnahme des Basisregisters für Unternehmen (im Folgenden Basisregister) mit dem Statistischen Bundesamt als registerführender Behörde (§ 1 Absatz 1 bis 3 Unternehmensbasisdatenregistergesetz).

↘ **Nach § 3 Absatz 2 Unternehmensbasisdatenregistergesetz im Basisregister als Unternehmen geführte und in den Quellregistern gespeicherte Einheiten:**

1. Kaufleute im Sinne des Handelsgesetzbuchs;
2. Genossenschaften im Sinne des Genossenschaftsgesetzes;
3. Partnerschaften im Sinne des Partnerschaftsgesellschaftsgesetzes;
4. Vereine im Sinne des Bürgerlichen Gesetzbuchs;
5. wirtschaftlich Tätige im Sinne der Abgabenordnung:

   a) natürliche Personen, die wirtschaftlich tätig sind,
   b) juristische Personen und
   c) Personenvereinigungen; sowie
6. weitere Unternehmen im Sinne des Siebten Buches Sozialgesetzbuch.

Als bundeseinheitliche Wirtschaftsnummer dient dabei die Wirtschafts-Identifikationsnummer nach § 139 c der Abgabenordnung (§ 2 Absatz 1 Unternehmensbasisdatenregistergesetz), welche das Bundeszentralamt für Steuern vergibt. Für die Umsetzung haben sich moderne technologische Ansätze wie maschinelles Lernen (ML) als mögliche Katalysatoren im Bereich der Informationsverknüpfung erwiesen (Schnell, 2021). Aus diesem Grund stellte auch in der im Statistischen Bundesamt durchgeführten Vorstudie zum Basisregister für Unternehmen maschinelles Lernen einen zentralen Baustein in den Untersuchungen dar. Sie hatte zum Ziel, Erkenntnisse darüber zu erlangen, wie für die Zusammenführung der bestehenden Einzelregister automatisierte Verfahren gewinnbringend eingesetzt werden können.

Das folgende Kapitel 2 stellt das Basisregister für Unternehmen allgemein vor. Kapitel 3 beschreibt die Vorstudie zum Basisregister, informiert zu deren Datenbasis, zur Effizienz der Datensatzverknüpfung und zeigt Ähnlichkeiten in den Quellregistern auf. Wie maschinelles Lernen in der Vorstudie eingesetzt wurde, erläutert Kapitel 4 mit Erläuterungen zum Datensatz, zur maschinellen Lernstrategie, zur Datensatzverknüpfung und Deep-Learning-Ansätzen. Abschließend bewertet Kapitel 5 die Vorstudie zum Basisregister und die mit ihr gewonnenen Erkenntnisse.

2

Das Basisregister für Unternehmen

Die Einführung des Basisregisters soll erstmals den zentralen und registerübergreifenden Zugriff auf Unternehmensbasisdaten unter den gesetzlich definierten Rahmenbedingungen ermöglichen. Es birgt somit erhebliches Potenzial, um Verwaltungsprozesse effizienter zu gestalten und die Bürokratiekosten der Unternehmen zu senken. Die deutsche Registerlandschaft umfasst rund 120 einzelne Register mit Unternehmensbezug, die alle zweckgebunden und weitgehend unabhängig voneinander agieren (BMWK, 2021). Dies bedeutet Pflegeaufwand für jedes einzelne Register und führt vor allem dazu, dass Daten zwischen den Registern inkonsistent sind. Das Basisregister soll definierte Registerinformationen mit dem Ziel der „Single Source of Truth“ als konsistente Datenbasis zusammenführen. Mehrfachmeldungen der Unternehmen zur Datenaktualisierung sollen vermieden und gleichzeitig ein effizienter Datenaustausch zwischen den Registern ermöglicht werden. Das Basisregister umfasst künftig für diesen Zweck die qualitätsgesicherten Stammdaten aller Unternehmen und führt in diesem Zusammenhang die eindeutige bundeseinheitliche Wirtschaftsnummer als Identifikator ein. Dies verspricht mit Inbetriebnahme nicht nur einen effizienten Datenaustausch in der Verwaltung, sondern entlastet auch die Unternehmen selbst, da Mehrfachmeldungen bei Veränderungen entfallen.

↘ **Nach § 3 Absatz 3 Unternehmensbasisdatenregistergesetz im Basisregister gespeicherte Stammdaten:**

1. Für den Rechtsverkehr verbindliche Angabe der Firma oder des Namens entsprechend der Eintragung im Handelsregister, Partnerschaftsregister, Genossenschaftsregister oder Vereinsregister,
2. für Verwaltungszwecke aktuelle Angabe der Firma oder des Namens entsprechend der Führung im Datenbestand der öffentlichen Stelle nach § 4 Absatz 1,
3. Verwaltungsanschrift unter Angabe von Straße, Hausnummer, Postfach, Postleitzahl, Ort und Länderkennzeichen,
4. Sitz (Ort),
5. inländische Geschäftsanschrift entsprechend der Eintragung im Handelsregister, Partnerschaftsregister, Genossenschaftsregister oder Vereinsregister unter Angabe von Straße, Hausnummer, Postleitzahl, Ort und Länderkennzeichen, soweit die Pflicht zur Eintragung besteht,
6. Rechtsform und
7. Haupttätigkeit nach Klassifikation der Wirtschaftszweige.

Verwaltungsstellen können nach § 5 Unternehmensbasisdatenregistergesetz nach der Inbetriebnahme des Basisregisters auf dieses zugreifen und aktuelle Stammdaten zu Unternehmen abrufen. Angebundene Register verfügen so über aktuelle Daten hoher Qualität, wodurch Kosten durch veraltete Informationen vermieden werden können. Das Statistische Bundesamt spielt dabei als erfahrene registerführende Stelle eine entscheidende Rolle und trägt dazu bei, die Register- und Verwaltungslandschaft in Deutschland zu modernisieren.

3

Vorstudie zum Basisregister

Die Daten für das Basisregister werden für den Betrieb aus mehreren Quellen (sogenannte Quellregister) zur Verfügung gestellt. Die Quellregister – namentlich das des Bundeszentralamts für Steuern, das zentrale Unternehmerverzeichnis der Deutschen Gesetzlichen Unfallversicherung und das gemeinsame Registerportal der Länder (Justiz) – führen ihre Datenbanken nach unterschiedlichen Kriterien und verfügen jeweils über einen voneinander abweichenden Umfang an Datensätzen beziehungsweise Unternehmen. Das Basisregister wird künftig all diese Daten zusammenbringen und vereinheitlichen. Um aus den bestehenden Datenbeständen erste Schlüsse auf die zu erwartenden Dateninhalte, die Datenqualität und die Herausforderungen bei der Verknüpfung ziehen zu können, hat das Statistische Bundesamt eine Vorstudie zum Basisregister durchgeführt. Dafür haben die oben genannten Quellen für das Bundesland Hamburg Daten bereitgestellt, die für die geplante Zusammenführung untersucht wurden. Um zu aussagekräftigen Ergebnissen zu gelangen war es wichtig, bei den Untersuchungen auf Echtdaten zurückzugreifen und die Konsolidierung in einem definierten Rahmen möglichst realitätsnah zu simulieren. Hierzu wurden ausschließlich Daten juristischer Personen herangezogen. Ziel war, Einblicke in die Datenqualität und den Datenumfang der einzelnen Quellen zu erhalten und Potenziale für die automatisierte Zusammenführung der Datenbestände zu identifizieren.

