Justus-Liebig-Universität Gießen
Fachbereich 07: Mathematik und Informatik, Physik, Geographie
Institut für Geographie
Professur für Wirtschaftsgeographie

Web Mining und Natural Language Processing als methodisches Komplement in der Wirtschaftsgeographie

Web mining and natural language processing in economic geography

Vom Fachbereich 07 der Justus-Liebig-Universität Gießen zur Erlangung des akademischen Grades Doktor der Naturwissenschaften (Dr. rer. nat.) genehmigte Dissertation

vorgelegt von
Lukas Julian Kriesch

Gießen, Januar 2023

Betreuung:

Prof. Dr. Stefan Hennemann

Professur für Wirtschaftsgeographie
Institut für Geographie, Justus-Liebig-Universität Gießen

Erstgutachten:

Prof. Dr. Stefan Hennemann

Zweitgutachten:

Prof. Dr. Christian Diller

Tag der Promotion:

15.05.2023

Zusammenfassung

Für wirtschaftsgeographische Forschung spielen räumlich und inhaltlich granular aufgelöste Daten eine zentrale Rolle, um Treiber und Barrieren sozioökonomischer Entwicklungen von Regionen besser verstehen zu können. Vor dem Hintergrund der zunehmenden Digitalisierung hat sich das Internet zu einer enorm umfassenden Datenquelle für unterschiedlichste Forschungsdisziplinen entwickelt. Insbesondere die Fähigkeit moderner Algorithmik auch unstrukturierte Textdaten semantisch auswerten zu können, ermöglicht es, enorm umfassende und gleichzeitig sehr detaillierte Informationen aus Webdaten gewinnen zu können.

In der Wirtschaftsgeographie hat eine Exploration dieser Verfahren bisher kaum stattgefunden, sodass es das übergeordnete Ziel dieser Disseration ist unstrukturierte Textdaten aus dem Internet für wirtschaftsgeographische Forschung nutzbar zu machen. Aufgrund des methodenexplorierenden Charakters der Arbeit führt diese zunächst in die Forschungsfelder Web Mining und Natural Language Processing ein, bevor die Methodiken anhand von Fallstudien konkret auf wirtschaftsgeographische Forschungsfragen projiziert werden.

Die Fallstudien skizzieren verschiedene Zugänge zu Webdaten, demonstrieren unterschiedliche Verfahren zur quantitativen Textanalyse, behandeln Texte unterschiedlicher Sprachen und umfassen sowohl Quer- als auch Längsschnittanalysen. Dabei liegt der Fokus auf der Entwicklung und Adaptierung von Modellen, die speziell im Kontext raumbezogener Forschung eingesetzt werden können. Im Rahmen der ersten Fallstudie wurde das offene Webrepositorium CommonCrawl genutzt, um eine flächendeckende, koordinatenscharfe Datenbank von Unternehmensdomains mittels Web Mining zu erstellen. Die geographische Analyse und der Vergleich mit amtlichen Statistiken zeigt, dass die extrahierten Daten in der Lage sind, die tatsächliche Unternehmenslandschaft in Deutschland zu repräsentieren. Fallstudie 2 nutzt diese Daten, um Unternehmen anhand ihrer Webseitentexte nach Technologienutzung zu klassifizieren. In der dritten Fallstudie wurde einschlägige wirtschaftsgeographische Literatur herangezogen, um abstrakte Themen in den Publikationen aufzudecken. Ferner konnten Entwicklungstrends und Zusammenhänge der Themen mittels Verfahren des Natural Language Processings quantifiziert werden.

Abschließend diskutiert die Arbeit weitere Potentiale und Herausforderungen der explorierten Methodiken. Die Diskussion beinhaltet ferner eine Gegenüberstellung der untersuchten Methodiken mit tradierten Verfahren der empirischen Sozialforschung. Aus dieser Erörterung heraus wurde ebenfalls beleuchtet, wie sich Web Mining und Natural Language Processing insbesondere in wirtschaftsgeographische Forschungsdesigns integrieren lassen und welche Perspektiven eine Methodenintegration ermöglicht.

Abstract

Spatially and contextually granular data play a central role in economic geography research in order to better understand the drivers and barriers of socio-economic developments in regions. In light of the increasing digitalisation, the internet has become an enormously comprehensive source of data for a wide range of research disciplines. In particular, the ability of modern algorithms to semantically evaluate even unstructured text data makes it possible to obtain enormously comprehensive and at the same time very detailed information from web data.

In economic geography, an exploration of these methods has hardly taken place so far, thus the overall aim of this dissertation is to make unstructured text data from the Internet more useable for economic geographic research. Due to the method-exploratory character of the thesis, it first introduces the research fields of web mining and natural language processing before projecting the methodologies onto concrete research questions in economic geography by means of case studies.

The case studies outline different approaches to web data, demonstrate different procedures for quantitative text analysis, deal with texts of different languages and include both cross-sectional and longitudinal analyses. The focus is on the development and adaptation of models that can be used specifically in the context of spatial research. In the first case study, I used the open web repository CommonCrawl to create a comprehensive, coordinate-sharp database of corporate domains by using web mining. The geographical analysis and the comparison with official statistics show that the extracted data are able to represent the actual business landscape in Germany. Case study 2 uses this data to classify companies by technology use based on their website texts. In the third case study, I used relevant economic geography literature to uncover latent themes in publications. Furthermore, I could quantify development trends and correlations of the topics using natural language processing techniques.

Finally, the thesis discusses further potentials and challenges of the explored methodologies. The discussion also includes a comparison of the explored methodologies with traditional methods of empirical social research. This discussion also shed light on how web mining and natural language processing can be integrated into research designs in economic geography in particular and which research perspectives arise from the integration of methods studied.

Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die mich während der Erstellung meiner Dissertation unterstützt haben.

Zunächst möchte ich mich bei meinem Betreuer Prof. Dr. Stefan Hennemann bedanken, ohne dessen Anregungen ich vermutlich nie mit dem Thema dieser Dissertation in Kontakt gekommen wäre. Insbesondere die Freiräume und die Ermutigung, ein Thema abseits der ausgetretenen Pfade zu bearbeiten, waren eine wichtige Motivation für mich. Unsere Gespräche auf intellektueller und persönlicher Ebene werden mir immer als bereichernder und konstruktiver Austausch in Erinnerung bleiben. Ich bedanke mich ebenfalls bei Prof. Dr. Christian Diller für die Betreuung als Zweitgutachter.

Ich danke ebenfalls meinen Kollegen Marie, Natalie, Lisa, Julia, Lisett, Moritz und Niklas für die schöne gemeinsame Zeit und den konstruktiven Austausch. Besonders dankbar bin ich Lisa, die zu einer wichtigen Freundin für mich geworden ist – egal, ob in Gießen, Bonn, Hanau, am Gardasee oder im Schwarzwald. Carsten Klaholz danke ich für technische Unterstützung, ohne dessen Engagement meine Webcrawler schnell zum Erliegen gekommen wären. Dr. Jörn Profe danke ich für die Unterstützung bei der Geokodierung.

Bedanken möchte ich mich ebenfalls bei meinen Freunden. Einerseits bei Michael, Julian und Niklas für die kurzweiligen Samstage während des Studiums. Andererseits bei meinen „Staanemer Jungs“ Huber, Thurner, Bieri, Philipp, Jan, Niklas, Daniel und Alex, die teils seit über 20 Jahren an meiner Seite sind und stets aufopferungsvoll für etwas geistige Zerstreuung und Ablenkung von wissenschaftlichen Themen sorgen. Besonders möchte ich nochmals bei Niklas bedanken, der seit der Grundschule jeden Bildungsweg mit mir gemeinsam beschritten hat und egal, ob als Schulfreund, Kommilitone, Mitbewohner, Getränkehändler oder Arbeitskollege eine wichtige Stütze für mich ist.

Tief verbunden und dankbar bin ich meiner Freundin Sarah, die mich nicht nur inhaltlich, moralisch und organisatorisch beim Verfassen dieser Arbeit unglaublich unterstützt hat, sondern auch abseits unseres gemeinsamen Arbeitszimmers immer für mich da ist. Ihr gebührt daher mein größter Dank. Danke, dass du an meiner Seite bist!

Abschließend möchte ich mich bei meinen Eltern und meiner Schwester bedanken. Danke für die enorme Unterstützung, das Vertrauen, die Geduld und das Vorleben wichtiger Werte, die mich erst in die Lage versetzt haben, diese Arbeit zu verfassen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

API	Application Programming Interface
BERT	Bidirectional Encoder Represenations from Transformers
Bi-LSTM	Bidirectional long short term memory
BOW	Bag of Words
BMBF	Bundesministerium für Bildung und Forschung
CBOW	Continious bag of words
CC	CommonCrawl
CNN	Convolutional neural network
CSS	Cascading Style Sheets
F&E	Forschung und Entwicklung
GIS	Geoinformationssysteme
Glove	Global word vectors
GLUE	General Language Understanding Evaluation
HTML	Hypertext Markup Language
IKT	Informations- und Kommunikationstechnologie
IPCC	Intergovernmental Panel on Climate Change
JSON	JavaScript Object Notation
KI	Künstliche Intelligenz
KMU	Kleine und mittlere Unternehmen
KNN	Künstliches Neuronales Netz
LDA	Latent Dirichlet Allocation
LSTM	Long Short-Term Memory
MAUP	Modifiable Areal Unit Problem
MINT	Mathematik, Informatik, Naturwissenschaften, Technik
ML	Maschinelles Lernen
NER	Named-Entity-Recognition
NLG	Natural Language Generation

NLP	Natural Language Processing
NLU	Natural Language Understanding
OSM	Open Street Map
PDF	Portable Document Format
RNN	Rekurrentes Neuronales Netz
Tanh	Tangens hyperbolicus
TF-IDF	Term Frequency-Inverse Document Frequency
TLD	Top-Level-Domain
URL	Uniform Resource Locator
WARC	Web Archive-Format
XML	Extensible Markup Language

1 Einleitung

Digitale Technologien bestimmen zu einem großem Teil das tägliche Leben moderner Gesellschaften. Digitale Informationen werden dabei größtenteils in Form von Text erstellt und über das Internet geteilt. Textdokumente werden losgelöst von ihrem Kontext nahezu ausschließlich digital publiziert und stellen sowohl im wissenschaftlichen als auch im nicht-wissenschaftlichen Diskurs eine zentrale Informationsquelle dar. Damit ist digitaler Text in der jüngeren Vergangenheit - auch für die Sozial- und Geisteswissenschaften - zu einer wertvollen Ergänzung von tradierten strukturierten Daten gereift. Der Umfang und die Geschwindigkeit, mit der digitaler Text sekündlich zumeist über das Internet publiziert wird, hat nunmehr Größenordnungen erreicht, die die menschlichen Verarbeitungskapazitäten bei weitem übersteigen. Es werden folglich computergestützte Verfahren benötigt, um die Informationsfülle digitaler Texte quantitativ strukturieren und analysieren zu können.

Für die Computerlinguistik stellt die Prozessierung natürlicher Sprache eine immense Herausforderung dar. Schließlich ist geschriebener respektive gesprochener Sprache eine enorme Dimensionalität inhärent. Buchstaben, Wörter oder Satzzeichen können nicht isoliert voneinander als statistische Merkmale betrachtet werden, sondern lassen sich erst unter Einbezug ihrer semantischen Bedeutung und des individuellen Kontexts in sinnvolle Informationen überführen. Eine sachgerechte Prozessierung von Texten muss folglich dieser Dimensionalität Rechnung tragen und ähnlich wie Menschen die Bedeutung eines Texts im Gesamtzusammenhang interpretieren. Die Verfahren zur Analyse natürlicher Sprache sind daher Wissenschaftsdisziplinen entlehnt, die vermehrt hochdimensionale Daten analysieren, wie die Computerwissenschaften, die Bioinformatik oder die Physik. In diesen Bereichen haben sich Verfahren entwickelt, die mittels performanter Algorithmik unter Einsatz von künstlicher Intelligenz (KI) und maschinellem Lernen (ML) aus immensen Datenbeständen wertvolle Informationen ableiten können. Insbesondere der Einsatz von tiefen neuronalen Netzen hat in der jüngsten Vergangenheit fächerübergreifend zu bemerkenswerte Erfolge geführt.

In der Biologie gelang es der Google-Tochter DeepMind die 3D-Struktur von über 365.000 Proteinen anhand ihrer Aminosäuresequenz vorherzusagen. Damit lösen die Autor:innen das über 50 Jahre alte Problem der Proteinfaltung (JUMPER et al. 2021). In der Meteorologie konnte mithilfe neuronaler Netze eine neue Qualität der Niederschlagsvorhersage erreicht werden. Die eingesetzten Deep-Learning-Modelle übertrafen in Vergleichsstudien systematisch die Vorhersagequalität etablierter physischer Simulationsmodelle (ESPEHOLT et al. 2021; RAVURI et al. 2021). In der Mathematik helfen neuronale Netze sowohl bei der Entdeckung neuer Theoreme und Algorithmen (DAVIES et al. 2021) als auch bei der Lösung komplexer Differentialsysteme (CHARTON et al. 2021). Auch in den Materialwissenschaften werden neuronale Netze zur Materialentdeckung eingesetzt,

beispielsweise zur Entwicklung von Katalysatoren, welche Methan effizient in Methanol umwandeln können (NANDY et al. 2022).

Die generischen Fähigkeiten neuronaler Netze, Muster und Abhängigkeiten in hochdimensionalen Massendaten erkennen zu können, eignen sich daher ebenfalls zur Analyse unstrukturierter, natürlicher Sprache. Triebkräfte und Vorreiter dieser Entwicklung sind dabei amerikanische Technologieunternehmen. Insbesondere Google Brain und Meta AI haben angesichts der gigantischen, hauseigenen Datenbestände Methoden zur informationstechnologischen Inwertsetzung dieser entwickelt. Dabei liegt seit rund zwei Dekaden der Fokus der Forschungsarbeiten auf der computergestützten Verarbeitung von natürlicher Sprache in Form von Text. Angesichts der Tatsache, dass ein Großteil der Informationen im Internet in unstrukturiertem Text gespeichert ist, offeriert die intelligente und automatische Prozessierung dessen enorme Potentiale zur Wissensgenerierung.

Für Computersysteme ist Text zunächst eine relativ informationsarme Ressource. Auf der granularsten Ebene verarbeitet der Prozessor eines Computers binäre Zahlensequenzen, sodass Text zunächst in ein für den Rechner interpretierbares Format überführt werden muss. Dieser Aufgabe widmet sich seit Beginn der 1950er Jahre das Forschungsfeld des Natural Language Processings (NLP). Die bedeutendsten Fortschritte dieser Disziplin sind dabei in der jüngeren Vergangenheit zu verzeichnen und gehen somit mit der rapiden Entwicklung der Digitalisierung - und der damit explosionsartig ansteigenden Menge von verfügbarem digitalem Text - einher. Die technischen Möglichkeiten Webmassendaten semantisch verarbeiten und analysieren zu können, ermöglichen dabei auch für Wissenschaftsdisziplinen jenseits der Computerlinguistik völlig neue Anwendungsperspektiven.

In der Geographie findet bisher keine intensive Auseinandersetzung mit diesen methodischen Neuerungen statt. Die Abstinenz moderner Verfahren der automatisierten Textanalyse im geographischen Methodenset ist vorwiegend mit den technischen und methodischen Anforderungen zu begründen, die den Einsatz von großen Textmengen durch NLP erfordern. (Human-)geographische Forschung bedient sich methodisch in der Regel den Standardverfahren der empirischen Sozialforschung, welche Text als Datenressource vor allem mit qualitativen Forschungsmethoden analysieren. Angesichts der weiter zunehmenden Relevanz des Internets als unstrukturierte Wissensdatenbank und der methodischen Erfolge der jüngeren Vergangenheit, Textdaten computerbasiert auswerten zu können, stellt die Exploration dieser Datenzugänge und Methoden ein Forschungsdesiderat dar. Übergeordnetes Ziel dieser Disseration ist es daher Verfahren und Methoden des NLP sowie des Web Minings zu verstehen und die methodischen Stärken dieser auf wirtschaftsgeographische Fragestellungen zu projizieren. Damit soll diese Arbeit einen Ausgangspunkt für weitere Explorationen der Methoden im Kontext der Geographie schaffen.

1.1 Forschungsziele und Forschungsfragen

Im Kontext raumbezogener Forschung hat eine dezidierte Auseinandersetzung respektive Anpassung moderner Text Mining-Verfahren bisher nicht stattgefunden. Da insbesondere die Geographie - unter anderem aufgrund der räumlichen Verankerung von Forschungsfragen - besondere Ansprüche an Analyseverfahren stellt, ist von einem besonderen Bedarf angepasster Methodiken auszugehen. Die Relevanz dieser Arbeit fußt somit auf zwei übergeordneten Überlegungen. Einerseits hat der Forschungszweig insbesondere in den letzten beiden Dekaden eine enorme Dynamik und Heterogenität entwickelt, sodass ein Überblick über theoretische Grundlagen, Verfahren und Begrifflichkeiten nötig ist, um eine Integration dieser Methoden in die sozialwissenschaftliche Forschungspraxis anzustoßen. Andererseits gilt es hierfür über eine reine Deskription von Verfahren und Perspektiven hinauszugehen und beispielhaft konkrete Beispiele aufzuzeigen und Datenzugänge, methodische Vorgehensweisen und Analysetechniken darzulegen.

Übergeordnetes Ziel der vorliegenden Dissertation ist vor diesem Hintergrund der Aufbau eines geordneten Methodenrahmens der computerlinguistischen Verfahren NLP und Web Mining für die Geographie. Neben der Vorstellung und Einordnung der vielfältigen Techniken zu Web Mining und NLP sollen konkrete Anwendungsbeispiele der Methodik skizziert und diskutiert werden. Diese umfassen dabei den vollständigen Forschungsprozess von Datenerhebung über Datenabruf und -analyse bis hin zu Datenauswertung und demonstrieren den Einsatz unterschiedlicher NLP-Algorithmen. Auf Basis dieser methodenexplorierenden Untersuchungen können anschließend die Integrationsmöglichkeiten der vorgestellten Methodiken in das Methodenspektrum der empirischen Sozialforschung beleuchtet werden.