3.1 Datenbasis

Eine vorbereitende Datenanalyse hat zu Beginn erste Einblicke in die Daten ermöglicht, Muster identifiziert und Hypothesen generiert. Die Datenlieferungen zur Vorstudie zeigen, dass jede Quelle über eine unterschiedliche Anzahl an Unternehmensdatensätzen in ihrer Datenbank verfügt. Gleichzeitig unterscheiden sich die zur Verfügung gestellten Datensätze quellspezifisch auch in der Definition, der Verfügbarkeit und der Formatierung der Attribute, welche für die Datensatzverknüpfung vorgesehen sind. Mit der bundeseinheitlichen Wirtschaftsnummer soll künftig ein quellenübergreifender Identifikator zur Konsolidierung geschaffen werden. Dieser existierte für die erste Zusammenführung der Daten jedoch nicht, sodass eine andere Verknüpfungsstrategie erforderlich war. Die Daten mussten anhand

Tabelle 1
Umfang der Vorstudiendaten

	Zur Verfügung stehende Datensätze
Bundeszentralamt für Steuern	347 493
Zentrales Unternehmerverzeichnis der Deutschen Gesetzlichen Unfallversicherung	151 884
Gemeinsames Registerportal der Länder (Justiz)	137 464
Insgesamt	636 841

Übersicht 1
Bewertung der Attribute für die Zuordnungsfindung hinsichtlich Güte und Verfügbarkeit

Attributgruppe	Verfügbarkeit der Attribute in Prozent der gesamten Datensätze	Einschätzung zur Verfügbarkeit und Nutzbarkeit der Attribute
Unternehmensbezeichnung	100	Immer verfügbar, je Amtsgericht und je Quellregister eindeutig, anfällig für Schreibfehler, bedingt normierbar
Adressinformationen	99	Hohe Verfügbarkeit, normierbar, für Clusterbildung geeignet
Registerstring[1]	30	Nur für juristische Personen verfügbar, dort jedoch per Definition eindeutig
Rechtsform	100	Immer verfügbar, kein einheitliches Schema, als Attribut für Clusterbildung geeignet

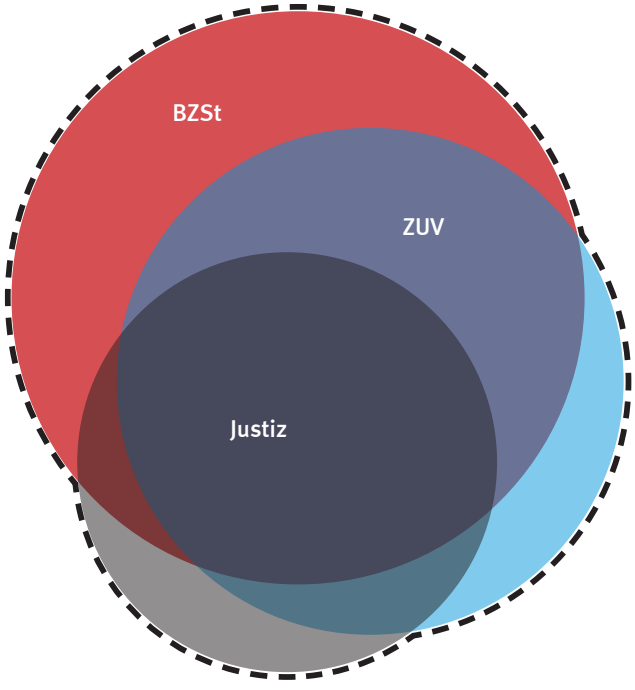
1 Der für die Vorstudie definierte Registerstring setzt sich aus dem Registergericht, an dem die Eintragung stattgefunden hat, der Registerart und der Registernummer zusammen. Der angegebene Wert bezieht sich auf alle Einheiten, jedoch sind viele Einheiten nicht eintragungspflichtig, weshalb fehlende Werte hier toleriert werden müssen. Dies ist gleichzeitig die größte Herausforderung bei der Zusammenführung, da der eindeutige Identifikator bisher fehlt und erst mit Vergabe der bundeseinheitlichen Wirtschaftsnummer (beWiNr.) eingeführt wird.

der übermittelten Attribute identifiziert und so miteinander verknüpft werden, dass ein konsolidierter Datensatz entstand, der korrekte vorhandene Informationen aus bis zu drei Quellen umfasst. Für die explorative Vorstudie wurden insgesamt 636 841 Datensätze ausgewertet, die sich auf die einzelnen Quellregister aufteilen. Dabei wurden mehrere Attribute zur Verknüpfung der Daten in Betracht gezogen und auf ihre Eignung hinsichtlich der Nutzbarkeit für ein Matching der Daten eingeschätzt. ↘ Tabelle 1

↘ Übersicht 1 zeigt, dass sich die verfügbaren Attribute unterschiedlich gut zur Verknüpfung der Datensätze zwischen den Quellregistern eignen. Hierbei sind neben der Vollständigkeit auch andere Faktoren wie Verfügbarkeit oder Eindeutigkeit entscheidend. Deshalb muss der Worst Case, das heißt die geringste Verfügbarkeit eines Attributes, angenommen werden. Außerdem sind die eingehenden Informationen nicht immer auf demselben Stand und können so zu Widersprüchen bei der Zuordnung führen. Um dieser Problematik zu begegnen, sollte die Vorstudie testen, inwiefern Algorithmen aus dem maschinellen Lernen in diesem Kontext geeignet sind, Zuordnungen hoher Güte mithilfe der verfügbaren Attribute vorzunehmen. Ziel ist, die automatische Zuordnungsgüte weiter zu erhöhen und – falls manuelle Entscheidungen zu treffen sind – die Sachbearbeitung durch passende Vorschläge zu unterstützen.

Aus der verfügbaren Datenbasis wird deutlich, dass beim Zusammenführen der übermittelten Informationen Herausforderungen für die Konsolidierung entstehen können. Kein Quellregister geht vollständig in der Menge eines anderen Quellregisters auf. Dabei kann einerseits noch nicht exakt bestimmt werden, wie groß der Abdeckungsgrad der Quellregister gegeneinander ist; dies ist schematisch in ↘ Grafik 1 dargestellt. Andererseits sind die Attribute nicht fehlerfrei und teilweise nicht übereinstimmend, sodass eine gewisse Fehlertoleranz und ein eindeutiges Regelwerk für die Verknüpfung übereinstimmender Datensätze erforderlich

Grafik 1
Schematische Darstellung des geplanten Registerumfangs als Venn-Diagramm



BZSt: Bundeszentralamt für Steuern; ZUV: Zentrales Unternehmerverzeichnis der Deutschen Gesetzlichen Unfallversicherung; Justiz: Gemeinsames Registerportal der Länder (Justiz).

sind. Sind Attribute nicht übereinstimmend, ist es Aufgabe des Basisregisters, den richtigen Eintrag je Attribut zu identifizieren und in das konsolidierte Unternehmen zu übernehmen. Hierbei soll der Automatisierungsgrad durch maschinelle Entscheidungen so groß wie möglich werden, ohne Qualitätseinbußen in Kauf zu nehmen. Das Basisregister wird nach der Aufbauphase Daten aus allen Quellregistern enthalten, sodass die Gesamtzahl an Unternehmen größer ist als in jedem Quellregister. Der äußere Rand des gesamten Venn-Diagramms stellt hierbei den Umfang des Basisregisters dar.