Da sich das übergeordnete Ziel dieser Arbeit in Teilziele aufgliedern lässt, werden die Forschungsfragen nach Untersuchungsebenen differenziert hergeleitet. Aus der bisherigen Darstellung ergeben sich hinsichtlich der zu betrachtenden Thematik folgende Forschungsfragen:

(1) Datengrundlage

> *Wie kann das offene Webrepositorium CommonCrawl (CC) als Datengrundlage für empirische geographische Untersuchungen genutzt werden?*

Zu Beginn jedes empirischen Untersuchungsvorhabens steht die Datenerhebung. Da Sozial- und Geisteswissenschaften noch keinen systematischen Zugang bzw. Umgang mit Webseiteninhalten entwickelt haben, gilt es zu untersuchen inwiefern das CommonCrawl Repositorium als Ausgangsbasis für webbasierte wissenschaftliche Analysen genutzt werden kann (COMMONCRAWL 2022). Bisherige Forschungsarbeiten greifen auf handkuratierte Domainlisten oder kommerzielle Datenbanken zurück, um Zugang zu Webdaten zu erhalten. Daher gilt es zu klären, ob das CC als

kostenfreie und umfangreiche Datenbank eine Alternative für die Forschungspraxis darstellt. Aufgrund des enormen Umfangs von Webinhalten ist ferner zu prüfen welche Möglichkeiten der Strukturierung und deskriptiven Analyse des Webrepositoriums bestehen. Auf Basis dieser Betrachtung soll diskutiert werden, inwiefern sich das CC als Ausgangspunkt für umfassende Web Mining-Vorhaben eignet.

(2) Datenabruf

Wie können systematisch Webmassendaten für Forschungszwecke aus dem Internet abgerufen werden?

Neben der Schaffung einer umfassenden Datengrundlage in Form von Domains stellt der Abruf von Webinhalten eine zweite zentrale Herausforderung für die Web Mining-Forschung dar. Auch an dieser Stelle bedarf es Verfahren, welche den enormen Umfang von Webinhalten bewältigen können. Darüber hinaus liegen Webtexte zumeist in unstrukturierter Form vor. Auch einzelne Webseiten sind nicht a priori attributiert. Dementsprechend stellt neben dem Abruf auch die Navigation auf Webseiten und die Auswahl von Inhalten eine zentrale Herausforderung dar, die im Forschungsdiskurs bisher nur unzulänglich beleuchtet wurde.

Folglich gilt es zu klären welche Verfahren zur Generierung von Webmassendaten geeignet sind, die Funktionsweise dieser zu verstehen und sie schlussendlich in Forschungsvorhaben zu integrieren.

(3) Datenselektion und –vorverarbeitung

Wie können Unternehmensdomains identifiziert und georeferenziert werden?

Da geographische Fragestellungen im Regelfall räumlich verankert sind, ist die Geokodierung der abgerufenen Domains und deren Inhalte eine notwendige Vorarbeit für weitere Analysen. Die Wirtschaftsgeographie nimmt dabei häufig Unternehmen in den Fokus der Betrachtung, um regionale wirtschaftliche Entwicklungen anhand der Unternehmensstrukturen erklären zu können. Webdaten von Unternehmen können daher für die Wirtschaftsgeographie ein wichtiges Komplement darstellen, um sowohl inhaltlich als auch räumlich granulare Einblicke in die Unternehmenslandschaft gewinnen zu können. Die Abstinenz eines etablierten, freien Zugangs zu Webdaten für die Wissenschaft allgemein stellt somit auch für die Wirtschaftsgeographie eine Barriere dar.

Entsprechend soll ein Verfahren entwickelt werden, welches in der Lage ist systematisch Adressdaten aus Webmassendaten zu extrahieren. Mittels der Adressen sollen anschließend Unternehmensdomains aus der Grundgesamtheit aller betrachteten Domains identifiziert werden. Zur Beantwortung dieser Forschungsfrage sollen Verfahren des Web Minings zur Datengenerierung mit Verfahren des NLP zur Datenveredelung methodisch miteinander verbunden werden. Weiterhin

soll geprüft werden, ob eine Geokodierung auf Basis der extrahierten Adressdaten möglich ist und welche räumlichen Verteilungsmuster sich aus den generierten Daten ergeben.

(4) Datenanalyse

Wie kann NLP eingesetzt werden, um Unternehmenswebseiten nach Technologienutzung zu klassifizieren?

Über die Generierung eines umfassenden, geokodierten Datensatzes von Unternehmensdomains hinausgehend soll im Rahmen dieser Arbeit geprüft werden, wie moderne Textverarbeitungsalgorithmik eingesetzt werden kann, um Textdaten automatisiert auswerten zu können. Da Webtexte von einer enormen Heterogenität und Unstrukuriertheit geprägt sind, stellt die Aufbereitung der Textdaten und die Extraktion von Fließtextpassagen eine weitere Herausforderung dar. Anschließend soll ein Textklassifikationsmodell trainiert werden, welches in der Lage ist auf Basis der aufbereiteten Texte eine Technologieklassifikation vorzunehmen. Beispielhaft sollen Unternehmen identifiziert werden, die KI aktiv nutzen bzw. entwickeln. Solche Technologieklassifikationen auf Unternehmensebene sind für die Wirtschaftsgeographie von besonderem Interesse, da kaum alternative Datenquellen bestehen, die eine derart feine Klassifikation ermöglichen.

Inwiefern können mittels NLP Themen innerhalb großer Textkorpora modelliert werden?

Neben der gezielten Klassifikation von Webseiteninhalten können weitere Verfahren des NLP genutzt werden, um Sammlungen von Textdokumenten semantisch zu verarbeiten. Daher soll geprüft werden inwiefern diese Verfahren geeignet sind, um Themen in wissenschaftlicher Literatur zu identifizieren und zu ordnen. Über eine Querschnittsbetrachtung hinaus soll erörtert werden, inwiefern NLP-Verfahren zur Analyse von Diskursveränderungen innerhalb wirtschaftsgeographischer Literatur genutzt werden können.

(5) Dateninterpretation

Wie können Verfahren des Web Mining und NLP in klassische wirtschaftsgeographische Forschungsdesigns eingebunden werden?

Abschließend werden die Analyseergebnisse sowohl singulär als auch im Rahmen einer subsumierenden Synthese interpretiert und diskutiert. Die Themenbereiche NLP und Web Mining eröffnen der geographischen Forschung einerseits neue Möglichkeiten zur Informationsgenerierung und –verarbeitung. Andererseits entsteht die Herausforderung, diese neuen Methoden in die bestehende Methodik zu integrieren. Des Weiteren gilt es zu erörtern, welchen Mehrwert die im Rahmen dieser Arbeit explorierten Methoden für die wirtschaftsgeographische Forschungspraxis darstellt. Neben diesen methodischen Aspekten ist zu diskutieren, inwiefern eine quantitative

Analyse von qualitativen Textdaten erkenntnistheoretisch eine Annäherung respektive Kombination qualitativer und quantitativer Ansätze ermöglichen kann. Daher wird erörtert, inwiefern sich die computergestützte, quantitative Textanalyse von bestehenden Verfahren unterscheidet, welche Integrationsmöglichkeiten bestehen und welcher weiterer Forschungsbedarf besteht.

1.2 Aufbau der Arbeit

Nachdem im ersten Kapitel die Relevanz der Untersuchung, die Forschungsziele und Forschungsfragen dargelegt wurden, gibt Kapitel 2 einen Überblick über die Datenquelle Text sowie die Implikationen von Big Data im Forschungskontext. Darüber hinaus werden Herausforderungen und Potentiale dieser neuen Datenquellen diskutiert. Kapitel 3 stellt das Forschungsfeld des Web Minings vor. Es werden zentrale Begrifflichkeiten erläutert, der Prozess des Web Minings beschrieben sowie unterschiedliche Arten des Web Minings dargelegt. Dieses Kapitel umfassst ebenfalls einen Überblick über bestehende Forschungsarbeiten, die mittels Webdaten raumbezogene Fragestellungen adressieren. Das vierte Kapitel stellt einerseits die Entwicklung des Forschungsfelds NLP vor. Andererseits wird die grundlegende Funktionsweise der wichtigsten Verfahren zur automatischen Textverarbeitung beschrieben.

Kapitel 5 führt in den empirischen Teil der Arbeit ein und stellt das Forschungsdesign vor. Kapitel 6 beleuchtet zunächst deskriptiv die Datenquelle CC, bevor der vollständige Forschungsprozess zur Identifikation und Geokodierung von Unternehmensdomains dargelegt wird. Kapitel 6 schließt mit einer Analyse der geographischen Verteilungsmuster der identifizierten Unternehmen. Kapitel 7 baut auf dem in Kapitel 6 generierten Datensatz auf und beschreibt das methodische Vorgehen zur Textklassifikation von Unternehmenswebseiten. Dieser Datensatz wird anschließend in ein klassisches ökonemtrisches Forschungsdesign eingebunden. In Kapitel 8 werden Artikel zentraler wirtschaftsgeographischer Zeitschriften genutzt, um mittels Techniken des NLP abstrakte Themen zu identifizieren und deren semantische Verwandtschaft zu quantifizieren. Darüber hinaus werden ebenfalls mittels NLP Veränderungen in Forschungsdiskurs im Zeitverlauf analysiert.

Kapitel 9 führt die Ergebnisse der einzelnen empirischen Kapitel unter Rückbezug auf die konzeptionellen Überlegungen der ersten vier Kapitel zusammen. Hierzu werden die eigenen empirischen Ergebnisse im Rahmen des in Kapitel 3.1 erläuterten Web Mining Prozesses diskutiert. Kapitel 9 schließt mit der Diskussion der empirischen Limitationen der Arbeit. In Kapitel 10 wird die Integration der im Rahmen dieser Arbeit explorierten Methoden in die empirische Sozialforschung diskutiert. Dazu werden zunächst Unterschiede und Gemeinsamkeiten zwischen Text Mining und klassischen qualitativen sowie quantitativen Forschungsdesigns hinsichtlich erkenntnistheoretischer Konzeption, Datengrundlagen, Untersuchungsumfängen, Gütekriterien und Qua-

litätssicherung erläutert. Das zehnte Kapitel schließt mit der Vorstellung einiger Beispiele zu integrativen Forschungsdesigns. Kapitel 11 gibt unter Reflektion der Forschungsergebnisse Handlungsempfehlungen sowohl für die wirtschaftsgeographische Forschungspraxis, die geographische Methodenausbildung als auch die (Hochschul)politik. In Kapitel 12 werden die Forschungsfragen abschließend beantwortet und ein Ausblick auf weitere Forschungsperspektiven gegeben.

2 Text als Datenquelle

Text stellt seit jeher eine zentrale Wissensressource für die Forschung dar. Mit der fortschreitenden Digitalisierung und der Ausbreitung des Internets in nahezu allen Lebensbereichen werden permanent neue Daten in Form von Text generiert bzw. produziert. Dabei sorgt die „umfassende Digitalität moderner Produktions- und Kommunikationskanäle" dafür, dass Text zu einer annähernd ubiquitär verfügbaren Datenquelle wird (BIEMANN et al. 2022: 8). Neben frisch publizierten Zeitschriftenartikeln, wissenschaftlichen Publikationen, Geschäftsberichten, Social Media Posts, Webseiten oder politischen Statements werden auch zunehmend historische Texte nachdigitalisiert (WIEDEMANN 2013).

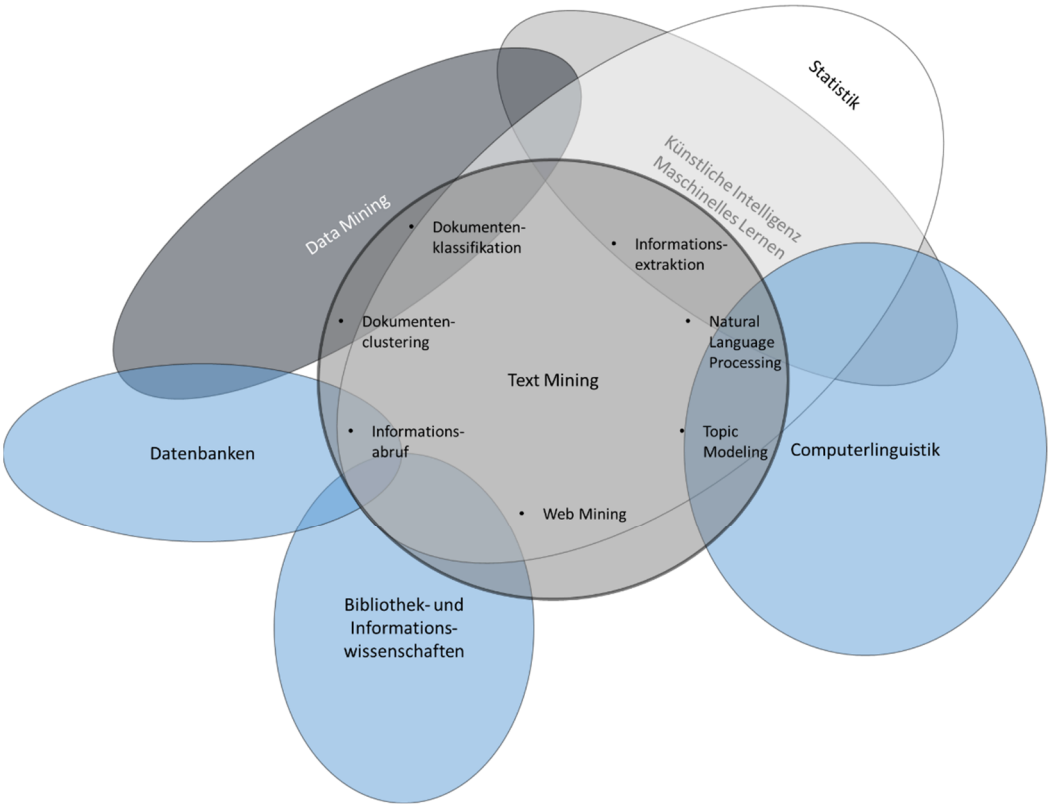
In der empirischen Sozialforschung wird Text vor allem in qualitativen Forschungsdesigns im Rahmen von Dokumentenanalysen und qualitativen Inhaltsanalysen betrachtet. Die Datenmengen, die heutzutage insbesondere im Internet verfügbar sind, übersteigen allerdings die Kapazitäten traditioneller Werkzeuge der qualitativen Sozialforschung. Zwar halten Computerprogramme auch vermehrt in die qualitative Forschung Einzug, jedoch eher zur Datenorganisation und –verwaltung (LEMKE und WIEDEMANN 2016). Dennoch können durch Computerunterstützung immer größere Datensätze verwertet und schlussendlich „[…] eine neue Stufe qualitativer Datenanalyse erreicht" werden (KUCKARTZ 2010: 13).

Rein computerbasierte Analysen sind empirischen Sozialforscher:innen daher eher aus der quantitativen Methodik zur Analyse von Sekundärstatistiken oder standardisierten Umfragen bekannt (PHILIPPS 2018). Ein markanter Unterschied zwischen numerischen Daten und Text als Datenquelle ist, dass gesprochener bzw. geschriebener Text von Natur aus hochdimensional ist. Die Interpretation eines Textdokuments ist massiv von grammatikalischen Strukturen, Interaktionen zwischen den einzelnen Wörtern sowie semantischen Bedeutungen der einzelnen Wörter abhängig (GENTZKOW et al. 2019).

Zur Analyse natürlicher Sprache werden daher statistische Verfahren aus Wissenschaftsdisziplinen angewandt, welche bereits seit Jahren mit hochdimensionalen Datensätzen arbeiten, wie die Bioinformatik, die Computerwissenschaft oder die Physik (GENTZKOW et al. 2019). Das umfassende Instrumentarium moderner Text Mining-Verfahren bietet nun die Möglichkeit digitalen Text in großen – menschliche Kapazitäten übersteigenden – Umfängen zu analysieren. Das Forschungsfeld des Text Minings geht explizit über die Stufe des reinen Datenmanagements hinaus und versucht Texte unter Einbezug des Kontexts automatisiert semantisch zu analysieren. Dabei ist Text Mining lediglich als Sammelbegriff für ein umfassendes Instrumentarium von Analyse- und Verarbeitungstechnologien zu verstehen, die zumeist aus großen unstrukturierten Textbeständen Informationen und Muster extrahieren (PUCHINGER 2016; BIEMANN et al. 2022). Wie aus Abbildung 1 hervorgeht, lassen sich mindestens sechs Forschungsdisziplinen identifizieren, die

unmittelbar das Forschungsfeld Text Mining speisen, sodass Text Mining-Forschung von einer enormen Komplexität geprägt ist.

Abbildung 1: Text Mining und Hintergrunddisziplinen.



Quelle: verändert nach MINER et al. (2012): 31.

Während der Umgang mit dem Wissensrohstoff Text aus den Bibliotheks- und Informationswissenschaften stammt, sind es Verfahren der Computerlinguistik und der KI, die eine maschinelle Verarbeitung von Textdaten ermöglichen. Da Text Mining schwerpunktmäßig auf die Analyse von Textmassendaten abzielt, spielen Verfahren des Data Minings eine zentrale Rolle im Kontext der Datenbeschaffung und des Datenmanagements in performanten Datenbanken. Den Zugang zu diesen enormen Mengen unstrukturierten Texts schafft häufig das Internet. Entsprechend spielt das Forschungsfeld des Web Minings insbesondere für Datenzugang und –abruf eine wichtige Rolle. Verfahren der Statistik und der linearen Algebra bilden für viele der genannten Disziplinen das methodische Rückgrat und sind somit als Querschnittsanforderung zu verstehen.

Abhängig von der Problemstellung kommen also algorithmische bzw. ki-basierte Anayseverfahren verschiedener Disziplinen zum Einsatz. Diese methodische Vielfalt erschwert einerseits den Aufbau eines geordneten Methodenrahmens, andererseits fehlen Sozial- und Geisteswissenschaftler:innen häufig die technischen Programmierkenntnisse, um Text Mining aktiv in der eigenen Forschung einzusetzen. Nichtsdestotrotz stehen auch Sozial- und Geisteswissenschaften aufgrund der Potentiale der neuesten methodischen Entwicklungen, vor der Herausforderung die

(teil-)automatisierte Textanalyse in den bestehenden Methodenkoffer zu integrieren. Die stärkere Verquickung von Geisteswissenschaften, Sozialwissenschaften, Informatik und Ingenieurwissenschaften wird daher auch vom deutschen Bundesministerium für Bildung und Forschung (BMBF) unter dem Titel „eHumanities" bzw. „Digital Humanities" seit über zehn Jahren in Form interdisziplinärer Forschungsprojekte gefördert (BMBF 2011, 2022). Die Projektergebnisse verdeutlichen, dass es trotz der generischen Wissensressource Text fachspezifischer Lösungen bedarf, die bestehende Verfahren anpassen oder gar neu erfinden (WIEDEMANN 2013; LEMKE und WIEDEMANN 2016).