3.2 Effiziente Datensatzverknüpfung

Ein grundlegendes Problem bei der Datensatzverknüpfung ist die Komplexität der Daten, da eine komplette Überprüfung zur Folge hätte, dass alle Einträge unter den Quellregistern miteinander verglichen werden müssten (Christen, 2012). Gerade bei anspruchsvolleren Ähnlichkeitsberechnungen lässt sich die vollständige Überprüfung nicht mehr gewährleisten. Daher wird in der Praxis häufig über das Blocking eine Vorauswahl an Einträgen getroffen, welche überhaupt für einen Vergleich infrage kommen. Zwar lässt sich über das Blocking die Menge der Vergleiche reduzieren, allerdings führt das im Umkehrschluss zu einem Ausschluss von potenziell zusammengehörenden Einträgen. Im Kompromiss aus Genauigkeit und Geschwindigkeit wurden unter der vorhandenen IT-Infrastruktur die Postleitzahl und die Hausnummer für das Blocking verwendet. Diese liegen ausreichend fehlerfrei vor und können somit als notwendige Voraussetzung für eine effiziente Verlinkung gesehen werden. Durch das Blocking ließ sich die Anzahl der notwendigen Vergleiche von 720 385 600 auf 463 690 (0,06 %) reduzieren.

3.3 Ähnlichkeiten in den Quellregistern

Ziel der Untersuchung war, die automatisierte Verknüpfung der unterschiedlichen Quellregister zu bewerten. Generell erfolgt die Verknüpfung mehrerer Datenbestände idealerweise für jeden Eintrag über einen eindeutigen Identifikator. In den Daten aus Quellregistern gibt es jedoch keinen solchen Identifikator, weshalb gängige Ähnlichkeitsmaße herangezogen wurden (de Bruin, 2022).

Neben schnell berechenbaren Ähnlichkeitsmaßen[1] wurden für die Verlinkung weitere Ähnlichkeiten, basierend auf dem Tf-idf-Maß[2] erzeugt. In der Verarbeitung natürlicher Sprache ist das ein übliches Vorgehen, so werden Texte häufig in N-Grams[3] repräsentiert und durch das Tf-idf-Maß bewertet.

Bei der Betrachtung der Ähnlichkeitsmaße fällt auf, dass sich beispielsweise die Levenshtein-Ähnlichkeit (Yujian/Bo, 2007) zwar schnell berechnen lässt, dafür jedoch in der Schwellenwert-basierten Einordnung fehleranfällig ist. Besonders bei den nicht normierten Straßennamen fallen die Fehlzuordnungen ins Gewicht. Das lässt erahnen, dass eine geeignete Normierung von zentraler Bedeutung sein wird. ↘ Grafik 2 auf Seite 72

Ebenfalls lässt sich in Grafik 2 erkennen, dass ein größeres N-Grams zu deutlich aussagekräftigeren Ähnlichkeitsmaßen führt, was eine verlässlichere Verlinkung der Quellregister ermöglicht. Zwar scheinen besonders längere Vektoren – wie sie durch größere N-Grams entstehen – erfolgversprechend für die Verlinkung anhand der Unternehmensnamen im Basisregister zu sein, allerdings benötigen diese besonders viel Rechenspeicher. Um diesem Effekt entgegenzuwirken wurde (wie in Abschnitt 4.4 beschrieben) versucht, durch einen Autoencoder[4] eine speichereffizientere Repräsentation der Vektoren zu gewährleisten.

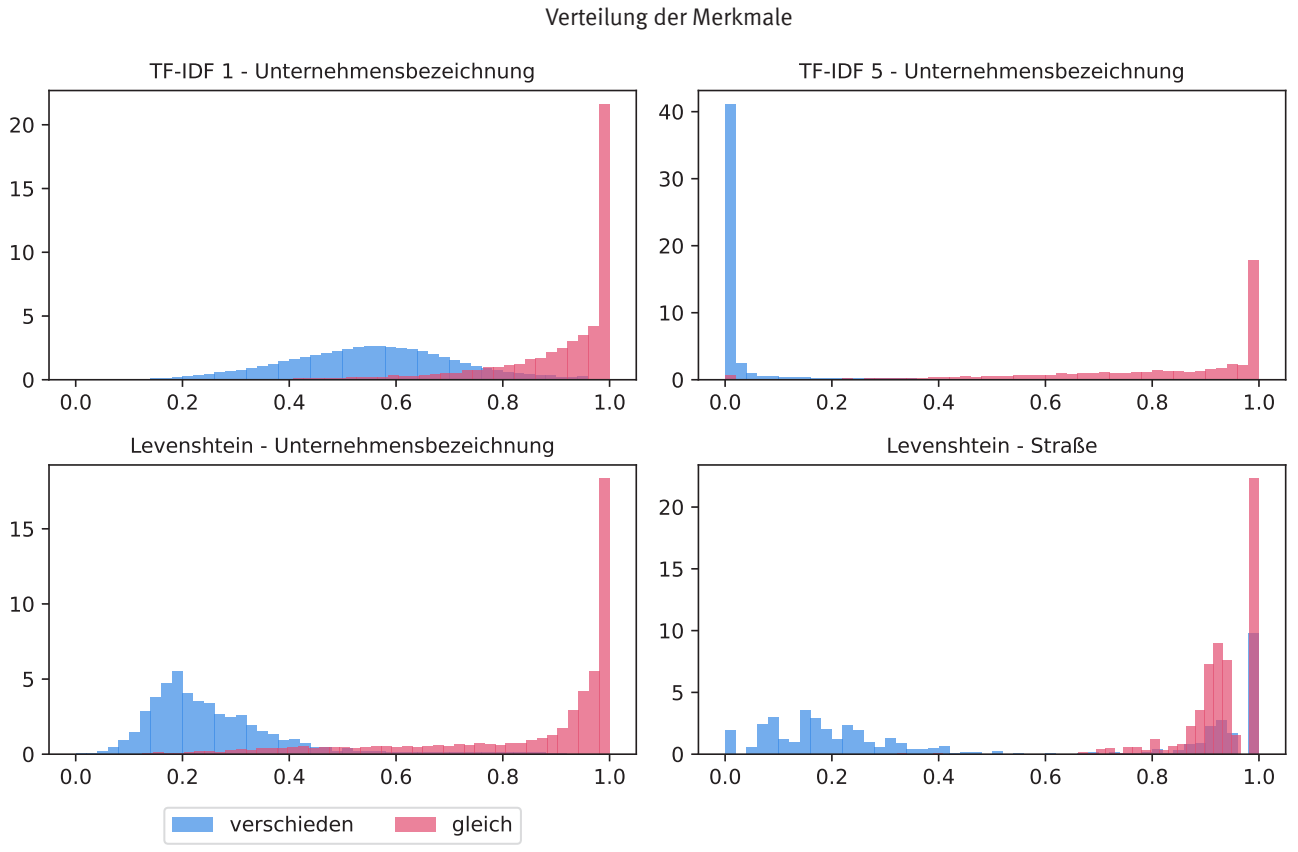
1 Eine umfassende Übersicht aller Ähnlichkeitsmaße enthält Grafik 4.