Insbesondere für die Wissenschaft werden der Wissensressource Text durch die (semi-)automatisierte Analyse großer digitaler Textmengen sowohl inhaltlich als auch methodisch neue Eigenschaften zuteil. Diese werden seit Beginn der 2000er Jahre unter dem Terminus Big Data gefasst und werden im folgenden Kapitel erläutert.

2.1 Der Begriff Big Data

Der Begriff Big Data wurde im wissenschaftlichen Kontext erstmalig 2001 definiert, wobei die Definition 2012 nochmals angepasst wurde (BEYER und DO LANEY 2012; LANEY 2001). Während Big Data 2001 noch mit „3 Vs" (Volume, Velocity und Variety) beschrieben wurde, wurde die Definition des Begriffs 2012 um weitere „2 Vs" erweitert (Veracity und Value) (BEYER und DO LANEY 2012). Die genaue Bedeutung der „5 Vs" wird im Folgenden kurz erläutert.

Unter **Volume** verstehen die Autor:innen die absolute Größe bzw. den Umfang von Big Data. Eine Grenze, ab welcher ein Datensatz das Kriterium erfüllt, ist jedoch nicht feststehend definiert. Vielmehr beschreibt Volume die Tatsache, dass der Umfang von Big Data gängige Rechnerkapazitäten übersteigt und effiziente Methoden der Datenverarbeitung nötig sind, um diese analysieren zu können. **Velocity** beschreibt die Geschwindigkeit, mit der Big Data generiert bzw. transferiert wird. Der permanente Zuwachs der Datenbestände erfordert demnach neue Verfahren, um die Daten adäquat speichern sowie analysieren zu können und schlussendlich mit der Generierungsgeschwindigkeit neuer Daten analytisch mithalten zu können.

Variety meint die Vielfalt von Datentypen und –strukturen, in der Big Data vorliegen kann. Insbesondere im Internet finden sich unterschiedlichste Datentypen. Angefangen mit stark unstrukturierten Datenformaten - wie Video- bzw. Audiodateien - über Rohtexte natürlicher Sprache bis hin zu stark strukturierten Datenbanken oder Tabellen liegen Daten im Internet in allen erdenklichen Strukturen vor. Die Daten können dabei entweder von Menschen oder Maschinen erzeugt worden und in unterschiedlichen Formaten gespeichert sein.

Die Eigenschaft **Veracity** wurde erst nachträglich in die Definition von Big Data aufgenommen und beschreibt die Genauigkeit respektive die Korrektheit der Informationen, die Big Data beinhaltet. Speziell im Internet sind Daten gespeichert, die eine extreme Verzerrung der Sachlage implizieren können. Weiterhin können Informationen auch schlicht falsch sein, da diese meist ungefiltert bzw. ungeprüft von Menschen und Maschinen in das Internet geladen werden.

Gemeinsam mit Veracity wurde auch die Eigenschaft **Value** nachträglich in die Definition von Big Data aufgenommen. Der Wert von Big Data bezieht sich dabei in erster Linie auf die Möglichkeit räumlich, organisatorisch und zeitlich hochaufgelöste Daten in nie dagewesenen Umfängen generieren zu können. Dies trifft vor allem auf prozessgenerierte Daten, wie Mobilfunkdaten, Verkehrsdaten, Konsum- und Kreditinformationen oder Nutzungsdaten elektronischer Geräte zu (MAYERL 2015). Basierend auf diesen umfassenden Informationen können sowohl Unternehmen als auch Politiker:innen oder Wissenschaftler:innen ggf. bessere Entscheidungen treffen, sodass Big Data ein enormer Wert zugesprochen wird (BEYER und DO LANEY 2012).

Entsprechend löste der Begriff Anfang der 2010er Jahre eine nahezu euphorische Diskussion über die Potentiale der Datenflut für die Wissenschaft aus. Einige Forschende schreiben Big Data gar die Funktion zu, einen Paradigmenwechsel respektive eine quantitative Revolution in der Wissenschaft auszulösen (MILLER 2010; KITCHIN 2014; WYLY 2014). (KITCHIN 2013: 263) fasst diese Potentiale folgendermaßen zusammen:

> „Big data holds the promise of a data deluge – of rich, detailed, interrelated, timely and low-cost data – that can provide much more sophisticated, wider scale, finer grained understandings of societies and the world we live in. It offers the possibility of shifting from data-scarce to data-rich studies; static snapshots to dynamic unfoldings; coarse aggregations to high resolutions; relatively simple hypotheses and models to more complex, sophisticated simulations and theories."

Wie aus obenstehendem Zitat hervorgeht, lassen sich aus den Eigenschaften von Big Data Potentiale ableiten, die es ermöglichen, sozialwissenschaftliche Prozesse in größeren Umfängen, feinerer Granularität und größerer Detailschärfe abzubilden und zu analysieren.

Für die sozialwissenschaftliche Forschung sind insbesondere die Textdaten, welche in den umfassenden digitalen Datentöpfen enthalten sind, von Interesse. Sie stellen eine bisher kaum genutzte reichhaltige Datenressource dar, die im Zeitalter von Big Data zunehmend an Bedeutung gewinnt. Vor diesem Hintergrund hat sich das Forschungsfeld des Text Minings herauskristalliert. Die Potentiale und Herausforderungen des Text Minings für den sozialwissenschaftlichen Methodenkoffer werden im folgenden Kapitel dediziert beleuchtet.

2.2 Metho(dolog)ische Herausforderungen der empirischen Sozialforschung

Aufgrund der epistemologischen Unterschiede zwischen qualitativen und quantiativen Forschungsdesigns besteht seit den 1970er Jahren bis heute ein „Methodenstreit" in sämtlichen sozialwissenschaftlichen Disziplinen (MENNELL 1975). Erst seit den 1990er Jahren subsumieren sich unter dem Begriff der Mixed Methods unterschiedliche Ansätze mit dem gemeinsamen Ziel die jeweiligen Stärken qualitativer und quantitativer Forschungsdesigns zu verbinden. KUCKARTZ (2014) beobachtet in der jüngeren Vergangenheit eine pragmatischere, anwendungsorientiertere Methodenwahl, die qualitative und quantitative Elemente miteinander verbindet. Diese Verbindung entsteht dabei jedoch nicht organisch, da quantitative und qualitative Ansätze jeweils unterschiedliche Sachverhalte und Untersuchungsgegenstände fokussieren (KELLE 2014). Vielmehr stützen sich Mixed Methods-Ansätze auf das Prinzip der Triangulation, sodass „sich qualitative und quantitative Forschungsergebnisse im besten Fall ergänzen, jedoch nicht unbedingt entsprechen oder sich widersprechen [können]" (KELLE 2014: 157). Dennoch verspricht sich FLICK (2011: 12) durch die Verbindung methodischer Zugänge einen „prinzipiellen Erkenntniszuwachs".

Die Diskussion um bessere empirische Methoden wird auch in der Wirtschaftsgeographie geführt. BATHELT und LI (2020) diskutieren fünf methodische Herausforderungen, die neue Methoden der wirtschaftsgeographischen Forschung speziell adressieren sollen. Die Autoren fragen erstens nach neuen Methoden, die die Kluft zwischen qualitativen und quanitativen Verfahren schließen können. Sie betonen, dass insbesondere in der Wirtschaftsgeographie der Methodendualismus wichtige inhaltiche Diskussionen blockiert hat. Exemplarisch nennen sie die Schwierigkeit, die Bedeutung statistischer Zusammenhänge in quantitativen Studien zu kontextualisieren und zu verstehen. Gleichzeitig hinterfragen sie, inwiefern die Ergebnisse qualitativer Studien eine generellere Aussagekraft aufweisen, die über die Betrachtung des spezifischen Einzelfalls hinausreicht.

Zweitens fragen BATHELT und LI (2020) nach neuen Verfahren, welche die Potentiale von Big Data für die Forschungspraxis greifbar machen. Speziell in der Wirtschaftsgeographie kann eine umfassende Betrachtung individueller Entitäten (Personen, Organisationen) die Grundlage neuer Theoriekonzepte sein respektive bestehende Theorieansätze neu fundieren. Vertreter:innen der relationalen Wirtschaftsgeographie fordern bereits seit einigen Jahren eine Veränderung der Betrachtungsweise von Regionen. So kritisieren BATHELT und GLÜCKLER (2003: 121), dass Wirtschaftsgeograph:innen Regionen häufig als „handelnde Akteure" betrachten. Die Autoren fordern daher eine stärkere Fokussierung realer Akteure wie beispielsweise Unternehmen oder Organisationen. Drittens fordern BATHELT und LI (2020) strengere Analysemethoden, welche insbesondere eine stärkere Verknüpfung von aggregierten Daten mit den zugrundeliegenden Individuen erlaubt. Viertens gilt es nach BATHELT und LI (2020) epistemologisch divergierende Verständnisse von Kausalität zu klären und die Stärken und Schwächen der jeweiligen Methodik hinsichtlich des

Erkenntnisgewinns zu berücksichtigen. Fünftens messen die Autoren der Einhaltung ethischer Standards im Forschungsprozess eine zunehmende Bedeutung bei. Einerseits aufgrund von Gefahren, die durch die zunehmende Verbreitung von „Fake News" ausgehen. Andererseits können groß angelegte Studien zu Big Data datenschutzrechtliche Probleme hervorrufen. Insbesondere im Kontext von Open-Access-Publikationen, die neben Ergebnissen auch Daten und Methoden veröffentlichen sind diese Bedenken besonders zu berücksichtigen.

2.3 Potentiale von Text Mining und Big Data für die empirische Sozialforschung

Der methodologische Trend quantitative und qualitative Verfahren weniger als konkurrierende, sondern stärker als ergänzende Werkzeuge wahrzunehmen, kann durch Text Mining weiter verstärkt werden (STULPE und LEMKE 2016). Als methodisches Komplement sieht PUCHINGER (2016: 118) in Text Mining die Möglichkeit die bestehende Kluft zwischen qualitativen und quantitativen Forschungsdesigns zu schließen. Insbesondere die Fähigkeit moderner Algorithmen Text semantisch verarbeiten zu können, verringert die erkenntnistheoretische Lücke zwischen menschlichem und computerbasiertem Textverständnis (WIEDEMANN 2016: 22). Wissenschaftstheoretisch ist die Methode des Text Minings somit keinem Forschungsparadigma eindeutig zuzuordnen. Während quantitative Verfahren hypothesenprüfend eingesetzt werden und qualitative Forschungdesigns sozialkonstruktivistisch-interpretativ – also theoriebildend - vorgehen, zielt Text Mining vordergründig auf die Entdeckung latenter Strukturen in Textdaten ab (MANDERSCHEID 2019). WIEDEMANN (2013) sieht aus Perspektive der qualitativen Forschung die Möglichkeit mittels Text Mining statt einer – potentiell verzerrten – manuellen Stichprobe von weniger als 100 Fällen, statistisch repräsentative Stichproben von mehr als 1000 Fällen oder gar vollständige Textkorpora von mehr als 100.000 Einheiten zu analysieren. Für geographische Forschungsarbeiten könnte diese Verbindung aus qualitativen und quantitativen Elementen erstmals die Möglichkeit bieten räumlich fein granulare Daten in großen Umfängen zu untersuchen. Darüber hinaus bietet Text Mining die Potentiale aus einem umfassenden Datenbestand mittels automatisierter Textanalyse die relevanten Fälle zu extrahieren, um diese anschließend dezidiert mittels qualitativer Auswertungsmethodik zu beleuchten. Genauso können Text Mining-Verfahren angewandt werden, um große Datensätze inhaltlich zu strukturieren.

Somit zeigt die Anwendung von Text Mining, dass qualitative und quantitative Elemente im Forschungsprozess kombiniert werden können und somit eine weitere Annäherung der Forschungsdesigns möglich ist (SCHNAPP und BLÄTTE 2018). Einen Vorschlag zur synergetischen Verbindung von Text Mining und qualitativen Untersuchungsdesigns machen STULPE und LEMKE (2016). Sie greifen die von MORETTI (2000) eingeführte Dichotomie des *close* und *distant reading* auf. Während MORETTI unter *close reading* die intensive Lektüre eines Texts versteht, sieht er in *distant*

reading die Möglichkeit mehr Literatur in kürzerer Zeit zu überblicken und somit einen vollständigeren Überblick über den Literaturkanon zu erhalten. STULPE und LEMKE (2016) argumentieren, dass Text Mining die Aufgabe des *distant reading* durch Häufigkeitsanalysen, Kookurrenzanalysen, Themenmodellierung oder automatisierte Textklassifikation übernehmen kann. Somit können große Textkorpora effizient strukturiert und Fälle für „ein interpretierendes und kontrollierendes Lesen" im Sinne des *close readings* identifiziert werden (STULPE und LEMKE 2016: 55).

Im Kontext raumbezogener Forschung haben sich in der jüngeren Vergangenheit sechs Vorteile des Text Minings gegenüber klassischen Methoden der empirischen Sozialforschung herauskristallisiert. Diese sind in Abbildung 2 dargestellt.

Abbildung 2: Vorteile von Text Mining gegenüber klassischen Methoden.

Quelle: Eigene Darstellung.

Erstens sind Indikatoren, die mittels Text Mining erhoben wurden, nahezu in Echtzeit verfügbar und können in beliebigen Intervallen aktualisiert werden. Während klassische Sekundärdatenquellen in eher groben Zeitintervallen neue Daten erheben, sind automatisch erstellte Datensätze in der Regel zeitlich höher aufgelöst (EINAV und LEVIN 2014; MILLER und GOODCHILD 2015). Somit offerieren automatisch zeitreferenzierte Daten vollkommen neue Möglichkeiten zur Durchführung von **Längsschnitts- und Panelanalysen**. Insbesondere spontan auftretende Phänomene wie Katastrophen oder Krisen und die Reaktionen auf selbige sind mit tradierten Erhebungsinstrumenten kaum messbar. Die hohe zeitliche Auflösung gepaart mit der engen Taktung von Aktualisierungsintervallen ermöglichen es darüber hinaus, Dynamiken und Diffusionsprozesse genauer verstehen zu können.

Neben zeitlich feiner aufgelösten Analysen lassen sich durch die Arbeit mit Webtexten auch **räumlich granulare Auswertungen** vornehmen. Gerade quantitative, geographische Studien leiden aufgrund der unterschiedlichen Aggregation amtlicher Statistik unter dem „Problem der veränderbaren Gebietseinheit" (MADELIN et al. 2009). Dieses beschreibt die Anfälligkeit statistischer Auswertungen und kartographischer Darstellungen für Verzerrungen in Abhängigkeit der gewählten Maßstabsebene. OPENSHAW (1984) stellte das „modifiable areal unit problem" (MAUP) bereits in den 1980er fest. Da sehr kleinräumige Daten wie z.B. geokodierte Punktdaten nicht aggregiert sind, stellen diese eine Möglichkeit dar dem MAUP zu begegnen. Darüber hinaus ermöglicht koordinatenscharfes Datenmaterial die Durchführung kleinsträumiger Analysen, die aufgrund der grobmaßstäbigen Aggregation amtlicher Statistik nicht möglich sind. GOODCHILD (2013: 281) argumentiert, dass diese Eigenschaft für geographische Fragestellungen besonders relevant ist, da Geograph:innen häufig mit limitierten Stichproben arbeiten, die aus einer Grundgesamtheit der realen, heterogenen Welt gezogen wurden. Somit stützt sich die Geographie häufig auf die Analyse einzelner Fallstudien deren Generalisierungsfähigkeit stark limitiert ist.

Ferner können mittels Text Mining **umfangreichere Stichproben** analysiert werden als es mit klassischen empirischen Methoden möglich wäre. Insbesondere für Untersuchungen auf Mikroebene sind aus zeitlichen, finanziellen oder personellen Gründen Primärerhebungen im Rahmen klassischer qualitativer bzw. quantitativer Forschungsdesigns nur in stark begrenzten Umfängen durchführbar. Diese Kapazitätsgrenzen bestehen beim Text Mining kaum, sodass theoretisch eine Analyse von mehreren Millionen Untersuchungseinheiten möglich ist. Hieraus ergeben sich völlig neue Möglichkeiten, einerseits zur Mikrofundierung bestehender Theorien oder Modelle, andererseits zur Entwicklung neuer Analysemethoden (MUNZERT und NYHUIS 2020; KINNE und LENZ 2021).

Außerdem sind die **Kosten** für Datenzugang und Erhebung von Textmaterial über das Internet relativ gering. Während der Zugang zu professionellen Datenbanken finanziell intensiv sein kann und eigene Primärerhebungen zumindest hohe zeitliche Aufwände bedeuten, können Webtexte automatisiert und kostenfrei erhoben werden (KINNE und AXENBECK 2020).

Darüber hinaus beinhalten Massendaten völlig neue Arten von Untersuchungsvariablen. Insbesondere Daten zu **Relationen** auf Individualebene sind in gängigen Strukturdaten nicht enthalten. EINAV und LEVIN (2014: 4) schreiben diesen Datentypen eine große Bedeutung zu und argumentieren, dass diese Eigenschaft „[…] ein erstaunlicher Segen für sozialwissenschaftliche Forscher […]" sein könnte. Die Diversität neuer Untersuchungsvariablen aus Massendaten verspricht daher die Möglichkeit die beschriebenen Sachverhalte auch für die wirtschaftsgeographische Forschung greifbarer zu machen.

Weiterhin weisen die Daten keine einheitliche bzw. dokumentierte Form oder Struktur auf. Diese Eigenschaft ist einerseits eine große methodische Herausforderung. Andererseits lässt sich aus diesen Rohdaten eine Vielzahl an Merkmalen auf Individualebene ableiten. Des Weiteren können unterschiedlichste Attribute aus Textdaten gewonnen werden, während fragebogenbasierte Ansätze auf eine limitierte Zahl an Items begrenzt sind (STICH et al. 2022). Durch die hohe Dimensionalität von Textdaten können sehr detaillierte Informationen gewonnen werden, sowohl zu Einzelbeobachtungen als auch zu Zusammenhängen. Somit bieten sie die Möglichkeit sonst ungreifbare, **unscharfe Sachverhalte**, wie Denkweisen, Wissen, Einstellungen oder Wahrnehmungen zu messen, die mittels anderer quantitativer Indikatorik kaum messbar wären.

2.4 Herausforderungen der Methodenintegration

Trotz der genannten empirischen und epistemologischen Potentiale von Text Mining in der empirischen Sozialforschung gehen mit der Methodenintegration auch Herausforderungen einher. Einerseits bestehen vor allem technische und methodische Herausforderungen da Vielfalt, Umfang und Geschwindigkeit, Dynamik und (Un-)Struktur von Big Data gänzlich neue Verfahren der Datenverarbeitung und –generierung erfordern. Traditionelle statistische Methoden sind für die datenarme Wissenschaft konzipiert, um signifikante Beziehungen aus kleinen, sauberen Stichproben mit bekannten Eigenschaften zu ermitteln (KITCHIN 2013). Für die Arbeit mit unstrukturierten, fehlerbehafteten und großen Datenmengen bedarf es sowohl hinsichtlich der Datengenerierung und –haltung als auch hinsichtlich der Analyse neue Verfahren, um aus ihnen gesicherte Erkenntnisse ableiten zu können (LI et al. 2016; KITCHIN 2014).

Andererseits ist Big Data nicht allein als technischer Begriff zu deuten, dem ausschließlich mit neuen technischen Verfahren begegnet werden kann. Es stellen sich insbesondere der Wissenschaft epistemologische, (datenschutz-)rechtliche und ethische Fragen, da Big Data ohne zentrale Kontrollinstanz und meist aus verschiedenen, teilweise unbekannten und unstrukturierten Datenquellen besteht (FLORIDI 2012; EKBIA et al. 2015; BOYD und CRAWFORD 2012).

Bereits die Datenerhebung von Big Data offenbart deutliche Unterschiede im Vergleich zu tradierten Methoden der empirischen Sozialforschung. Während quantitative Daten klassisch im Rahmen eines überwachten Erhebungsprozesses von Wissenschaftseinrichtungen, staatlichen Behörden oder im Falle von Primärerhebungen von Forschenden nach fest definierten wissenschaftlichen Standards erhoben werden, ist Big Data in den seltensten Fällen ausschließlich für wissenschaftliche Zwecke generiert worden (LIU et al. 2016a). Denn viele der für die Sozialwissenschaft relevanten Datensätze sind im Besitz von großen Technologieunternehmen wie Google, Meta und Twitter oder werden von kommerziellen (Unternehmens-)datenbanken vertrieben. Die Privatisierung der digitalen Datenbestände erschwert einerseits den empirischen Zugang im For-

schungsprozess, andererseits verschränken sich Kommerzialisierungsgedanken der Inhaber:innen mit den Idealen eines reproduzierbaren und offenen Wissenschaftsdiskurses (SCHNAPP und BLÄTTE 2018). Umso wichtiger ist es für die Forschung, Mittel und Wege zu finden, um zu Big Data proprietäre Datenzugänge zu schaffen, welche kostenlos und frei zugänglich sind.

Darüber hinaus sind Textdaten auf Plattformen der sozialen Medien oder auf Webseiten im Normalfall von den betreffenden Personen oder Organisationen selbst verfasst und können daher durch den Selbstauskunftscharakter verzerrt sein. Weiterhin kann die Informationsdichte über unterschiedliche Webseiten hinweg stark variieren. Während beispielsweise manche Organisationen sehr umfassende Internetauftritte unterhalten, sind die anderer Organisationen teilweise auf eine Seite beschränkt (KINNE und AXENBECK 2020). RAMMER und ES-SADKI (2022) argumentieren, dass Unternehmen Schlüsselbegriffe eventuell nach unterschiedlichen, bzw. eigenen Definitionen auslegen könnten. Dies hat zur Folge, dass Konsistenz, Vollständigkeit und Genauigkeit der Textdaten und somit Validität, Objektivität und Reliabilität der Analyse leiden können. Darüber hinaus können die Aktualisierungsintervalle der Webseiten ebenfalls stark variieren.

Bevor Big Data analysiert werden kann, sind umfangreiche Schritte notwendig, um die unstrukturierten, aus unterschiedlichen Quellen stammenden Daten in ein verarbeitbares Format, zu überführen. BOLLIER und CHARLES (2010) hinterfragen daher, inwiefern Big Data eine Objektvität zugesprochen werden kann oder ob die Daten bereits durch den Bereinigungsprozess verzerrt werden. Gleichzeitig beinhalten insbesondere Textdaten irrelevante bzw. inhaltsleere Passagen, sodass die Aufbereitung und Filterung eine notwendige Bedingung für eine fundierte Analye darstellt (DENNY und SPIRLING 2018; RÜDIGER et al. 2017).

Ferner sind auch hinsichtlich der Repräsentativität von Webmassendaten fragestellungsabhängige Limitationen existent. Der Begriff Big Data zielt unter anderem darauf ab, dass durch die Masse an Daten eine nahezu flächendeckende Abdeckung der zugrundeliegenden Population erreicht werden kann. Allerdings sorgt die reine Masse an analysierten Daten nicht notwendigerweise für eine höhere Repräsentativität oder eine bessere Datenqualität (BOLLIER und CHARLES 2010; CHESHIRE und BATTY 2012). Im Falle von Webdaten verbessert sich die Repräsentativität stetig. So unterhalten im Jahr 2021 78 % der Unternehmen in der europäischen Union (EU) eine eigene Webseite, während es zehn Jahre zuvor noch 68 % waren. Gleiches gilt für den Internetzugang von Haushalten in der EU. Während 2011 lediglich 72 % der Haushalte einen Internetzugang hatten, sind es im Jahr 2021 bereits 92 % (EUROSTAT 2022). Nichtsdestotrotz sind auch Webdaten per se nicht probabilistisch, sondern müssen im Vorfeld der Analyse hinsichtlich ihrer Selektivität geprüft werden (BERĘSEWICZ et al. 2018). Außerdem bilden Webdaten - trotz der zunehmenden Verlagerung sämtlicher Aktivitäten ins Internet - in der Regel nur einen Teil der Gesamtpopulation ab. Daher ist fragestellungsabhängig zu prüfen, inwiefern aus der untersuchten Population

generalisierbare Erkenntnisse ableitbar sind und welche Implikationen bei der Interpretation zu berücksichtigen sind.

Letztlich gilt es insbesondere bei der Verwendung umfassender Web Mining-Techniken die rechtlichen Rahmenbedingungen zu berücksichtigen. Zwar kommt ein Rechtsgutachten von VOGEL und HILGENDORF (2019) zu dem Entschluss, dass Web Mining für wissenschaftliche Zwecke ein legales Erhebungswerkzeug darstellt. Allerdings sind einige Voraussetzungen im Vorfeld des Forschungsvorhabens zu prüfen. So ist im Urheberrechtsgesetz explizit für das Web Scraping im Rahmen wissenschaftlicher Untersuchungen eine Schrankenregelung verankert, die Text und Data Mining erlaubt. Forschende dürfen demnach Webdokumente für wissenschaftliche, nicht-kommerzielle Zwecke vervielfältigen, beispielsweise zur Generierung von Textkorpora. Allerdings müssen die Daten nach Abschluss der Forschungsarbeiten gelöscht bzw. in einer Bibliothek oder einem Archiv gespeichert werden. Ferner dürfen die Daten lediglich innerhalb eines abgegrenzten Personenkreises zur gemeinsamen Forschung geteilt und nur zur Qualitätskontrolle übergeben werden (VOGEL und HILGENDORF 2019).

Zusammenfassend lässt sich festhalten, dass die Integration von Web Mining in den Methodenkoffer der empirischen Sozialforschung gleichermaßen eine lohnende und fordernde Aufgabe darstellt. Zwar besteht seit über einer Dekade ein Forschungsdiskurs über die Relevanz von Big Data für die Forschung. Allerdings ist der Forschungskanon zu methodischen Standards und Prinzipien bisher kaum ausdifferenziert, sodass es einer umfassenderen und tiefgreifenderen Auseinandersetzung mit der Methode bedarf, bevor diese standardmäßig eingesetzt werden kann. Die Exploration dieser Methodik schneidet dabei einige, der Geographie unbekannte, Forschungsbereiche an. Einerseits stellen Rohdaten aus dem Internet eine bisher kaum genutzte Datenressource dar, sodass sowohl konzeptionelle Überlegungen als auch technische Details geklärt werden müssen, bevor eine empirische Nutzung dieser erfolgen kann. Andererseits sind Verfahren der automatisierten Textanalyse großer Textmengen ebenfalls nicht im geographischen Methodenrepertoire verankert, sodass auch hier eine methodische Auseinandersetzung der empirischen vorgeschaltet werden muss. Darüber hinaus sind insbesondere aus geographischer Perspektive mit der Analyse von räumlich verorteten Textdaten erkenntnistheoretische Implikationen verbunden, deren (Integrations-)Bedingungen ebenfalls ein Forschungsdesiderat darstellen.

3 Web Mining

Da das Internet nicht als komfortabel abrufbare Datenbank entworfen wurde, stellt die Informationssammlung und –verarbeitung eine wesentliche Herausforderung dar. Gleichzeitig dient das Internet für moderne Text Mining-Vorhaben als zentrale Ressource von digitalem Text. In den vergangenen beiden Dekaden hat sich das Forschungsfeld des Web Minings aus dem übergeordneten Fachgebiet des Data Minings herausgebildet und weiterentwickelt. Web Mining lässt sich dementsprechend definieren als: "[…] the use of data mining techniques to automatically discover and extract information from Web documents and services" (KOSALA und BLOCKEEL 2000: 2). Eine Webseite ist über einen sogenannten Uniform Resource Locator (URL) erreichbar. Häufig teilen sich mehrere Webseiten den gleichen URL-Stamm, welcher auch Domain genannt wird. Der Domainname wird durch einen Punkt von der Domainendung bzw. der sogenannten Top-Level-Domain (TLD) getrennt. Web Mining stellt somit einen Teilbereich des Text Mining dar, der explizit Textdaten aus dem Internet für fortführende Analysen nutzt.

Web Mining beinhaltet einen kompletten Prozess, der relevante Webinhalte identifiziert, diese aufbereitet und anschließend analysiert, um die zumeist unstrukturierten Daten in sinnvolle Informationen zu übersetzen (vgl. Abbildung 2). Damit besteht zwischen Web Mining und Web Scraping definitorisch eine gewisse Schnittmenge. Web Scraping beschreibt den automatisierten Datenabruf von Informationen aus dem Internet. Die Analyse und Verarbeitung der extrahierten Inhalte sind kein Teil des Web Scrapings, da Web Scraping häufig zur Sammlung strukturierter Daten eingesetzt wird, die keine gesonderte Veredelung benötigen (KÜHNEMANN 2021). Web Scraping lässt sich nach STATEVA et al. (2018) nochmals in zwei Subkategorien untergliedern.

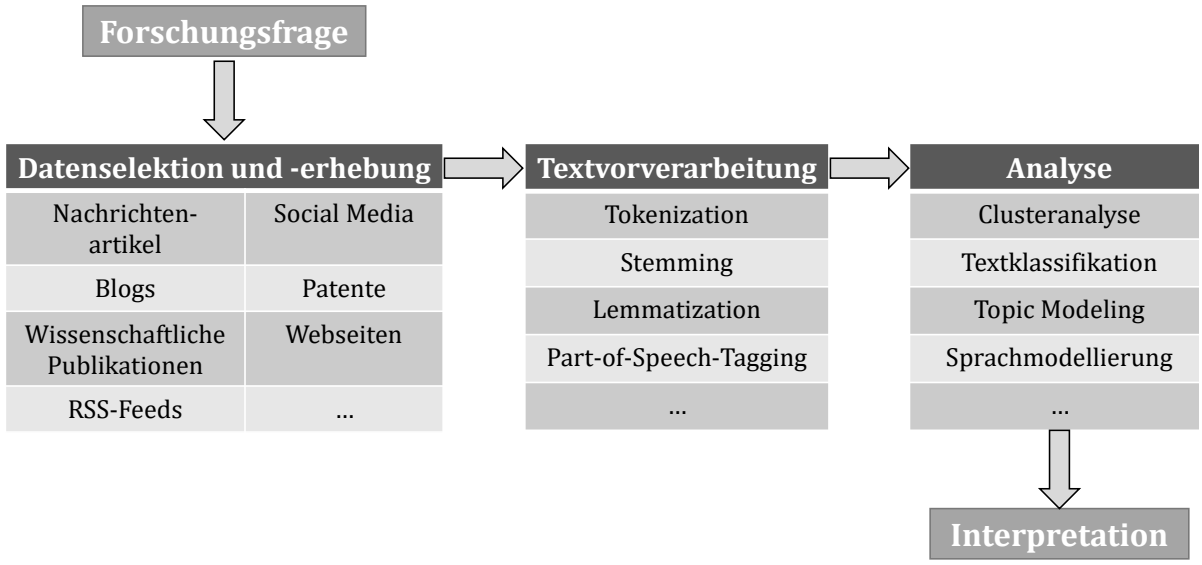
Spezifisches Web Scraping kommt zum Einsatz, wenn sowohl Inhalt als auch Struktur der Webseite vollständig bekannt sind. In diesem Fall müssen Web Scraping-Programme lediglich die gewünschten Informationen der Seite extrahieren und abspeichern. Ein typisches Beispiel für *spezifisches Web Scraping* ist der Abruf von Preisinformationen von Produktwebseiten. Sind Struktur und Inhalt der Webseiten einmal bekannt, kann der Prozess beliebig oft für beliebig viele Produkte wiederholt werden. Sind Aufbau und Inhalt der zugrundeliegenden Webseite unbekannt, da diese beispielsweise im Rahmen eines Crawlingprozesses besucht wurde, sprechen STATEVA et al. (2018) von *generischem Web Scraping*. In diesem Fall können die Webseiteninformationen nicht systematisch verarbeitet werden, sondern müssen zuerst einen Veredelungsprozess durchlaufen. Zur effizienten Prozessierung unbekannter Webseiteninhalte kommen mittlerweile vermehrt überwachte und unüberwachte Verfahren des ML zum Einsatz, um bereits während des Crawlings Inhalte filtern zu können (UZUN 2020; VOGELS et al. 2018; KENEKAYORO et al. 2014; MAGHDID 2019).

Ein weiterer, in diesem Kontext abzugrenzender, Begriff ist das Web Crawling. Web Crawling bezeichnet die automatische Sammlung und Indexierung von Webseiten. Dies geschieht über das systematische Verfolgen von Webseitenverlinkungen. Web Crawling wird im großen Stil von Suchmaschinen angewandt, um Suchbegriffe mit Webseiten verknüpfen zu können (XU et al. 2011). Web Crawling kommt jedoch auch im Forschungskontext zum Einsatz, um weitere Webseiten zu sammeln, die im Anschluss per Web Scraping abgerufen werden sollen (MITCHELL 2018).

3.1 Der Web Mining Prozess

Der Prozess des Web Minings lässt sich dabei nach KOSALA und BLOCKEEL (2000) in vier übergeordnete Aufgabenbereiche gliedern. Abbildung 3 gibt einen Überblick über den Prozess. Die einzelnen Prozessabfolgen werden im folgenden Teil der Arbeit detailliert erläutert.

Abbildung 3: Web Mining Prozess.



Quelle: Eigene Abbildung nach KAYSER und BLIND (2017: 210).

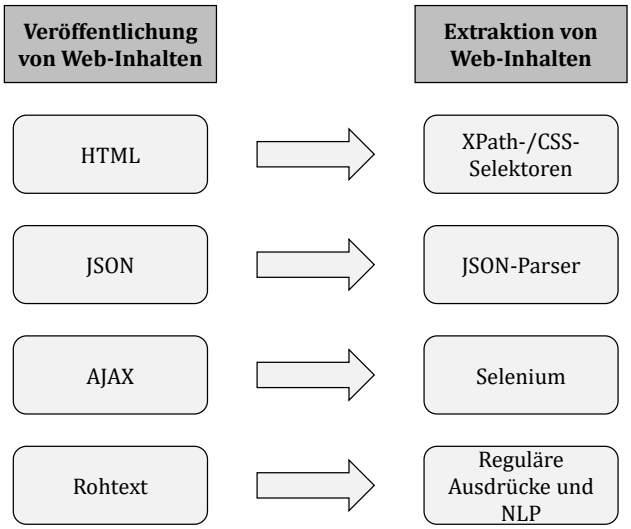
3.1.1 Datenselektion und -erhebung

Die Datenbeschaffung stellt den ersten Arbeitsschritt eines Web Mining-Vorhabens dar. Übergeordnet kann die Datenbeschaffung auf drei verschiedene Arten erfolgen. Falls bereits zu Beginn des Vorhabens eine vollständige Liste relevanter Webseiten vorliegt, können diese direkt mittels generischem bzw. spezifischem Scraping abgerufen werden. Deutlich komplexer wird das Vorhaben, wenn diese Liste a priori nicht bekannt ist, sodass dem Scraping ein Crawlingprozess vorgeschaltet werden muss. Beispielsweise können Suchmaschinen genutzt werden, um durch die Suche nach einem Unternehmensnamen die zugehörige Webseite zu erhalten (BARCAROLI et al. 2016). Eine dritte Variante der Datenbeschaffung stellen Web-APIs (englisch: application programming interface; MITCHELL (2018)) dar.

Web-APIs sind Programmierschnittstellen, die von Serverbetreiber:innen betrieben werden können. Sie erlauben einen strukturierten, dokumentierten und automatisierten Zugriff auf Datenbanken, die von Webseitenbetreiber:innen bereit gestellt werden (FULLER 2008).

Darüber hinaus können Daten auf Webseiten in unterschiedlichen Formaten veröffentlicht werden, welche unterschiedliche Verfahren erfordern, um die Daten abzugreifen. Abbildung 4 gibt einen Überblick über unterschiedliche Veröffentlichungsformate von Webinhalten und jeweils passende Extraktionsverfahren.

Abbildung 4: Technologien zur Veröffentlichung und Extraktion von Webdaten.



Quelle: Eigene Abbildung nach MUNZERT und NYHUIS (2020: 382).

Web-APIs tauschen Daten in der Regel über das JSON-Format (JavaScript Object Notation) bzw. für Geodaten über das GeoJSON-Format aus. Für das Parsing – also das Auslesen und Weiterverarbeiten – von JSON-Daten stehen in unterschiedlichen Programmiersprachen verschiedene Softwarepakete zur Verfügung. Im Falle des spezifischen Web Scrapings kann die Syntax des HTML-Codes (englisch: Hypertext Markup Language; (WOLF 2021)) genutzt werden, um mit HTML-Parsern bestimmte Inhalte (z.B. Tabellen, Überschriften oder Textbausteine) einer statischen Webseite auszulesen.