2 Tf-idf steht für Term Frequency – Inverse Document Frequency und lässt sich mit „Vorkommenshäufigkeit – inverse Dokumenthäufigkeit“ übersetzen. Dieses statistische Maß wird eingesetzt, um die Relevanz der N-Grams in Bezug auf den gesamten Datensatz zu bewerten.

3 N-Grams, auch Q-Grams genannt, sind die Fragmente (in diesem Fall Buchstaben), in welche ein Text zerlegt wurde. Bei diesem Verfahren werden N aufeinanderfolgende Buchstaben zusammengefasst. Durch die Überführung eines Wortes in N-Grams lassen sich Muster in den Zeichenketten erkennen und ähnliche zusammengesetzte Wörter identifizieren.

4 Autoencoder gehören den künstlichen neuronalen Netzen an und können wie in diesem Fall genutzt werden, um durch nicht lineare Operationen effiziente Kodierungen zu erzeugen (Bank und andere, 2020).

Grafik 2
Gegenüberstellung der Verteilung von vier Ähnlichkeitsmaßen



Tf-idf: Term Frequency – Inverse Document Frequency (Vorkommenshäufigkeit – inverse Dokumenthäufigkeit).
Idealerweise sollten die zusammengehörigen Einträge (rot) eine Säule nahe bei 100 bilden und äquivalent die ungleichen Einträge eine blaue Säule nahe bei null.

4

Einsatz von maschinellem Lernen in der Vorstudie zum Basisregister für Unternehmen

4.1 Datensatz zum maschinellen Lernen

Unter den untersuchten Ähnlichkeitsmaßen eignete sich keines einzeln, um eine verlässliche Verknüpfung der Quellregister zu gewährleisten. Aus diesem Grund wurden Experimente mit Algorithmen aus dem maschinellen Lernen durchgeführt. Beim maschinellen Lernen lassen sich aus mehreren Merkmalen, welche in diesem Fall die generierten Ähnlichkeitsmaße aus den Attributen Unternehmensbezeichnung und Straßenname sind, nicht lineare Zusammenhänge ableiten. Das trainierte Modell soll damit in der Lage sein, automatisiert verlässlichere Verknüpfungen zwischen den Quellregistern vorzunehmen, als es mit einem einfachen Schwellenwert der Fall wäre. Modelle aus dem überwachten maschinellen Lernen benötigen für das Training jedoch Labels, welche in diesem Fall dem nicht vorhandenen eindeutigen Identifikator entsprechen. Einige Einheiten enthalten durch ihre Eintragungspflicht Informationen über Registernummer, Registerart und Amtsgericht, anhand deren Kombination sich eine eindeutige Verknüpfung durchführen lässt. Unter der Annahme, dass die Attributinhalte der Einheiten mit vorhandenem Registerstring auch denen ohne Registerstring und damit der Grundgesamtheit entsprechen (siehe Übersicht 1), wurden die Informationen über die Zusammengehörigkeit durch Übereinstimmung

des Registerstrings gewährleistet und entsprechend aus den Trainingsdaten entfernt. Dieser Schritt ist für das überwachte maschinelle Lernen in diesem Fall die bestmögliche Annäherung an die in den Realdaten tatsächlich fehlenden Identifikatoren. Das Modell wird somit gezwungen, anhand der übrigen Merkmale eine Verknüpfung zu erstellen, wenngleich die Labels, also die Zusammengehörigkeit, zur Evaluierung des Modells bekannt sind. Im tatsächlichen Einsatz würden die Attribute des Registerstrings, wenn vorhanden, zusätzlich für die Datensatzverknüpfung verwendet werden, da sie eine definitorisch eindeutige Zuordnung zulassen und sehr effizient sind. Bedingt durch die Kapazität der nutzbaren IT-Infrastruktur hat sich die Vorstudie für die Untersuchungen im Bereich des maschinellen Lernens auf die Datenbestände des Bundeszentralamts für Steuern und des zentralen Unternehmerverzeichnisses der Deutschen Gesetzlichen Unfallversicherung beschränkt.

Mit diesem aufbereiteten Datenbestand soll anhand der verbleibenden Attribute, bestehend aus Unternehmensbezeichnung, Postleitzahl, Hausnummer und Straßenname, die Verknüpfung der Quellregister umgesetzt werden. Insgesamt lassen sich aus den Datenbeständen des Bundeszentralamts für Steuern und des zentralen Unternehmerverzeichnisses der Deutschen Gesetzlichen Unfallversicherung durch dieses Vorgehen jeweils 26 840 gelabelte Proben erstellen. Davon besteht für die Hälfte eine Verknüpfung durch den erzeugten Identifikator zum anderen Datenbestand.

4.2 Maschinelle Lernstrategie

Beim überwachten maschinellen Lernen werden unterschiedliche Modelle mit gelabelten Daten trainiert und anschließend auf für das Modell unbekannten Daten getestet, um eine Aussage über die Transferleistung des Modells zu ermöglichen. Für eine zuverlässige Bewertung der ML-Algorithmen wurde die Datensatzverknüpfung mit einer fünffachen Kreuzvalidierung durchgeführt und das Hyperparametertuning[5] in einer erneuten fünffachen Kreuzvalidierung verschachtelt. Für die Analysen wurden insgesamt acht unterschiedliche ML-Algorithmen zur Klassifikation herangezogen (siehe Grafik 3). Für die Hyperparameteroptimierung wurden etwa 60 Konstellationen je Klassifikator verwendet, wobei die rechenaufwendige Support Vector Maschine aus Performancegründen nur in vier Konstellationen verglichen werden konnte. Darüber hinaus wurde mit einem Oversampling[6] gearbeitet, um Verzerrungen im Modell zu reduzieren (Géron, 2019; Hasti und andere, 2009). Es wurde in einer Python-Umgebung gearbeitet und größtenteils scikit-learn als ML-Bibliothek (Pedregosa und andere, 2011) und das Python Record Linkage Toolkit (de Bruin, 2024) für die Datensatzverknüpfung verwendet.

4.3 Datensatzverknüpfung durch maschinelles Lernen

↘ Grafik 3 zeigt einen Vergleich der Ergebnisse der trainierten ML-Modelle aus der fünffachen Kreuzvalidierung. Da bei der Datensatzverknüpfung vor allem falsche Verknüpfungen vermieden werden sollen, wird der positive prädiktive Wert (PPV)[7] als entscheidendes Maß für die Evaluation herangezogen. Das Random-Forest-Modell erweist sich mit einem PPV von 0,978 am leistungsstärksten. Im Einklang mit dem hohen PPV fällt der Standardfehler beim Random-Forest-Modell mit 0,003 am geringsten aus und unterstreicht damit die Aussagekraft des Modells. Neben dem PPV wird auch der F1-Wert[8] dargestellt, um einen allgemeinen Eindruck über die Performance des Modells zu gewährleisten. Hier liegt der Random-Forest-Klassifikator mit 0,988 nur knapp hinter XGBOOST mit einem Wert von 0,989. Basierend auf PPV und dem F1-Wert kann das Random-Forest-Modell für die Datensatz-Verlinkung als geeignetes Modell betrachtet werden.

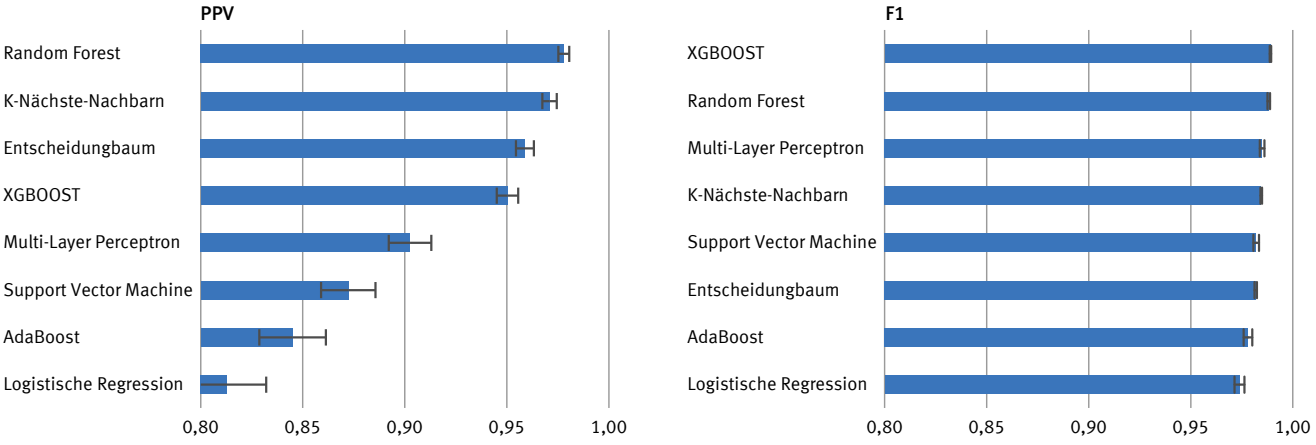
5 Das Hyperparametertuning ist entscheidend für den Erfolg eines Modells. Dabei wird das Modell mit unterschiedlichen Voreinstellungen (das heißt Hyperparametern) auf denselben Daten trainiert und die beste Hyperparameter-Konstellation für die weitere Anwendung übernommen. Da es viele mögliche Konstellationen für jedes Modell gibt, wurde hier mit einer zufälligen Suche (englisch: Random Search) gearbeitet.

6 Das Oversampling (auch Upsampling genannt) lässt sich mit überrepräsentierter Stichprobe übersetzen und wird verwendet, um Klassenungleichgewichte zu adressieren. Es hilft die Fähigkeit des Modells zur genauen Vorhersage der Minderheitsklasse zu verbessern, indem es mehr Beispiele zum Lernen zur Verfügung stellt.

7 Der positive prädiktive Wert (auch positiver Vorhersagewert; englisch: Precision oder Positive Predictive Value) ist die Wahrscheinlichkeit, dass eine Verlinkung mit positivem Testergebnis tatsächlich einer Verlinkung entspricht. Der Wert ergibt sich aus folgender Formel: PPV = richtig-positiv / (richtig-positiv + falsch-positiv)

8 Der F1-Wert ist eine Kennzahl, die sich zur Bewertung von Klassifizierungsmodellen eignet. Er berechnet sich aus dem harmonischen Mittel aus dem positiven Vorhersagewert und der Sensitivität.

Grafik 3
Modellvergleich aus fünffacher Kreuzvalidierung



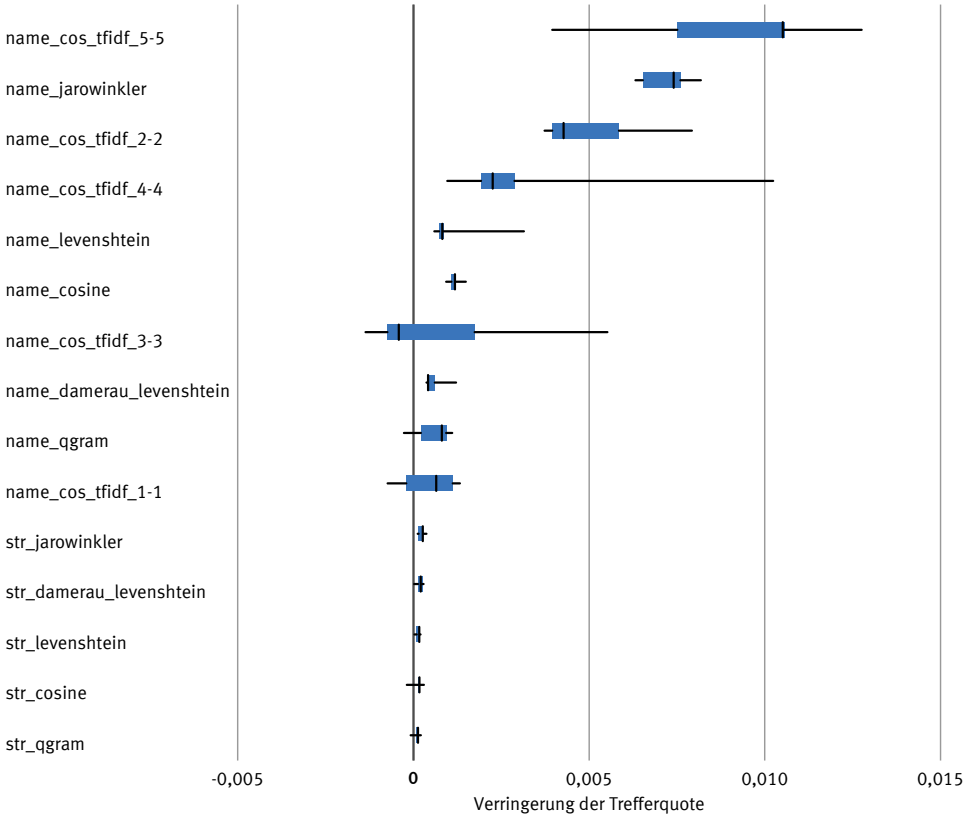
Anmerkungen: Modellvergleich durch PPV und F1-Wert mit +/-2 Standardfehler in gemittelter fünffacher Kreuzvalidierung. – Zu beachten ist, dass die Skala nicht bei 0 beginnt. – Der PPV ist das entscheidende Maß für die Modellwahl, da er die Wahrscheinlichkeit schätzt, dass eine vom Modell vermutete Verlinkung tatsächlich korrekt ist. Die Modelle mit höherem PPV weisen darüber hinaus auch geringere Standardabweichungen über die Kreuzvalidierung auf. Unter Betrachtung des F1-Wertes verändert sich insbesondere die Einordnung des Random-Forest-Modells nur geringfügig.