Dynamische Webseiten verändern hingegen ihre Inhalte in Abhängigkeit der Interaktion der Nutzenden mit der Webseite. Daher kann in diesem Fall nicht der statische HTML-Code verwendet werden, um die dynamisch generierten Daten auszulesen. Diese dynamischen Daten werden im Regelfall über die AJAX-Technologie (Asynchronous JavaScript and XML) eingespielt. Das Selenium-Framework imitiert das reale Browsingverhalten eines Menschen und ist somit in der Lage mit diesen dynamischen Webseiten zu interagieren (MUTHUKADAN 2022). Daher kann das Selenium-Framework auch verwendet werden, um automatisiert Texte einzugeben, Formulare auszufüllen, Elemente anzuklicken oder über die Webseite zu scrollen. Letztlich kann die gezielte

Ansteuerung von Textinhalten auch durch die Verwendung regulärer Ausdrücke erreicht werden. Somit können Muster und Regelmäßigkeiten der natürlichen Sprache in Form von Programmcodes definiert und anschließend extrahiert werden.

3.1.2 Informationsselektion und –vorverarbeitung

Anschließend müssen auf den identifizierten Webseiten die gewünschten Inhalte selektiert und vorverarbeitet werden, um die gewonnenen Daten in Informationen zu verarbeiten. Die Informationsveredelung unterscheidet sich zwischen generischem und spezifischem Web Scraping erneut stark. In der Regel wird beim Scraping zunächst der vollständige HTML-Code abgespeichert. Im HTML-Code einer Webseite liegen sämtliche Informationen in strukturierter Form vor, die durch sogenannte HTML-Tags sehr präzise ausgewählt werden können. Diese Struktur wird beim spezifischen Web Scraping genutzt, um die Inhalte einzelner HTML-Tags exakt auslesen zu können.
Beim generischen Web Scraping wird aus dem HTML-Code zunächst der Rohtext extrahiert und anschließend strukturiert. Dies geschieht in der Regel automatisiert durch Anwendung von Techniken des NLP. Die Textvorverarbeitung ist ein extrem erfolgskritischer Faktor bei Web Mining-Projekten (UYSAL und GUNAL 2014; VIJAYARANI MOHAN 2015; RÜDIGER et al. 2017). Gleichzeitig hängt die Wahl der Textaufbereitungsmethoden enorm von Art und Quelle der Texte sowie den Analysezielen des Vorhabens ab. Gängige Datenvorverarbeitungsmethoden sind dabei:

a. Tokenization:

   Viele NLP-Anwendungen betrachten als kleinste Untersuchungsentität die Wörter eines Satzes. Entsprechend wird Tokenization genutzt, um einzelne Wörter innerhalb eines Satzes zu identifizieren. Dieser Schritt gilt dabei als Grundlage für sämtliche aufsetzende Analysen (WEBSTER und KIT 1992; MANNING et al. 2014).

b. Stemming/Lemmatization:

   Stemming und Lemmatization werden eingesetzt, um die unterschiedlichen morphologischen Variationen eines Wortes zu reduzieren. Während Stemming durch das Abschneiden von Wortendungen Flexionen entfernt, führt Lemmatization die flektierten Wörter auf ihre Grundform zurück (JIVANI 2011). Der zentrale Unterschied zwischen Lemmatization und Stemming besteht darin, dass die lexikalischen Kategorien der Wörter bei der Lemmatization erhalten bleiben (BALAKRISHNAN und LLOYD-YEMOH 2014).

c. Entfernung von Stopwörtern:

   Stopwörter werden aus Textdokumenten entfernt, um den Vokabularumfang des Textkorpus‘ zu reduzieren. Die am häufigsten auftretenden Wörter in Texten sind Artikel, Präpositionen oder Pronomen, die keine oder nur eine untergeordnete Rolle für die Bedeutung eines Satzes haben. Eine Entfernung dieser sorgt entsprechend für eine deutliche Reduktion der Dimensionalität eines Textdokuments ohne dabei die inhaltliche Aussage des Textdokuments zu verändern (VIJAYARANI et al. 2015).

d. Part-of-speech-Tagging:

Part-of-speech-Tagging wird eingesetzt, um den Wörtern eines Textdokuments ihre jeweilige Wortart zuzuordnen (SCHMID 1999). Damit erfüllt Part-of-speech-Tagging eine Disambiguierungsaufgabe. Abhängig von ihrer Wortart können gleiche Wörter unterschiedliche Bedeutungen haben (z.B.: „Räumen Sie den Saal."; „In den Räumen brennt Licht"). Durch die Zuordnung von Wortarten durch Einbezug des Kontexts können diese Mehrdeutigkeiten aufgelöst werden (JURAFSKY und MARTIN 2019). Für das Part-of-speech-Tagging werden unterschiedliche Verfahren eingesetzt, z.B. rekurrente neuronale Netze oder Transformermodelle (vgl. Kapitel 4). Auf Basis der zugeordneten Wortarten können anschließend Textdokumente effizient gefiltert werden.

e. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF ist eine deskriptive Statistik, anhand welcher Gewichte für jedes Wort eines Textkorpus' berechnet werden können. Der erste Teil des Maßes beschreibt die Häufigkeit des Auftretens eines Wortes innerhalb eines Textdokuments (Term-Frequency). Der zweite Teil ist die inverse Häufigkeit des Wortes innerhalb des gesamten Textkorpus'. Durch die Multiplikation der beiden Statistiken entsteht ein harmonisiertes Maß zur Gewichtung von Wörtern in Textdokumenten. Auf Basis der TF-IDF-Statistik können ebenfalls Stopwörter identifiziert bzw. besonders charakteristische Wörter eines Textdokuments hervorgehoben werden. Damit ist die Berechnung und Filterung des TF-IDF eine gängige Vorverarbeitungsmethode für nachgelagerte Analysen wie beispielsweise Textklassifikation oder Textzusammenfassung (VIJAYARANI et al. 2015).

f. N-Gramme:

N-Gramme oder Mehrworteinheiten sind nach BUBENHOFER (2017: 70) „Paare von Worteinheiten (auf der Basis von Wortformen, Lemmata oder anderen sprachlichen Einheiten), die innerhalb einer bestimmten Distanz zueinander kookkurrieren und eine statistisch feststellbare Bindung zueinander aufweisen". Die Bindung der N-Gramme wird mittels statistischer Signifikanztests geprüft (MIKOLOV et al. 2013b; BOUMA 2009). Signifikant kookkurierende Wörter werden anschließend zu einem feststehenden Begriff zusammengefasst. Diese können theoretisch beliebig viele Einzelwörter umfassen. N-Gramme, die aus zwei Einzelwörtern bestehen, werden auch Bigramme genannt, während Wortpaare aus drei Einzelwörtern auch als Trigramme beschrieben werden.

3.1.3 Generalisierung

Die abgeleiteten Informationen werden genutzt, um Muster in den Daten zu erkennen. Dies kann sowohl auf einzelnen Webseiten geschehen, als auch über den gesamten Datensatz hinweg. Hierbei werden häufig Verfahren des ML und des NLP genutzt (vgl. Kapitel 4). Zielabhängig existieren

mittlerweile viele unterschiedliche Generalisierungsmethoden. Die zentralsten Verfahren sind unter anderem:

a. Sprachmodellierung

Das Training großer Sprachmodelle auf Basis von Webseitentexten ist eine grundlegende Methode des NLP, da auf vortrainierten Sprachmodellen viele Folgeanwendungen wie z.B. Textklassifikation oder Textgenerierung aufsetzen. Sprachmodellierung verfolgt demnach das Ziel einem Computersystem ein grundlegendes Verständnis der Sprache des zugrundeliegenden Trainingskorpus‘ beizubringen. Während des Trainingsprozesses werden einzelne Wörter der Trainingsdaten maskiert und das Sprachmodell versucht jeweils das fehlende Wort vorherzusagen (DEVLIN et al. 2019; JOZEFOWICZ et al. 2016a; LIU et al. 2019). Detaillierte Einblicke in Verfahren der Sprachmodellierung gibt das nächste Kapitel dieser Arbeit.

b. Textübersetzung

Die automatisierte Textübersetzung nimmt ebenfalls eine wichtige Rolle im Bereich des Textminings ein. Während frühere Ansätze statistische Verfahren zur Textübersetzung nutzen (LOPEZ 2008), haben sich insbesondere tiefe neuronale Netze als performante Werkzeuge zur automatisierten Textübersetzung erwiesen (BAHDANAU et al. 2015). Viele Systeme, die zur Sprachmodellierung trainiert wurden, können ebenso zur Textübersetzung eingesetzt werden. Entsprechend bestehen unterschiedliche Deep-Learning-Verfahren zur maschinellen Textübersetzung. Deep-Learning-Verfahren basieren dabei auf neuronalen Netzen, die durch Einsatz unterschiedlicher Zwischenschichten komplexere Strukturen verarbeiten können (SCHMIDHUBER 2015). Beispielhaft zu nennen sind rekurrente neuronale Netze (CHO et al. 2014), gefaltete neuronale Netze (GEHRING et al. 2017) oder Modelle mit Attention-Mechanismen (VASWANI et al. 2017).

c. Textklassifikation

Eine weitere zentrale Aufgabe des Textminings stellt die Textklassifikation dar. Dabei werden Texte vordefinierten Kategorien zugeordnet. Die Klassifikation kann dabei auf Dokumentenebene, Absatzebene, Satzebene oder Sub-Satzebene erfolgen (KOWSARI et al. 2019). Die Verfahren zur Textklassifikation lassen sich erneut grob in eher tradierte Ansätze des ML und moderne Deep-Learning-Verfahren differenzieren.

Unter den tradierten Ansätzen lassen sich Methoden der logistischen Regressionsverfahren (COX und SNELL 2018), Bayes-Klassifikatoren (QU et al. 2018; KIM et al. 2006), Support-Vector-Machines (SIMON und KOLLER 2001), Nächste-Nachbarn-Klassifikatoren (HAN et al. 2001; JIANG et al. 2012) und Entscheidungsbäume (SAFAVIAN und LANDGREBE 1991) subsumieren. In die Gruppe der moderneren Ansätze fallen erneut Deep-Learning-basierte Systeme (LIU et al. 2016c; SUTSKEVER et al. 2014; JOULIN et al. 2016). Besonders hervorzuheben sind die Einflüsse

von Klassifikationsverfahren, die auf vortrainierten Sprachmodellen basieren (YANG et al. 2019; LIU et al. 2019).

d. Topic Modeling

Verfahren des Topic Modelings verfolgen das Ziel, unüberwacht abstrakte Themen aus einer Sammlung von Dokumenten zu extrahieren. Im Kontext des Textminings werden sie eingesetzt, um latente Strukturen in großen Textkorpora aufzudecken. Ein weit verbreitetes Verfahren zur Themenextraktion ist das statistische Wahrscheinlichkeitsmodell Latent Dirichlet Allocation (LDA) (BLEI et al. 2003). Dieses nutzt ein dreistufiges bayessches Modell, welches die Kookkurrenz der enthaltenen Wörter betrachtet. Somit ignorieren diese Verfahren die semantische Bedeutung der enthaltenen Wörter und modellieren abstrakte Themen lediglich statistisch. Auch im Bereich des Topic Modelings haben sich im Laufe der Zeit neuere KI-basierte Verfahren etabliert, die Word-Embeddings (vgl. Kapitel 4) verwenden, um semantische Ähnlichkeiten zwischen Wörtern berücksichtigen zu können (DIENG et al. 2020; ANGELOV 2020; MOODY 2016; GROOTENDORST 2022).

e. Named-Entity-Recognition

Named-Entity-Recognition wird zur Erkennung von benannten Entitäten innerhalb eines Texts verwendet. Jedem Wort eines Satzes wird dabei ein Label aus einer vordefinierten Menge von Labels zugewiesen. Eine einzelne Entität kann sich über mehrere Wörter erstrecken (LAMPLE et al. 2016). Benannte Entitäten können dabei unterschiedlichst definiert sein. Die am häufigsten benannten Entitäten sind Personennamen, Namen von Organisationen oder Ortsbezeichnungen.

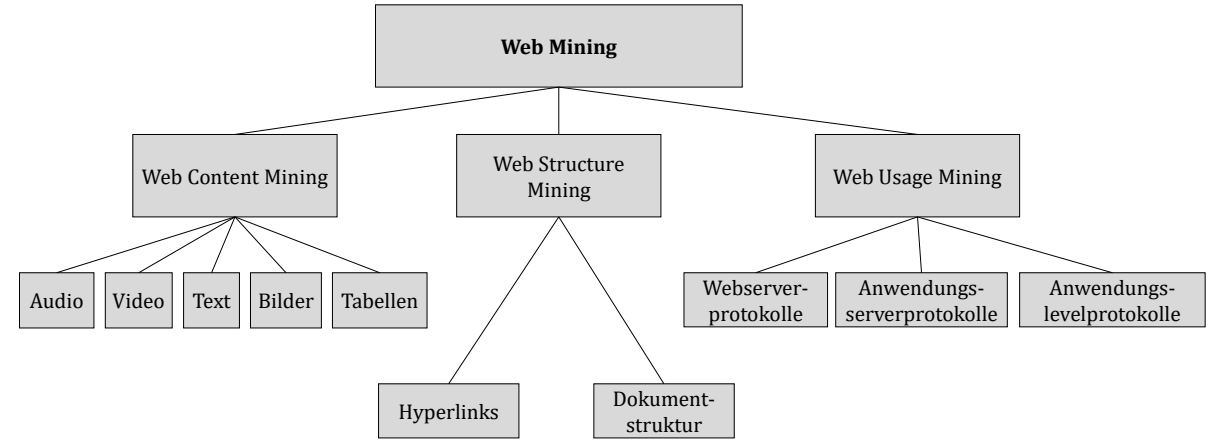
3.1.4 Interpretation

Die identifizierten Muster werden validiert und interpretiert. Die Interpretation und Validierung ist dabei massiv von der vorliegenden Fragestellung sowie den verwendeten Daten abhängig. Häufig erfolgt die Validierung der identifizierten Muster entweder anhand manueller annotierter Trainingsdaten oder durch den Vergleich mit amtlichen Statistiken bzw. traditionell erhobenen Daten oder bestehenden Studien. Nach der Validierung können die generierten Daten als proprietärer Datensatz analysiert oder als Indikator in ein umfassendere übergeordnete Forschungsdesigns integriert werden.

3.2 Die Web Mining Taxonomie

Die erläuterte Verfahrensabfolge kann im Rahmen des Web Minings zu unterschiedlichen Zwecken angewandt werden. In der Fachliteratur wird allgemein in drei Subdisziplinen des Web Minings unterschieden: Web Content Mining, Web Structure Mining und Web Usage Mining (KOSALA und BLOCKEEL 2000; MINER et al. 2012; DUSHYANT et al. 2013). Abbildung 5 zeigt die Taxonomie des Web Minings auf. Die jeweiligen Subdisziplinen werden im Folgenden näher erläutert.

Abbildung 5: Taxonomie des Web Minings.



Quelle: Eigene Darstellung nach DUSHYANT et al. (2013: 99).

3.2.1 Web Content Mining

Unter Web Content Mining wird die Extraktion von Informationen aus Webseiteninhalten verstanden. Webseiteninhalte, die mittels Web Content Mining addressiert werden, sind beispielsweise Texte, Audio-, Video- und Bilddateien oder strukturierte Daten wie z.B. Tabellen oder Listen. Textdaten werden dabei am häufigsten verwendet, um Webseiten zu klassifizieren oder Themen zu identifizieren. Aus der Betrachtung des Forschungskanons geht hervor, dass Web Content Mining die populärste Gattung des Web Minings für die wissenschaftliche Forschung darstellt. Mittels Web Content Mining werden unterschiedliche Arten von Webseiten untersucht. Laut einer Studie von SAMUEL et al. (2019) analysieren über 50 % der Forschungsarbeiten, die mit Web Content Mining arbeiten, allgemeine Webseiten. Etwa jede fünfte Studie beleuchtet den Inhalt von Nachrichtenwebseiten mittels Web Content Mining. Weitere Anwendungsfälle sind Social Media-Webseiten, Suchmaschinen oder Bildungswebseiten.

Insbesondere in den Bibliotheks- und Informationswissenschaften wird Web Mining immer häufiger genutzt, um mittels Topic Modeling latente Themen in wissenschaftlichen Publikationen aufzudecken. Die Ergebnisse dieser Studien zeigen auf, dass unüberwachtes Topic Modeling durchaus in der Lage ist latente Themenstrukturen zu erkennen und diese den jeweiligen Publikationen zuzuordnen (YAU et al. 2014; SUOMINEN und TOIVANEN 2016). Längsschnittstudien untersuchen darüber hinaus Veränderungen im wissenschaftlichen Diskurs anhand von dynamischer Themenmodellierung (ZHANG et al. 2017).

Ein weiterer großer Teil des Forschungskanons nimmt Nachrichtenartikel in den Fokus der Betrachtung. Hierzu bestehen unterschiedliche Ansätze, die versuchen, technologisches Trendscouting auf Basis von Nachrichtenartikeln zu betreiben (IGLESIAS et al. 2016; KAYSER und BLIND 2017; RADINSKY und HORVITZ 2013; JAHANBIN und RAHMANIAN 2020). Während der Coronapandemie sind

darüber hinaus einige Arbeiten entstanden, die die Entwicklung der medialen Berichterstattung mittels Topic Modeling verfolgen (ORDUN et al. 2020; MELO und FIGUEIREDO 2021).

Im Kontext raumbezogener Forschung beschäftigen sich aktuelle Studien vor allem mit der Ableitung von unternehmensbezogenen Informationen aus verschiedenen Webquellen. Diese Arbeiten betonen häufig die Vorteile von Web Mining in Bezug auf Aktualität, Validität, Abdeckung und Verzerrung gegenüber herkömmlichen Datenbanken (ARORA et al. 2013; KINNE und LENZ 2021; YOUTIE et al. 2012; GÖK et al. 2015). Darüber hinaus werden Unternehmenswebseiten zur Messung eines breiten Spektrums von Innovationsaktivitäten genutzt. Neuere Studien nutzen Daten von Unternehmenswebseiten, um das Innovationsverhalten von Unternehmen zu untersuchen (GÖK et al. 2015; LI et al. 2018; KINNE und LENZ 2021; YOUTIE et al. 2012), Kommerzialisierungsstrategien aufzudecken (ARORA et al. 2013), Geschäftsbeziehungen und Kooperationen zu ermitteln (ABBASIHAROFTEH et al. 2021), die Verbreitung von Normen und Richtlinien zu beleuchten (MIRTSCH et al. 2021; GARECHANA et al. 2017) oder die Exportorientierung von Unternehmen zu analysieren (BLAZQUEZ und DOMENECH 2018). Unternehmenswebseiten sind geeignet, um unternehmensbezogene Informationen abrufen zu können, da Unternehmen ihren Internetauftritt nutzen, um potentielle Kund:innen und Geschäftspartner:innen zu akquirieren und ihr Image zu pflegen. Entsprechend sind Firmen interessiert daran ein möglichst vollständiges und präzises Bild des eigenen Profils auf den Webseiten darzustellen (HERNÁNDEZ et al. 2009).

Der aktuell noch wenig ausdifferenzierte Forschungskanon besteht derzeit hauptsächlich aus spezifischen Fallstudien, welche die prinzipielle Eignung von Webdaten zur Beantwortung regionalökonomischer Fragestellungen nutzen. Beispielsweise untersuchen GÖK et al. (2015) die Webseiten von 296 nachhaltigen Unternehmen aus dem vereinigten Königreich im Hinblick auf ihre Forschungs- und Entwicklungsaktivitäten (F&E). Die Autor:innen argumentieren, dass Unternehmen heute ihre Webseiten nutzen, um Informationen über Produkte, Dienstleistungen, Strategien, Partnerschaften und Personalentscheidungen zu veröffentlichen. Sie kommen zu dem Schluss, dass Website-Inhalte eine vielversprechende Datengrundlage für die Analyse von F&E-Aktivitäten sind.

Einen ähnlichen Ansatz verfolgen LI et al. (2018), die in ihrer Studie die Einflüsse der Triple Helix, also der Kooperationsprozesse zwischen Industrie, Staat und Wissenschaft, auf 271 amerikanische kleine und mittlere Unternehmen (KMU) nachzeichnen. Die Autor:innen argumentieren, dass Web Mining genutzt werden kann, um Triple-Helix-Beziehungen auf der Mikroebene zu analysieren. Sie erklären weiter, dass die generierten webbasierten Indikatoren als Ergänzung zu herkömmlichen Sekundärstatistiken verwendet werden können und einen zusätzlichen, tieferen Erklärungsgehalt haben. YOUTIE et al. (2012) führten eine weitere kleinere Fallstudie durch. Ziel die-

ser war es, das Innovationsverhalten von 30 amerikanischen KMU aus dem Bereich der Nanotechnologie anhand von Web-Archivdaten zu untersuchen. Sie kommen zu dem Schluss, dass die Ergebnisse ihrer Studie aufgrund der geringen Stichprobengröße nur begrenzt aussagekräftig sind, aber die Autor:innen verdeutlichen dennoch das Potential von Web Mining zur Generierung von Unternehmensdaten. STICH et al. (2022) nutzen Topic Modeling, um abstrakte Themen auf britischen Unternehmenswebseiten zu identifizieren und diese georeferenzieren zu können. Die Autor:innen sehen in Webdaten einen vielversprechenden Ansatz zur Mikrofundierung von Clustertheorien.

DAAS und VAN DER DOEF (2021) nutzen Webseitentexte, um 4765 Unternehmen hinsichtlich ihrer Innovativität zu klassifizieren. Die Vorhersagen des Klassifikationsmodells wurden den Ergebnissen einer umfassenden Innovationserhebung gegenübergestellt. Trotz der hohen Vorhersagegenauigkeit des finalen Modells von 88 %, nennen die Autor:innen einige methodische Limitationen ihrer Studie. Beispielsweise berücksichtigt das Modell lediglich 584 Wortstämme, sodass es anfällig für sprachliche Veränderungen ist. Daher schlagen sie eine Kombination aus Web Mining und Expert:innenmeinungen vor, um eine stabilere Klassifikation zu ermöglichen.

SCHWIERZY et al. (2022) verwenden die Webseitentexte deutscher Unternehmen zur Kartierung von Unternehmen, die sich mit dem Thema 3D-Druck beschäftigen. Das zugrundeliegende Klassifikationssystem ist dabei in der Lage Unternehmen anhand des Webseitentexts in vier Kategorien einzuteilen: Hersteller, Dienstleister und Händler sowie Unternehmen, die lediglich über die Technologie informieren. Die Autor:innen sehen in der Methodik vor allem Vorteile gegenüber der Verwendung von Patentdaten. Einerseits ermöglichen Webseitentexte eine Identifizierung von Technologieanwender:innen, während Patentdaten lediglich die Innovatore:innen abbilden. Andererseits ermöglicht diese Granularität der Analyse tiefergreifende Untersuchungen zu Diffusions- und Adaptionsprozessen.

PAPAGIANNIDIS et al. (2018) greifen auf Webtexte von Unternehmen zurück, um feiner aufgelöste Industrieklassifikationen zu erhalten. Die Autor:innen betonen insbesondere drei zentrale Vorteile ihres Ansatzes gegenüber klassischer Industrieklassifikationen: Erstens sind bestehende Klassifikationen häufig zu grob, um ähnliche von gleichen Wirtschaftsbereichen zu unterscheiden. Zweitens fällt es Unternehmen schwer sich exakt einem Wirtschaftsbereich zuzuordnen, sodass drittens Unternehmen häufig allgemeine Kategorien wie z.B. Unternehmensdienstleistungen als Geschäftsbeschreibung wählen.

HÉROUX-VAILLANCOURT et al. (2020) führen eine Validierungsstudie zum Vergleich der durch Webseitentexte generierten Variablen mit Ergebnissen einer klassischen Umfrage durch. Die signifikanten Korrelationen der Studie bestätigen die Möglichkeit, Webdaten als Ergänzung oder Ersatz für klassische statistische Variablen zu nutzen.

Die umfangreichste georeferenzierte Analyse deutscher Unternehmenswebseiten von KINNE und LENZ (2021) umfasst 685.057 Unternehmen. Die Datengrundlage dieser Studie ist die größte deutsche Unternehmensdatenbank, das Mannheimer Unternehmenspanel (BERSCH et al. 2014). Dieses bezieht zweimal jährlich Unternehmensdaten von dem Verband der Vereine Creditreform e.V., bereitet die Datensätze auf und reichert sie mit weiteren Informationen an. Allerdings sind die Daten nicht frei zugänglich, sondern werden primär von Forschenden des Zentrums für europäische Wirtschaftsforschung genutzt. Diese Arbeiten zeigen unter anderem wie Webdaten zu einem besseren Verständnis der komplexen Wechselwirkungen zwischen Innovation und Raum beitragen können. KINNE und RESCH (2018) demonstrieren beispielsweise, dass Webdaten geeignet sind, um die Mikrostandortmuster von Softwareunternehmen zu identifizieren. KINNE und LENZ (2021) nutzen diesen Datensatz, um innovative Unternehmen anhand ihrer Webseiten-Texte zu identifizieren. Die Ergebnisse dieser Studie weisen große Ähnlichkeiten mit Vergleichsdaten aus dem Mannheimer Unternehmenspanel auf. Die Ergebnisse unterstreichen damit den aktuellen Stand der Forschung und weisen auf weitere große Potentiale der webdatenbasierten Innovationsforschung hin.

In diesem Kontext sind außerdem Arbeiten zu nennen, die Patentdokumente mittels Text Mining analysieren (TSENG et al. 2007). Dieser Forschungskanon beschäftigt sich beispielsweise mit der automatisierten Klassifikation von Patenten mittels Topic Modeling (YUN und GEUM 2020; VENUGOPALAN und RAI 2015; HU et al. 2014) und der Ableitung von Patentlandkarten zur Analyse von technologischen Wettbewerbstrends (YOON et al. 2013). Weiterhin bietet NLP die Möglichkeit aus Patenttexten neue Metriken zur Untersuchung von Innovationen und deren Nutzung abzuleiten (ARTS et al. 2021; BALSMEIER et al. 2018).

3.2.2 Web Structure Mining

Web Structure Mining wird genutzt, um relationale Informationen über das Internet zu gewinnen. Dabei werden Webgraphen erzeugt, bei denen Webseiten als Knoten und Verbindungen zwischen Webseiten (Hyperlinks) als Kanten eines Netzwerks fungieren. Mittels Web Structure Mining werden Strukturinformationen über das Internet gewonnen (PARK 2003). Beispielsweise können somit zentrale Webseiten oder Webseitencluster identifiziert werden. Ein prominenter Anwendungsfall von Web Structure Mining stellt der PageRank-Algorithmus dar (PAGE et al. 1999). Der speziell für Googles Suchmaschine entwickelte Algorithmus nutzt die Verlinkungen von Webseiten untereinander, um zentrale Webseiten zu identifizieren und somit die Ergebnisreihenfolge einer Suchabfrage festzulegen. Der PageRank-Algorithmus hat sich seit der Einführung zu einem zentralen Werkzeug der Netzwerkanalyse entwickelt und wird in unterschiedlichsten Kontexten angewandt, z.B. zur Bewertung der Relevanz wissenschaftlicher Literatur (SENANAYAKE et al. 2015).

Weitere Anwendungsfälle von Web Structure Mining sind beispielsweise Analysen von Unternehmensverlinkungen (KRÜGER et al. 2020; ZHU et al. 2020; ABBASIHAROFTEH et al. 2021). Diese werden dabei häufig als Indikator für eine geschäftliche Beziehung zwischen den Firmen genutzt. Aus diesen Netzwerken lassen sich wiederrum Subnetzwerke und zentrale Akteure ableiten. ZHU et al. (2020) zeigen, dass Unternehmensverlinkungen häufiger zwischen Unternehmen existieren, die auch räumlich eine gewisse Nähe aufweisen. HELLMANZIK und SCHMITZ (2017) nutzen Verlinkungen zwischen Akteuren der Finanzbranche als Indikator der virtuellen Nähe. Die Autor:innen zeigen anhand von Aktieninvestitionen, dass virtuell verwandtere Staaten auch finanziell stärker integriert sind. TRANOS et al. (2022) verwenden Verlinkungen zwischen Unternehmen, um interregionale Handelsströme zu messen. Der verwendete Random Forest-Klassifikator ist in der Lage zuverlässig die realen Handelsströme anhand der Hyperlinks vorherzusagen.

Verlinkungen können nicht nur zwischen Unternehmen analysiert werden. MEIJERS und PERIS (2019) identifizieren beispielsweise Bezeichnungen niederländischer Orte, um auf deren Basis digitale Lage- und Funktionsbeziehungen zwischen Städten modellieren zu können. Die Ergebnisse zeigen, dass räumlich nahe gelegene Orte auch im digitalen Raum stärkere Beziehungen untereinander aufweisen, jedoch auch intensive Beziehungen über größere Distanzen bestehen. Ähnlich gehen DEVRIENDT et al. (2008) vor. Die Autor:innen bilden auf Basis der Kookkurrenz von Städtenamen ein digitales Netzwerk, welches transnationale Verbindungen darstellt. KEßLER (2017) nutzt die Linkstruktur von Wikipedia, um Verbindungen zwischen den Wikipediaeinträgen deutscher Gemeinden herzustellen. Die Ergebnisse zeigen, dass die Linkstruktur deutliche Überlappungen mit der Raumordnungstruktur Deutschlands aufweist.

VAUGHAN et al. (2006) führen eine qualitative Untersuchung mit 280 nordamerikanischen Unternehmen aus der IT-Branche durch, um die Bedeutung von Verlinkungen zwischen Unternehmen besser verstehen zu können. Die Autor:innen kommen zu dem Ergebnis, dass Verlinkungen mehrheitlich für Geschäftsbeziehungen wie z.B. Kunden-Lieferanten-Beziehungen, Sponsoringpartnerschaften oder Pressemeldungen stehen.

Zu einer ähnlichen Erkenntnis gelangen VAUGHAN und WU (2004), die die Hyperlinks der 100 größten IT-Unternehmen Chinas untersuchen. Die Verlinkungen von Unternehmen können entsprechend genutzt werden, um Unternehmen zu charakterisieren. KRÜGER et al. (2020) zeigen beispielsweise, dass die Innovationsleistung eines Unternehmens signifikant von dessen Verlinkungen zu anderen Unternehmen abhängig ist. KATZ und COTHEY (2006) demonstrieren am Beispiel europäischer und kanadischer Innovationssysteme, dass auch komplexere Akteurskonglomerate mittels Web-Indikatorik abgebildet werden können. Weitere Studien untersuchen mittels Web

Structure Mining die Vernetzung von Tourismusdestinationen (RAISI et al. 2018), Nachrichtenökosystemen (SJØVAAG et al. 2019) oder akademischen internationalen Netzwerken (ORTEGA und AGUILLO 2008; THELWALL et al. 2005).

3.3.3 Web Usage Mining

Web Usage Mining analysiert die Nutzung bestimmter Webseiten, beispielsweise über Aufrufzahlen oder Suchanfragen. Während Web Content Mining die Inhalte und Web Structure Mining die Strukturen als Primärdaten betrachten, zielt Web Usage Mining darauf ab, aus dem Verhalten von Internetnutzer:innen Informationen abzuleiten (KUMAR und KUMAR 2021). Web Usage Mining wird in erster Linie von Webseitenbetreiber:innen genutzt, um Informationen über Verweildauer, Einkaufsverhalten, Klickverhalten und andere Nutzungsindikatoren zu gewinnen. Die Daten werden anschließend verwendet, um vorherzusagen, welche Webseite vom Nutzer:innen als nächstes besucht wird (CHATTERJEE et al. 2017), den Besucher:innen passende Inhalte zu präsentieren (SUADAA 2014) oder auf Basis der Browsinghistorie neue Inhalte vorzuschlagen (CHO und KIM 2004). Diese Informationen bilden häufig die Grundlage für die Entwicklung weiterer digitaler Marketingmaßnahmen.

Eine gängige Quelle wissenschaftlicher Forschung stellt *Google Trends* dar (JUN et al. 2018). Die Daten zu Suchanfragen an Google ermöglichen somit beispielsweise die Erkennung von Krankheitswellen (GINSBERG et al. 2009; FELDMAN 2004; CARNEIRO und MYLONAKIS 2009), die Analyse von Handelsverhalten auf Finanzmärkten (PREIS et al. 2013) sowie die Kurzfristvorhersage wirtschaftlicher Indikatoren (CHOI und VARIAN 2012) und Arbeitslosigkeit (ASKITAS und ZIMMERMANN 2009). Eine weitere häufig genutzte Grundlage wissenschaftlicher Forschung zu Web Usage Mining sind Daten aus den sozialen Medien (ROUSIDIS et al. 2020), die ebenfalls vielfältige Analysemöglichkeiten bieten. Die Daten werden unter anderem genutzt, um Preisänderungen auf dem Immobilienmarkt vorherzusagen (ZAMANI und SCHWARTZ 2017), Mobilitätsverhalten in Städten zu schätzen (ABBASI et al. 2017) oder den Ausgang politischer Wahlen zu prognostizieren (CAMERON et al. 2016).

4 Evolution des Natural Language Processings

Nachdem in den vorangegangenen Kapiteln die konzeptionellen Grundlagen des Web Mining und der neuen Rolle von Text als Datenquelle in der empirischen Sozialforschung beleuchtet wurden, gilt es nun ein Verständnis für die technischen und algorithmischen Details der automatisierten, quantitativen Textverarbeitung zu schaffen. Das vorliegende Kapitel stellt daher nach einem knappen Abriss der historischen Entwicklung des NLP die methodischen Entwicklungen der letzten beiden Dekaden vor. Diese reichen von den ersten statistischen Vektorverfahren über Word-Embeddings und verschiedene Ausbaustufen neuronaler Netze bis hin zu modernsten Transformermodellen.

4.1 Entwicklung des Natural Language Processings

Unter dem Oberbegriff NLP subsumieren sich unterschiedliche Techniken, die das gemeinsame Ziel verfolgen einem Computer eine menschenähnliche Verarbeitung natürlicher Sprache bzw. natürlicher Texte beizubringen. Der Obergriff lässt sich nochmals in die Subkategorien Natural Language Understanding (NLU) und Natural Language Generation (NLG) aufteilen. NLU beschreibt die Fähigkeit von Computern gesprochene bzw. geschriebene Sprache verstehen und interpretieren zu können. NLG baut auf Verfahren des NLU auf und meint die Fähigkeit von Computern Text zu produzieren. Häufig werden NLG-Verfahren verwendet, um eine Antwort auf eine Nutzer:inneneingabe zu geben. Der Prozess der NLG umfasst daher neben des NLU die Selektion passender Inhalte, die textuelle und linguistische Organisation der Outputsequenz sowie die abschließende Ausgabe des produzierten Texts (KHURANA et al. 2022).

Das Methodenspektrum des NLP reicht dabei von einfachen Aufgaben wie Tokenizing, also die Identifikation einzelner Wörter innerhalb eines Satzes, bis hin zu komplexen Funktionen, wie beispielsweise die automatische Beantwortung von Fragen oder die Ableitung von Knowledge Graphs aus Textsammlungen. Häufig bauen die einzelnen Techniken aufeinander auf, um z.B. unstrukturierten Text für ambitionierte Textverarbeitungsalgorithmen vorzubereiten. Das Forschungsfeld des NLP blickt auf eine dynamische Entwicklungsgeschichte zurück.

JONES (1994) gliedert die Historie des Forschungsfeldes in vier Phasen. Die erste Phase beginnt in den 1950er Jahren und fokussiert sich primär auf die maschinelle Übersetzung von Texten. Aufgrund der primitiven EDV-Ausstattung liegt das Hauptaugenmerk damaliger Forschung auf einfachen Wort-für-Wort Übersetzungen. Entsprechend werden in dieser ersten Phase der NLP-Forschung keine semantischen Zusammenhänge berücksichtigt, sondern vielmehr die Syntax unterschiedlicher Sprachen fokussiert. In der zweiten Phase der NLP-Forschung in den 1960er und 1970er Jahren entstehen basierend auf früher KI-Forschung erste Ansätze, die die Semantik von

Sprache berücksichtigen. In dieser Phase werden erste rudimentäre Frage-Antwort-Systeme entwickelt, die auf Basis hinterlegter Datenbanken Fragen innerhalb eines abgesteckten Themengebiets beantworten können (WINOGRAD 1972; WOODS 1977). Daher wird diese zweite Phase als KI-gestützt und semantisch orientiert beschrieben (JONES 1994).