Anhand der Permutation Importance[9] wird die Bedeutung der Merkmale für das erfolgreichste Modell (Random-Forest-Klassifikator) veranschaulicht. Es fällt auf, dass besonders die Tf-idf-Maße eine entscheidende Rolle für den Erfolg des Modells einnehmen, wobei das 5-Gram besonders wichtig für den Erfolg ist. Der Straßenname scheint hingegen in nicht normierter Form für den Erfolg des Modells nur von vernachlässigbar geringer Bedeutung zu sein. ↘ Grafik 4

↘ Grafik 5 auf Seite 76 verdeutlicht den Mehrwert des Random-Forest-Modells für die automatisierte Verknüpfung zwischen den Quellregistern. Eine Alternative zur Verwendung des ML-Modells, welches sich mehrerer Merkmale bedient, stellt eine rein Schwellenwert-basierte Zuordnung auf Basis der Kosinus-Ähnlichkeit aus den 5-Grams dar. Es ist deutlich zu erkennen, dass die Kurve des PPV beim ML-Modell deutlich schneller ansteigt und dass diese mit einer höheren Sensitivität einhergeht, wodurch eine höhere Abdeckung der relevanten Verknüpfungen erreicht wird. Das ML-Modell verknüpft bei einem PPV von 0,95 im direkten Vergleich 95,1 % der zusammengehörigen Daten, wohingegen beim Tf-idf-Maß nur 77,1 % der Daten zusammengeführt werden. Ähnlich verhält es sich bei einem PPV von 0,99, welcher im Gegensatz zum rein Tf-idf-basierten Verfahren nicht mehr bei einer Sensitivität von 0,571, sondern bei einem Wert von 0,883 liegt. Grafik 5 verdeutlicht daher, dass sich durch maschinelles Lernen anhand der verschiedenen Merkmale mehr Daten mit einer höheren Präzision automatisiert verarbeiten lassen. Dies ist insbesondere vor dem Hintergrund bemerkenswert, dass im Basisregister Fehlzuordnungen zwingend vermieden werden sollen und somit eine hohe Präzision erforderlich ist. Anhand der PPV-Werte könnten perspektivisch verschiedene Sicherheitsniveaus innerhalb der Verarbeitung der Daten abgeleitet werden.

9 Die Permutation Importance ist eine Technik zur Bewertung der Bedeutung von Merkmalen in einem maschinellen Lernmodell. Es wird gemessen, wie stark die Trefferquote des Modells abnimmt, wenn die Werte eines bestimmten Merkmals zufällig geändert werden. Dies hilft zu verstehen, welche Merkmale am einflussreichsten für die Vorhersagen des Modells sind.

Grafik 4
Permutation Importances an den Testdaten
Fünffache Kreuzvalidierung



Anmerkungen: Die Permutation Importances (Technik zur Bewertung der Bedeutung von Merkmalen in einem maschinellen Lernmodell) wurden für die Merkmale anhand der Testdaten in fünffacher Kreuzvalidierung berechnet. Für die Unternehmensbezeichnung (name) und die Straßennamen (str) wurden verschiedene Ähnlichkeiten berechnet: qgram: q-gram, jarowinkler: Jaro-Winkler, levensthein: Levenshtein, damerau_levenshtein: Damerau-Levenshtein, cosine: Kosinus. Für die Unternehmensbezeichnung wurden außerdem die Kosinus-Ähnlichkeiten aus dem Tf-idf-Maß für N-Grams von 1 (cos_tfidf_1-1) bis 5 (cos_tfidf_5-5) berechnet.

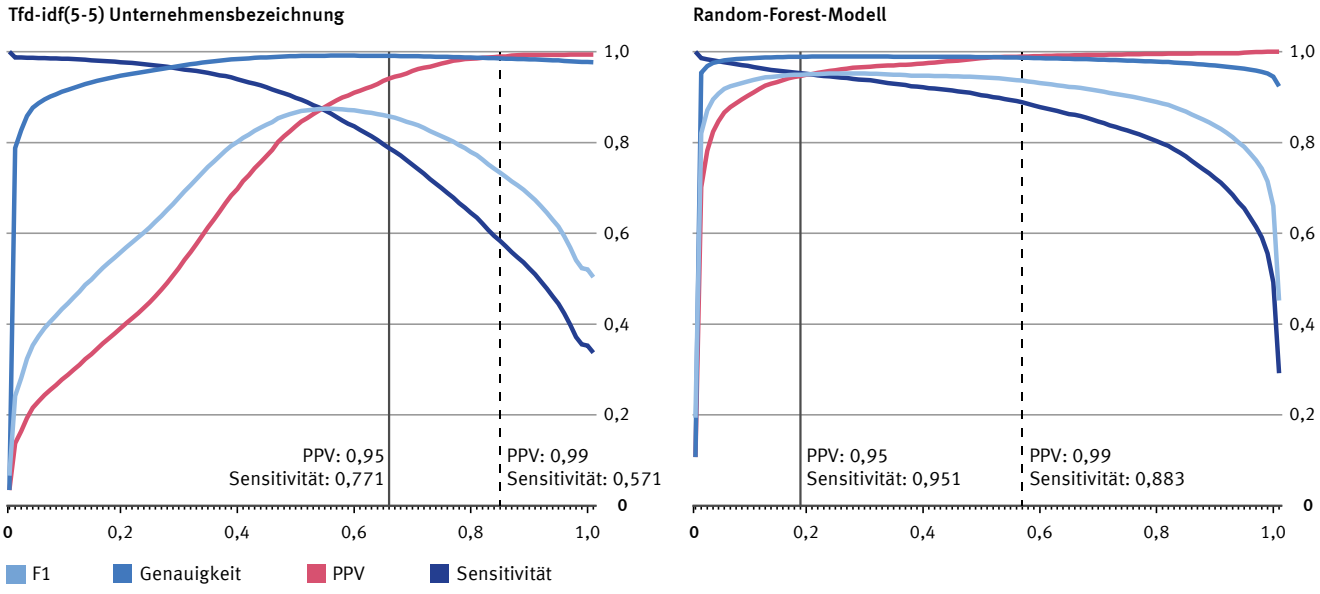
4.4 Deep-Learning-Ansätze

Als besonders erfolgversprechend erwiesen sich für das Random-Forest-Modell die Ähnlichkeitsmaße durch den Tf-idf-Algorithmus. Diese Art der Kodierung hat bereits lange Tradition in der Verarbeitung natürlicher Sprache. In den letzten zehn Jahren werden Zeichenketten jedoch zunehmend mit künstlichen neuronalen Netzen verarbeitet, da in diesem Forschungsbereich zahlreiche bahnbrechende Erfolge erzielt wurden (Li und andere, 2020; Mueller/Thyagarajan, 2024; Santos und andere, 2024; Shuangli und andere, 2019; Vaswani und andere, 2017; Xu und andere, 2022). Aufgrund der gegebenen Hardware konnten die Analysen in diesem Themenfeld nicht vollumfänglich durchgeführt werden, was insbesondere an fehlenden GPUs[10] liegt. Sie mussten daher auf einen späteren Zeitpunkt verschoben werden, wenn geeignete IT-Kapazitäten bereitstehen.