Die dritte Entwicklungsphase des NLP in den 1970er und 1980er Jahren wird als grammatikalisch-logische Phase beschrieben. Basierend auf grammatikalischen Theorien der Linguistik sowie der Ausrichtung der KI-Forschung hin zu logischer Wissensdarstellung entstehen NLP-Anwendungen wie automatisierte Nachrichtenübersetzungen oder semantische Datenbankabfragen. Die vierte Phase der NLP-Entwicklung Ende der 1980er Jahre ist geprägt von dem rasanten Wachstum von Rechnerleistung sowie verfügbarem maschinell lesbarem Text und markiert den Ausgangspunkt der statistischen Sprachverarbeitung mittels ML, sodass sich NLP ab den 1990er Jahren immer stärker mit den Begriffen Big Data und Text Mining verwebt (HIRSCHBERG und MANNING 2015). Infolgedessen werden vor allem statistische Modelle (z.B. logistische Regressionsverfahren, Support-Vector-Machines) genutzt, um Text quantitativ verarbeiten zu können (JOACHIMS 1998; FAN et al. 2009).

Um diese unterschiedlichen Verfahren des ML anwenden zu können, werden die Texte im Vorfeld in statistische, numerische Repräsentationen übersetzt (DAELEMANS und HOSTE 2002; OLSSON 2009). Die Übersetzung von Text in numerische Features wird in einschlägiger Literatur als Textvektorisierung (engl. Text Vectorization) bezeichnet. Das simpelste Konzept zur numerischen Textrepräsentation ist das *bag-of-words* (BOW) Verfahren. Dieses wird bereits in den 1950er Jahren entwickelt (HARRIS 1954) und aufgrund seiner Einfachheit, Effizienz und Genauigkeit noch bis in die 2000er Jahre angewandt (LE und MIKOLOV 2014). Jedes Wort eines Textkorpus wird dabei als einzigartiges Feature verstanden. Jedem Wort des Korpus wird ein numerisches Feature zugewiesen, sodass ein Textkorpus aus n Vokabeln als n-dimensionaler Vektor repräsentiert wird. Die Länge des Vektors wird dabei fix bestimmt von dem Umfang des Vokabulars des Textkorpus.

Tabelle 1 veranschaulicht das BOW-Verfahren. Das Beispiel verwendet ein sieben Wörter umfassendes Vokabular, aus welchem verschiedene Sätze gebildet werden können.

Tabelle 1: Beispiel für eine Bag-of-words Kodierung.

Satz	die	der	hörsaal	ist	universität	geschlossen	in
die universität ist geschlossen	1	0	0	1	1	1	0
der hörsaal ist in der universität	0	2	1	1	1	0	1
ist der hörsaal der universität geschlossen	0	2	1	1	1	1	0
der hörsaal ist geschlossen	0	1	1	1	0	1	0
in der universität ist der hörsaal	0	2	1	1	1	0	1

Quelle: Eigene Darstellung.

Die Zahlen in der Tabelle geben die Auftrittshäufigkeit des jeweiligen Wortes innerhalb des jeweiligen Satzes an. Somit kann jeder Satz in einen Wortvektor der Länge sieben übersetzt werden. Beispielsweise lässt sich der Satz *die universität ist geschlossen* in den Wortvektor 1001110 übersetzen.

Große Textmengen enthalten in der Regel viele unterschiedliche Vokabeln, sodass die Vektoren extrem umfangreich werden. Diese Eigenschaft macht die Verarbeitung der Vektoren sehr rechenintensiv. Weiterhin sind die Vektoren dünnbesetzt und enthalten entsprechend wenig Informationsgehalt. Beispielsweise geben die Vektoren eines BOW-Modells keine Auskunft über die Ähnlichkeit einzelner Wörter. Informationen über die Reihenfolge der Wörter gehen dabei ebenfalls verloren, sodass beispielsweise zwei unterschiedliche Sätze, die aus den gleichen Wörtern bestehen, in gleiche numerische Repräsentation übersetzt werden. Dabei kann die semantische Bedeutung der Sätze vollkommen unterschiedlich sein. Trotz dieser Limitationen sind ML-Verfahren, die auf BOW-Features basieren, beispielsweise in der Lage rudimentär die Ähnlichkeit von Dokumenten zu bestimmen und damit Texte in vordefinierte Kategorien einzuteilen. Innerhalb der BOW-Verfahren lassen sich einige Varianten subsumieren, wie beispielweise Häufigkeitsvektoren, die die Häufigkeit des Auftretens eines Wortes innerhalb eines Dokuments berücksichtigen. Ein populäres Verfahren, um einerseits die Häufigkeit eines Wortes innerhalb eines Dokuments zu berücksichtigen und andererseits diese zu normalisieren, ist die TF/IDF-Technik (vgl. Kapitel 3.1.2).

4.2 Word-Embeddings

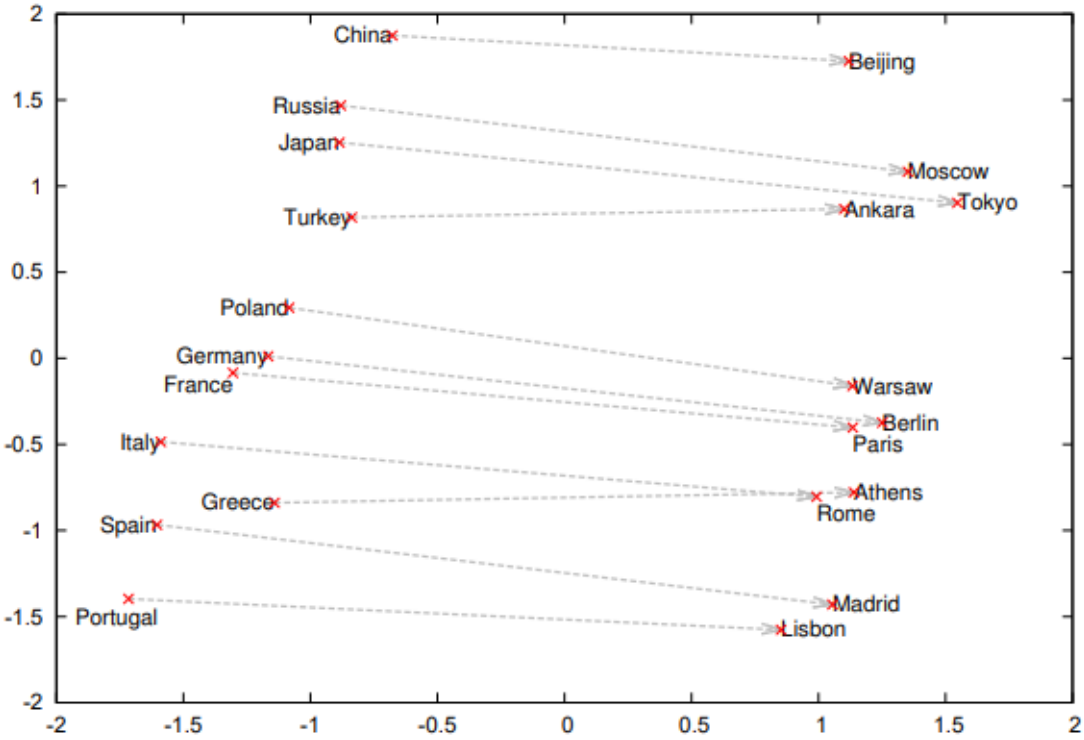
Um den genannten Nachteilen der BOW-Modelle zu begegnen, wird ab den 1980er Jahren an kontiniuerlichen Repräsentationen von Wörtern geforscht (RUMELHART et al. 1988). Die numerischen Repräsentationen werden dabei auf Basis großer nicht-annotierter Textkorpora erlernt und beziehen für die Berechnung der Vektoren neben dem Auftreten der Wörter auch deren Kontext mit ein (DEERWESTER et al. 1990; LUND und BURGESS 1996; SCHÜTZE 1992). Mathematisch wird zunächst eine hochdimensionale Wort-Kontext-Matrix erzeugt. Die Informationen zur Kookkurrenz von Wörtern und ihrem Kontext werden anschließend durch Matrizenfaktorisierung reduziert. Übergeordnetes Ziel der Dimensionsreduktion ist es, die Varianz der Ausgangsmatrix möglichst gut zu erhalten. Während die Länge von dünnbesetzten BOW-Vektoren durch den Umfang des Vokabulars des zugrundeliegenden Textkorpus determiniert ist, haben die dichtbesetzten, kontinuerlichen Vektoren von Word-Embeddings eine fixe Länge (JURAFSKY und MARTIN 2019).

Dieser Ansatz ermöglicht es erstmals auch die semantische Ähnlichkeit zwischen Wörtern zu quantifizieren. Semantisch ähnliche Wörter, wie Synonyme, Hyponyme oder Hyperonyme liegen im Vektorraum ebenfalls nah beieinander, während zwischen semantisch unähnlichen Wörtern größere Distanzen liegen. Der semantische Gehalt der Word-Embeddings ist dabei von der Größe

des gewählten Kontextfensters abhängig. Engere Kontextfenster sorgen für ein Erlernen der Wortbedeutungen, während weitere Kontextfenster eher Themen darstellen (RODRIGUEZ und SPIRLING 2022). Somit kann das Prinzip der Word-Embeddings auch auf Absätze übertragen werden und es können Dokument-Embeddings berechnet werden (AI et al. 2016; DAI et al. 2015).

Abbildung 6 zeigt die mittels Word-Embeddings quantifizierten Beziehungen zwischen Staaten und ihren Hauptstädten. Es wird deutlich, dass Word-Embeddings in der Lage sind automatisch solche latenten Beziehungen aus großen Textmengen abzuleiten, da sich die Lagebeziehungen zwischen den Nationalstaaten und den jeweiligen Hauptstädten stark ähneln.

Abbildung 6: Hauptkompentenprojektion von Wortvektoren.



Quelle: MIKOLOV et al. (2013b).

Da Word-Embeddings in der Lage sind Wörter und ihre semantische Bedeutung in Form von Vektoren quantifizieren zu können, ist es möglich Wörter miteinander zu verrechnen. Ein Beispiel stellt die Addition bzw. Subtraktion von Wortvektoren dar. Tabelle 2 zeigt Beispiele zur Addition von ausgewählten Wortvektoren auf. Wie daraus hervorgeht, können durch Vektoradditionen ebenfalls semantische Beziehungen zwischen Wörtern quantifiziert werden. Die Addition bzw. Subtraktion von Wortvektoren demonstriert beispielhaft wie Word-Embeddings im Hintergrund funktionieren. Beispielsweise entsteht durch die Addition des Wortvektors *German* mit dem Wortvektor *airlines* das Wort *airline Lufthansa*. An diesem Beispiel zeigt sich, dass in den verwendeten Wortvektoren die Informationen gespeichert sind, dass *Lufthansa* eine *deutsche Airline* ist.

Tabelle 2: Beispiele für die Addition von Wortvektoren.

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Quelle: Eigene Darstellung nach MIKOLOV et al. (2013a: 7).

Die Forschung zur statistischen Sprachmodellierung mittels kontinuierlichen Wortrepräsentationen entwickelt im Laufe der Zeit immer komplexere Verfahren, um Wörter als kontinuierliche Vektoren formalisieren zu können (ELMAN 1990; BENGIO et al. 2000). Insbesondere künstliche, neuronale Netze (KNN), die zunächst in einer vorgeschalteten Verarbeitungsschicht die Wortvektoren trainieren, um diese anschließend verarbeiten zu können, zeigen große Erfolge. Sie ermöglichen es der umfassenden Dimensionalität von Textdaten zu begegnen. Das erste Sprachmodell, welches neuronale Netzwerke zur Prozessierung von Textdaten einsetzt wird von BENGIO et al. (2000) vorgestellt. Die Arbeit von BENGIO et al. (2000) markiert damit den Ausgangspunkt moderner NLP-Forschung. Fundamentale Fortschritte erfährt die NLP-Forschung Anfang der 2010er Jahre mit dem Durchbruch Deep-Learning basierter Algorithmen und der Entwicklung unterschieder Gattungen neuronaler Netzwerke (LECUN et al. 2015; GOODFELLOW et al. 2016). Diese wissenschaftlichen Erfolge der Sprachmodellierung fungieren in der Folge als Rückgrat vieler ebenso erfolgreicher nachgelagerter NLP-Anwendungen wie z.B. parsing, tagging oder named-entity-recognition (COLLOBERT und JASON 2008; COLLOBERT et al. 2011; MIKOLOV et al. 2013b; TURIAN et al. 2010)

Die drei populärsten Ansätze zur Berechnung von Word-Embeddings sind Word2Vec (MIKOLOV et al. 2013a), GloVe (PENNINGTON et al. 2014) und fastText (BOJANOWSKI et al. 2017). Die Grundprinzipien und Unterschiede der jeweiligen Frameworks werden folgend vorgestellt.

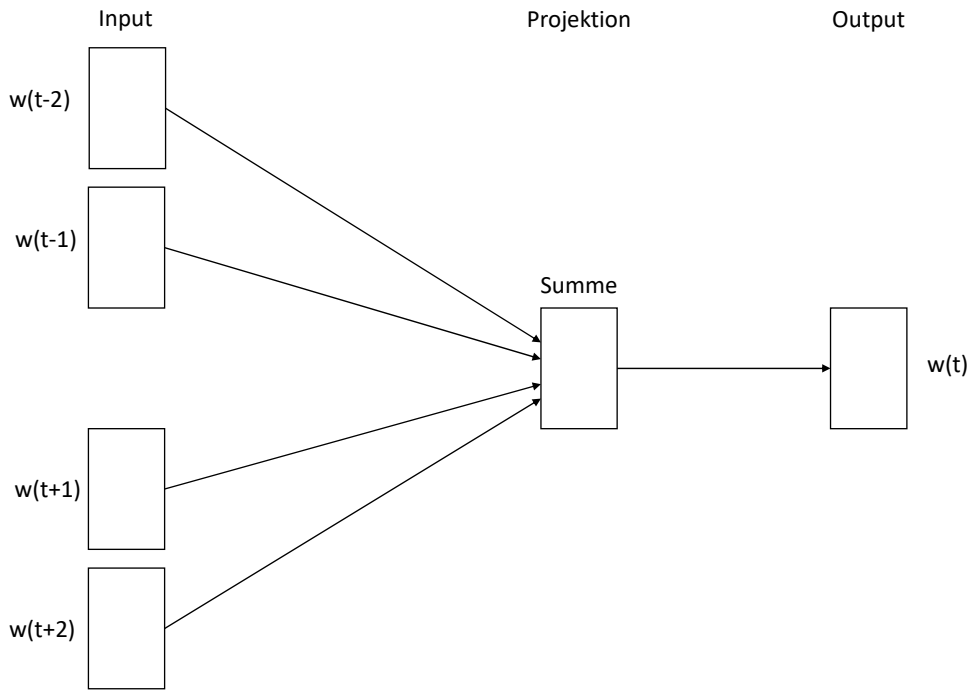
4.2.1 Word2Vec

Den Ausgangspunkt zur Erforschung moderner Word-Embeddings markieren MIKOLOV et al. (2013a) und präsentieren zwei innovative Algorithmen, um Wörter effizienter und aussagekräftiger in numerische Repräsentationen übersetzen zu können. Der *continious-bag-of-words* (CBOW)-Algorithmus kann als Weiterentwicklung des bekannten BOW-Algorithmus beschrieben werden. Für das Training der Word-Embeddings werden große Mengen nicht-annotierter Daten verwendet. Die Semantik eines Wortes wird dabei über seinen Kontext erfasst. Mit dem CBOW-Algorithmus wird versucht ein Zielwort anhand der umliegenden Wörter vorherzusagen und iterativ ein immer besseres Verständnis der Zielsprache zu gewinnen. Die Länge des Kontextfensters

- also die Anzahl benachbarter Wörter, die zur Vorhersage des Zielwortes verwendet werden – kann frei gewählt werden. Üblicherweise werden die Wörter vor dem Training in Kleinbuchstaben transformiert, um Duplikate zu vermeiden. Anschließend werden jeweils Wortpaare gebildet und als Trainingsdaten verwendet.

Abbildung 7 zeigt die formalisierte Modellarchitektur des CBOW-Modells. Die Inputs sind in diesem Beispiel jeweils die beiden vor- und nachstehenden Wörter des vorherzusagenden Wortes. Diese werden gemittelt und somit aus dem Kontext das Zielwort ermittelt.

Abbildung 7: Modellarchitektur CBOW.

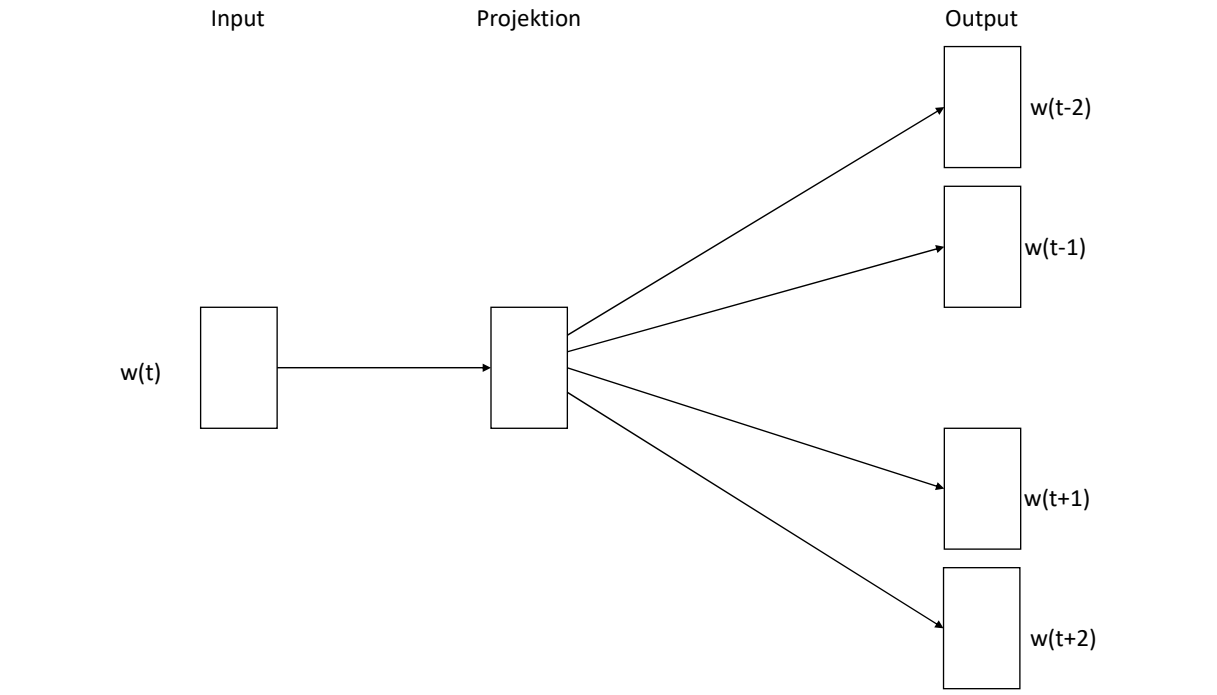


Quelle: Eigene Darstellung nach MIKOLOV et al. (2013a: 5).