Auf Basis von künstlichen neuronalen Netzen wurden zwei Ansätze erarbeitet, anhand derer weitere Ähnlichkeitsmaße aus den Zeichenketten extrahiert werden. Im ersten Ansatz sollte die Dimensionierung der erfolgreichen N-Grams reduziert werden, damit diese in einem umfangreicheren Ausmaß mit geringerem Speicher- und Rechenaufwand verarbeitet werden können. Eine potenzielle Lösung hierfür könnten Autoencoder sein, welche den Ähnlichkeitsvektor unter Berücksichtigung

10 GPU (Graphics Processing Unit) bedeutet zu Deutsch Grafikprozessor. Durch GPUs lassen sich künstliche neuronale Netze besonders effizient verarbeiten.

Grafik 5
Schwellenwert-basierter Performancevergleich zweier Klassifikatoren



Anmerkungen: In der Gegenüberstellung wird die rein Schwellenwert-basierte Klassifikation des aussagekräftigsten Merkmals (Tf-idf(5,5) Unternehmensbezeichnung) mit dem erfolgreichsten ML-Modell (Random-Forest-Modell) verglichen. In der Gegenüberstellung ist der PPV die entscheidende Metrik.

der wesentlichen Informationen reduziert abbilden. Es wurde zwar ein kleines Netzwerk mit effizienten Parametern gewählt, allerdings ließen sich die Berechnungen in keinem zurzeit zeitlich vertretbaren Rahmen umsetzen.

Um neue Merkmale zu generieren, wurde One-Shot-Learning[11] durch einen Transformer-basierten[12] Encoder[13] in einem Metric-Learning-Ansatz[14] erprobt. Anhand dieser Netzwerkarchitektur wird das Modell optimiert, Zeichenketten über die Kosinus-Ähnlichkeit[15] richtig zuzuordnen. Aufgrund der eingeschränkten Hardware wurde hierbei erneut bewusst eine Architektur mit möglichst wenigen Parametern genutzt. In ersten Analysen an Teilstichproben ließ sich zwar eine Transferleistung des trainierten Modells erkennen, um diese Beobachtung allerdings zu bestätigen, müssten weitere Untersuchungen durchgeführt werden.

11 One-Shot-Learning ist eine Methode aus dem maschinellen Lernen. Üblicherweise wird ein ML-Modell mit möglichst vielen Beispielen aus einer Klasse trainiert. Stehen beispielsweise nur wenige oder nur ein Beispiel für eine Klasse bereit, kann One-Shot-Learning eine geeignete Lösung darstellen, um die Übereinstimmung zweier Proben zu verifizieren. Im vorliegenden Fall steht nur eine Unternehmensbezeichnung für jede Unternehmenseinheit zur Verfügung, weshalb das One-Shot-Learning als Methode gewählt wurde.

12 Transformer sind Modelle, die zu den künstlichen neuronalen Netzen gehören (Vaswani und andere, 2017). Basierend auf dieser Architektur wurden im Bereich des Neuro-Linguistischen Programmierens (NLP) populäre Modelle wie GPT, Bert, Claude, Gemini oder Mistral entwickelt (Zhao und andere, 2024).

13 Bei dem Encoder handelt es sich um einen essenziellen Block aus der Transformer-Architektur. In diesem Fall wurde der Encoder so gebaut, dass er Zeichenketten in semantisch aussagekräftige Vektoren überführt, anhand welcher Aussagen über die Ähnlichkeit der Zeichenketten getätigt werden können.

14 Das Metric Learning wird in diesem Fall verwendet, um das passende Abstandsmaß zwischen Datenpunkten zu erlernen.

15 Mit der Kosinus-Ähnlichkeit lassen sich zwei Vektoren vergleichen.

5

Fazit

Die Experimente veranschaulichen, dass trotz der teils herausfordernden Datenlage weitere Arbeiten an einer automatisierten Verknüpfung der Quellregister gewinnbringend sein können und Verfahren des maschinellen Lernens hierbei eine wichtige Rolle zukommen wird.

Um den Unsicherheiten in den Daten entgegenzuwirken, wurden die Attribute schrittweise verarbeitet. Dabei erwies es sich als geeignetes Vorgehen, zu Beginn die robusteren Attribute (Postleitzahl und Hausnummer) für das Blocking heranzuziehen und dann mit anspruchsvolleren Ähnlichkeitsmaßen in der fehleranfälligen Unternehmensbezeichnung die Zuordnung zu verfeinern. Die Straßennamen erwiesen sich hingegen in den Experimenten in nicht normierter Form als ungeeignet. Künftig könnten daher geeignete Normierungsregeln und eine entsprechende Qualitätssicherung potenziell vorteilhaft für die Verknüpfung der Straßennamen sein. Das Blocking durch die Postleitzahl und die Hausnummer führte zu einem Verlust an Verknüpfungen, dieser wurde in der Vorstudie allerdings aus Ressourcengründen hingenommen. Daher wäre es denkbar, dass sich künftig das Blocking optimieren lässt, indem die Daten vor dem Blocking bereinigt werden, und dass durch eine umfangreichere IT-Infrastruktur Verknüpfungen durch weniger Blocking durchgeführt werden können. Wie auch in Grafik 2 deutlich wurde, wäre in weiteren Untersuchungen eine Vorverarbeitung in Form einer Adressnormierung insbesondere für die Straßennamen gewinnbringend. Darüber hinaus könnten künftig Labels aus dem Echtbetrieb die Datenqualität und den Umfang der Trainingsdaten bereichern.

Die Experimente zu den unterschiedlichen Ähnlichkeiten verdeutlichen, wie gewinnbringend es ist, wenn mehrere ähnlichkeitsbasierte Merkmale und insbesondere auch 5-Grams in Verbindung mit dem Tf-idf-Algorithmus in der Entscheidungsfindung hinzugezogen werden. Anhand dieser Ähnlichkeitsmerkmale erwies sich das Random-Forest-Modell unter Berücksichtigung des PPV mit einem Wert von 0,978 als verlässlichstes ML-Modell mit gleichzeitig niedrigem Standardfehler.

Besonders interessant dürfte für den laufenden Betrieb die Einbindung von Schwellenwerten bei der Aussage des ML-Modells sein. Wie Grafik 5 zeigt, lassen sich dadurch unter der Berücksichtigung individueller Genauigkeiten entsprechend viele Verknüpfungen automatisiert verarbeiten. Mithilfe des Random-Forest-Modells lassen sich somit 95,1 % der Verknüpfungen mit einem positiven prädiktiven Wert von 0,95 vollautomatisiert auffinden.

Bei der Erzeugung der Merkmale wie auch beim Training der Modelle wurde im Hinblick auf den Bedarf einer tagesaktuellen Verarbeitung der Daten deutlich, dass die IT-Infrastruktur entscheidend für Durchsatz, Methodik und Genauigkeit ist. Unter diesen Gesichtspunkten lässt sich der Einsatz von maschinellem Lernen in der Vorstudie zum Basisregister als gewinnbringende Möglichkeit für die automatisierte Verknüpfung der Quellregister betrachten. Gleichwohl gilt es zu beachten, dass die Daten nicht registrierter Unternehmen (das sind unter anderem natürliche Personen) in der Vorstudie nur simuliert werden konnten, sodass hier weiterer Forschungsbedarf besteht.

LITERATURVERZEICHNIS

Aghamohamadi, Zhina/Rezaei Ghahroodi, Zahra. *Record Linkage with Machine Learning Methods*. In: Journal of Statistical Sciences. Jahrgang 16. Ausgabe 1/2022, Seite 1 ff. [Zugriff am 8. Mai 2024]. Verfügbar unter: jss.irstat.ir

Bank, Dor/Koenigstein, Noam/Giryes, Raja. *Autoencoders*. In: CoRR, abs/2003.05991. 2020. [Zugriff am 8. Mai 2024]. Verfügbar unter: arxiv.org

BMWK (Bundesministerium für Wirtschaft und Klimaschutz). *Gesetzentwurf zur Umsetzung des Basisregisters für Unternehmensstammdaten mit bundeseinheitlicher Wirtschaftsnummer*. 2021. [Zugriff am 16. Januar 2024]. Verfügbar unter: www.bmwk.de

Christen, Peter. *The Data Matching Process*. In: Christen, Peter. Data Matching. Data-Centric Systems and Applications. Berlin, Heidelberg 2012. DOI: 10.1007/978-3-642-31164-2_2

de Bruin, Jonathan. *Record Linkage Toolkit Documentation*. 2022. [Zugriff am 8. Mai 2024]. Verfügbar unter: buildmedia.readthedocs.org

Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Zweite Auflage. 2019. [Zugriff am 8. Mai 2024]. Verfügbar unter: www.oreilly.com

Hasti, Trevor/Tibshirani, Robert/Friedman, Jerome. *The Elements of Statistical Learning*. Zweite Auflage. 2009.

Yujian, Li/Bo, Liu. *A normalized Levenshtein distance metric*. In: IEEE transactions on pattern analysis and machine intelligence. Jahrgang 29. Ausgabe 6/2007, Seite 1091 ff. DOI: 10.1109/TPAMI.2007.1078

Li, Pengpeng/Luo, An/Liu, Jiping/Wang, Yong/Zhu, Jun/Deng, Yue/Zhang, Junjie. *Bidirectional Gated Recurrent Unit Neural Network for Chinese Address Element Segmentation*. In: isprs International Journal of Geo-Information. Ausgabe 9/2020, Seite 635 ff. DOI: 10.3390/ijgi9110635

Mueller, Jonas/Thyagarajan, Aditya. *Siamese Recurrent Architectures for Learning Sentence Similarity*. In: Proceedings of the AAAI Conference on Artificial Intelligence. Jahrgang 30. Ausgabe 1/2016. doi.org

Pedregosa, Fabian/Varoquaux, Gaël/Gramfort, Alexandre/Michel, Vincent/Thirion, Bertrand/Grisel, Olivier/Blondel, Mathieu/Prettenhofer, Peter/Weiss, Ron/Dubourg, Vincent/Vanderplas, Jake/Passos, Alexandre/Cournapeau, David/Brucher, Matthieu/Perrot, Matthieu/Duchesnay, Édouard. *Scikit-learn: Machine Learning in Python*. In: Journal of Machine Learning Research. Ausgabe 12/2011, Seite 2825 ff. [Zugriff am 13. Mai 2024]. Verfügbar unter: scikit-learn.org

LITERATURVERZEICHNIS

Santos, Rui/Murrieta-Flores, Patricia/Calado, Pável/Martins, Bruno. *Toponym matching through deep neural networks*. In: International Journal of Geographical Information Science. Jahrgang 32. Ausgabe 2/2018, Seite 324 ff.
DOI: 10.1080/13658816.2017.1390119

Schnell, Rainer. *Maschinelles Lernen für Record Linkage*. Technischer Bericht für das Statistische Bundesamt. 2021.

Shan, Shuangli/Li, Zhixu/Quiang, Yang/Liu, An/Xu, Jiajie/Chen, Zhigang. *DeepAM: Deep Semantic Address Representation for Address Matching*. In: Shao, Jie und andere. Web and Big Data. 2019. Lecture Notes in Computer Science. Ausgabe 11641. doi.org

SPD; Bündnis 90/Die Grünen und FDP. *Mehr Fortschritt wagen – Bündnis für Freiheit, Gerechtigkeit und Nachhaltigkeit*. Koalitionsvertrag 2021-2025. [Zugriff am 16. Mai 2024]. Verfügbar unter: www.bundesregierung.de

Vaswani, Ashish/Shazeer, Noam/Parmar, Niki/Uszkoreit, Jakob/Jones, Llion/Gomez, Aidan N./Kaiser, Lukasz/Polosukhin, Illia. *Attention Is All You Need*. In: CoRR, abs/1706.03762. 2017. [Zugriff am 14. Mai 2024]. Verfügbar unter: arxiv.org

Xu, Liuchang/Mao, Ruichen/Zhang, Chengkun/Wang, Yuanyuan/Zheng, Xinyu/Xue, Xingyu/Xia, Fang. *Deep Transfer Learning Model for Semantic Address Matching*. In: Applied Sciences. Ausgabe 12.19/2022. doi.org

Zhao, Wayne Xin/Zhou, Kun/Li, Junyi/Tang, Tianyi/Wang, Xiaolei /Hou, Yupeng/Min, Yingqian/Zhang, Beichen/Zhang, Junjie/Dong, Zican/Du, Yifan/Yang, Chen/Chen, Yushuo/Chen, Zhipeng/Jiang, Jinhao/Ren, Ruiyang/Li, Yifan/Tang, Xinyu/Liu, Zikang/Liu, Peiyu/Nie, Jian-Yun/Wen, Ji-Rong. *A Survey of Large Language Models*. In: CoRR, abs/2303.18223. 2023. [Zugriff am 14. Mai 2024]. Verfügbar unter: arxiv.org

RECHTSGRUNDLAGEN

Abgabenordnung in der Fassung der Bekanntmachung vom 1. Oktober 2002 (BGBl. I Seite 3866; 2003 I Seite 61), die zuletzt durch Artikel 14 des Gesetzes vom 27. März 2024 (BGBl. I Nr. 108) geändert worden ist.

Gesetz zur Errichtung und Führung eines Registers über Unternehmensbasisdaten und zur Einführung einer bundeseinheitlichen Wirtschaftsnummer für Unternehmen (Unternehmensbasisdatenregistergesetz – UBRegG) vom 9. Juli 2021 (BGBl. I Seite 2506), das zuletzt durch Artikel 1 des Gesetzes vom 22. Dezember 2023 (BGBl. I Nr. 404) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Juni 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24003-4, ISSN 1619-2907