Das Skip-gram-Modell funktioniert exakt entgegengesetzt zu der Berechnung des CBOW-Modells. Ziel des Modells ist es, den Kontext eines Wortes anhand eines Wortes vorherzusagen. Abbildung 8 veranschaulicht die Modellarchitektur des Skip-gram-Modells analog zu der des CBOW-Modells. Hierbei fungiert ein Wort eines Satzes w(t) als Input und das Modell wird anschließend trainiert, um die umliegenden Wörter des Inputwortes vorherzusagen.

Aufgrund der unterschiedlichen Modellarchitektur und Intuition der Modelle haben diese auch unterschiedliche Stärken und Schwächen. Das CBOW-Modell lässt sich schneller trainieren und ist in der Lage häufig vorkommende Wörter besser numerisch zu repräsentieren. Das Skip-gram-Modell hingegen zeigt bessere Ergebnisse beim Training mit kleineren Datensätzen und erzielt bessere Ergebnisse bei der Vorhersage seltenener vorkommender Wörter (MIKOLOV et al. 2013a).

Abbildung 8: Modellarchitektur Skip-gram.



Quelle: Eigene Darstellung nach MIKOLOV et al. (2013a: 5).

4.2.2 Glove

Glove steht für „Global Word Vectors“ und wird 2014 von Forschenden der Stanford Universität vorgestellt (PENNINGTON et al. 2014). Die Autoren kritisieren, dass Word2Vec-Modelle globale Informationen zur Kookkurrenz von Wörtern und Kontexten nicht berücksichtigen. Kommen Wörter innerhalb des Korpus in bestimmten Kontexten häufiger vor als in anderen, wird diese globale Information für die Berechnung der Word-Embeddings nur latent berücksichtigt. Die Idee hinter dem Glove-Modell ist es folglich, häufig auftretende Kontext-Word-Kombinationen stärker zu gewichten als seltener auftretende Kombinationen. Glove nutzt daher die traditionellen Wort-Kontext-Matrizen als Grundlage, um globale Kontextinformationen zu Wörtern berücksichtigen zu können. Diese Informationen werden innerhalb des Modells mit den lokalen Informationen log-bilinearer Modelle (wie z.B. Word2Vec) verrechnet.

4.2.3 Fasttext

Fasttext wird 2017 von der Facebook AI Research Gruppe entwickelt und kann als Weiterentwicklung des Word2Vec-Modells verstanden werden. Ein bis dato ungelöstes Problem der NLP-Forschung war es Wörter, die nicht oder nur sehr selten in den Trainingsdaten vorkommen, adäquat numerisch repräsentieren zu können. BOJANOWSKI et al. (2017) stellen Fasttext daher vor, um diesem Problem zu begegnen. Während Glove und Word2Vec als kleinste Untersuchungseinheit Wörter betrachten, setzt Fasttext eine Ebene tiefer an. Einzelne Wörter werden dabei nochmals

in kleinere Einheiten - sogenannte character-n-grams - zerlegt, die in der Regel drei bis sechs Buchstaben lang sind. Mit diesem Verfahren gehen zwei übergeordnete Vorteile einher.

Erstens kann Fastttext durch das Erlernen von character n-grams auch unbekannte Wörter besser einordnen, da in der Regel einzelne character n-grams des Wortes bekannt sind. Entsprechend kann die Bedeutung eines eigentlich unbekannten Wortes aus der Zusammensetzung einzelner character-n-grams erschlossen werden. Das ist vor allem bei Konjugationen oder Flexionen von Wörtern der Fall. Der zweite Vorteil ist, dass Fasttext im Vergleich zu Word2Vec und Glove weniger Trainingsdaten benötigt, um robuste Ergebnisse zu erzielen, da durch das Training auf n-gram-Ebene mehrere Trainingsentitäten aus einem Wort erzeugt werden.

Tabelle 3 zeigt anhand dreier Beispiele die Zerlegung einzelner Wörter in n-gramme. Die deutsche Sprache enthält einige zusammengesetzte Substantive, sodass insbesondere eine Subwortbetrachtung dieser einen zusätzlichen semantisches Verständnis liefern kann. Beispielsweise ermöglicht die Teilung der Wortes *Autofahrer* in die character-n-grams *fahr*, *fahrer* und *auto* Informationen zu den einzelnen Subworten aus anderen Kontexten bei der numerischen Einbettung zu berücksichtigen.

Tabelle 3: Beispiele für character n-grams.

Wort	**n-gramme**		
autofahrer	fahr	fahrer	auto
freundeskreis	kreis	kreis	freund
grundwort	wort	wort>	grund
sprachschule	schul	hschul	sprach
Tageslicht	Licht	gesl	tages

Quelle: Eigene Abbildung verändert nach BOJANOWSKI et al. (2017: 143).

Ein weiterer Vorteil von Fasttext ist die Verarbeitungsgeschwindigkeit bei nachgelagerten NLP-Anwendungen, wie beispielsweise Textklassifikation. JOULIN et al. (2016) zeigen in einer Vergleichsstudie auf, dass das Training eines Fasttext-Klassifikationsmodells bis zu 15.000 Mal schneller verläuft als mit bis dato bekannten Standardverfahren. Die Klassifikationsgüte wurde in dieser Benchmarkstudie durch das deutlich schnellere Training nicht beeinflusst.

4.3 Rekurrente neuronale Netze

Trotz der verschiedenen Modellarchitekturen sind Word-Embeddings allgemein noch einige Nachteile inhärent, die eine performantere Textverarbeitung behindern. Einerseits sorgt die fixe Größe des Kontextfensters dafür, dass keine individuellen Kontexte beim Training von Word-Embeddings einbezogen werden. Andererseits können Word-Embeddings einem Wort lediglich eine semantische Bedeutung in Form eines Vektors zuordnen. Kontextabhängig können einem Wort

insbesondere bei Polysemen unterschiedliche Bedeutungen zuwiesen sein (JURAFSKY und MARTIN 2019).

Prädestiniert für das ML von solchen nicht-linearen, komplexen, semantischen Zusammenhängen sind KNN. KNN haben das menschliche Gehirn hinsichtlich Aufbau und Funktionsweise zum Vorbild. Insbesondere die Fähigkeit, abstrakte Strukturen und Interdependenzen zu erlernen und das erlernte Wissen nach Abschluss der Lernphase in unterschiedlichen Kontexten einsetzen zu können, stellt ein relevantes Merkmal von KNN dar (JURAFSKY und MARTIN 2019).

Diese Generalisierungsfähigkeit ist für die Anwendung von KNN in NLP besonders relevant, da sie so in der Lage sind ein allgemeines Sprachverständnis zu entwickeln und auf dessen Basis unterschiedliche Aufgaben zu bearbeiten. KNN weisen darüber hinaus gegenüber einfacher Algorithmen eine höhere Fehlertoleranz hinsichtlich der Eingabedaten auf. Sie besitzen die Fähigkeit quasi automatisch die relevanten Elemente in den Eingabedaten zu erkennen und diese zu verarbeiten (GRAUPE 2013).

Um diese erfolgskritischen Eigenschaften von KNN bei der Datenverarbeitung auch für NLP nutzbar zu machen, wird mit unterschiedlichen bestehenden Modellarchitekturen neuronaler Netze experimentiert. Nachdem zunächst sogenannte gefaltete (engl. convolutional) neuronale Netze (CNN) eingesetzt werden, setzen sich im Laufe der Zeit zunehmend rekurrente Modellarchitekuren (rekurrente neuronale Netze, RNN) durch (GOLDBERG 2017b). RNN verfügen über eine erfolgskritische Eigenschaft, die insbesondere bei der Verarbeitung von Textsequenzen enorm wichtig ist. Sie verarbeiten im Gegensatz zu CNN oder KNN Inputsequenzen nicht hierarchisch, sondern sequentiell (ELMAN 1990). Abbildung 9 zeigt die Unterschiede im Informationsfluss zwischen einem RNN und einem *feedforward* neuronalen Netz (FNN).

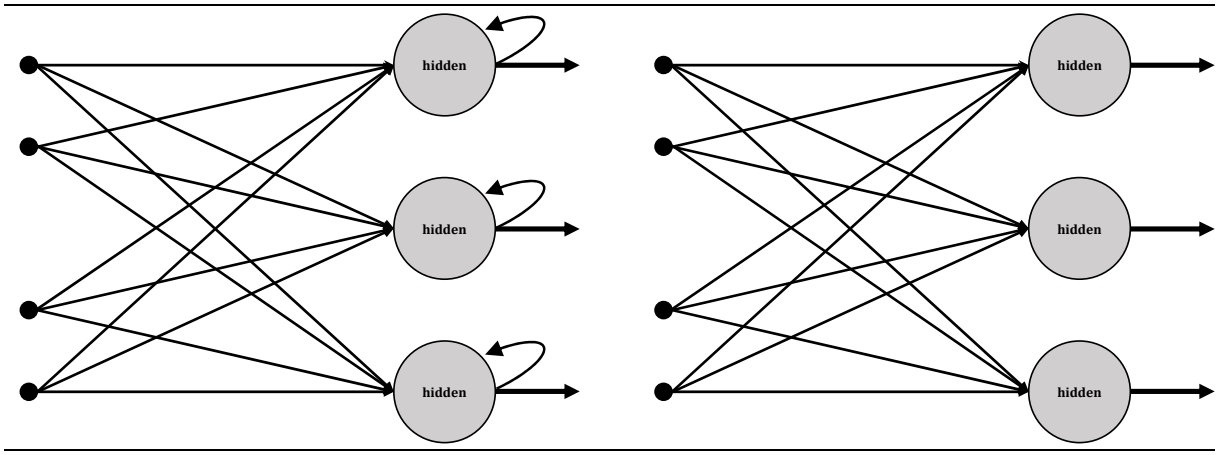
Allgemein bestehen KNN aus Neuronen, die in Schichten angeordnet und miteinander verbunden sind. KNN prozessieren dabei numerische Inputs mittels mathematischer Funktionen und geben als Ergebnis ebenfalls einen numerischen Output aus. Zwischen der Eingabe- und der Ausgabeschicht liegen verborgene Schichten, sogenannte *hidden layer* (SCHMIDHUBER 2015). Während des Lernprozesses haben KNN unterschiedliche Möglichkeiten sich an die vorliegende Problemstellung anzupassen. Dies geschieht durch Veränderungen einzelner Elemente des KNN, beispielsweise durch die Schaffung bzw. Löschung von Verbindungen zwischen Neuronen, die Neujustierung von Verbindungsgewichten oder die Hinzunahme bzw. Löschung ganzer Neuronen (HINTON 1992).

Ferner können Neuronen über sogenannte Aktivierungsfunktionen verfügen. Nicht-lineare Aktivierungsfunktionen sorgen dafür, dass Neuronen komplexe nicht-lineare Beziehungen zwischen Input und Output modellieren können. Die am häufigsten genutzten Aktivierungsfunktionen sind

unter anderem die binäre Schrittfunktion, die Sigmoid-Funktion, die hyperbolische Tangentenfunktion, verschiedene Formen der rectified linear unit Funktion oder die Softmax-Funktion (SHARMA et al. 2020).

Während FNN jeweils nur den aktuellen Input verarbeiten, verfügen RNN durch Verbindungen zwischen aufeinanderfolgenden Inputs über eine Art Kurzzeitgedächtnis. Somit werden bei der Verarbeitung des aktuellen Inputs auch die Outputs der vorherigen Datenpunkte betrachtet.

Abbildung 9: Informationsfluss in RNN (links) vs. Informationsfluss in FNN (rechts).



Quelle: Eigene Darstellung nach JURAFSKY und MARTIN (2019: 185).

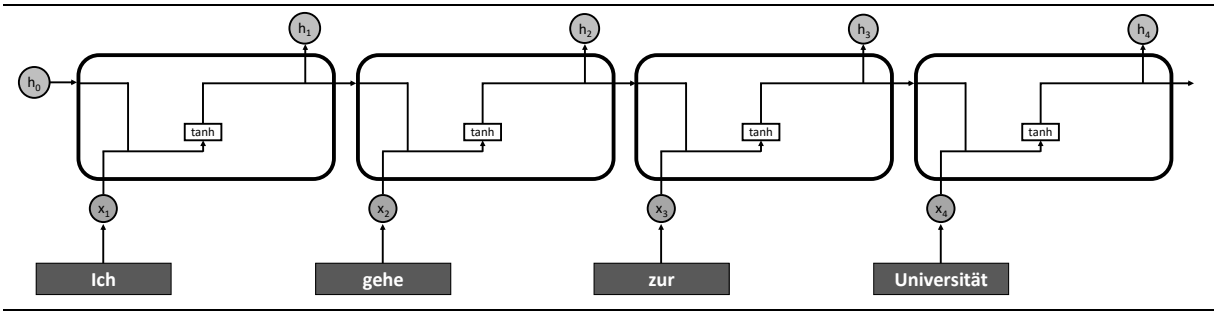
Damit können RNN Texte besser verarbeiten, da die Inputentitäten – beispielsweise die Wörter eines Satzes – nicht isoliert voneinander, sondern stets in Abhängigkeit der vorherigen Wörter betrachtet werden. Somit können RNN der Notwendigkeit Rechnung tragen, dass das gleiche Wort kontextabhängig eine unterschiedliche semantische Bedeutung haben kann bzw. bei Textübersetzungen ein Input-Wort in mehrere Output-Wörter übersetzt werden muss. Weiterhin sind RNN bedingt durch ihre Architektur in der Lage auch längere Inputsequenzen, wie mehrere Sätze oder ganze Dokumente, verarbeiten zu können (GOLDBERG 2017b).

Weiterentwicklungen der RNN verfügen über eine Encoder-Decoder-Verarbeitungseinheit, um Sequenzen unterschiedlicher Länge verarbeiten zu können (CHO et al. 2014). Diese Eigenschaft ist besonders für Textübersetzungen oder Frage-Antwort-Systeme relevant, da Inputs (z.B. die Wörter eines Satzes) sich nicht Wort für Wort in eine andere Sprache (Output) übersetzen lassen. Sequence to Sequence Modelle speichern daher über einen Encoder die Informationen der einzelnen Input-Elemente (z.B. Wörter) und formen diese zu einem Vektor um.

Der Decoder übersetzt anschließend die encodierten Informationen in den Output. Die Länge des Outputs ist dabei gänzlich unabhängig von der Länge des Inputs. Die Übersetzung eines Texts kann also aus mehr oder weniger Wörtern bestehen, als die Inputsequenz (SUTSKEVER et al. 2014). Abbildung 10 zeigt schematisch den Aufbau und den Informationsfluss innerhalb eines RNN-Encoders.

Jedes Wort der Inputsequenz wird durch einen Embedding-Layer in einen Wortvektor x_t übersetzt und mittels einer Aktivierungsfunktion (z.B. der hyperbolischen Tangentenfunktion) verrechnet (JOZEFOWICZ et al. 2015; LE et al. 2015). Zusätzlich zu dem jeweiligen Wortvektor der Inputsequenz x_t fließt in die Berechnung der sogenannte *hidden-state* h_t der vorherigen Wörter mit ein. Damit fungiert der *hidden state* als eine Art Kurzzeitgedächtnis, da bei der Verarbeitung eines Input-Tokens ebenfalls vorherige encodierte Input-Repräsentationen berücksichtigt werden.

Abbildung 10: Informationsfluss in einem ausgerollten RNN.



Quelle: Eigene Darstellung nach OLAH (2015).

Das Resultat dieser Berechnung ist ein neuer *hidden state*, der einerseits als Input für die Decoder-Einheit dient, andererseits gleichzeitig an das nächste Modul des Encoders überführt wird. Somit gelingt es mittels RNN den individuellen Kontexts eines Wortes bei der Textverarbeitung zu berücksichtigen.

Bedingt durch ihre Architektur sind RNN einige Nachteile inhärent. Diese liegen insbesondere in der Art und Weise begründet wie KNN während des Trainings lernen. Jedes Wort einer Inputsequenz wird auf Basis der vorherigen Wörter der Inputsequenz vorhergesagt. Nach jedem Trainingslauf wird die Vorhersage des KNN mit dem erwarteten Ergebnis verglichen. Die Stärke der Abweichung zwischen Vorhersage und erwartetem Ergebnis wird mittels einer Verlustfunktion berechnet. Der entstehende Fehlerwert wird anschließend genutzt, um die Gradienten der jeweiligen Neuronen - also die Anpassungen der Gewichte der jeweiligen Neuronen - zu berechnen (GOLDBERG 2017a).

Auf Basis der Gradienten werden dann die internen Gewichte des Netzes berechnet. Dies geschieht über einen Backpropagationmechanismus, der zunächst die Gewichte der letzten Verarbeitungsschichten neu justiert. Die Gradienten werden jeweils unter Berücksichtigung der Gradientenänderungen der vorherigen Schicht berechnet, sodass die Gewichtsjustierungen von Schicht zu Schicht exponentiell zu- bzw. abnehmen.

Bei RNN führt dieses Verhalten dazu, dass innerhalb der frühen Schichten eines Netzes keine Gewichtsjustierungen mehr stattfinden. Bezogen auf die Verarbeitung von Texten bedeutet dies, dass RNN insbesondere bei längeren Texten Abhängigkeiten zwischen Wörtern zu Beginn des

Texts mit später im Satz auftretenden Wörtern nicht erlernen können. Diese Problematik wird auch als Problem verschwindender bzw. explodierender Gradienten bezeichnet (HOCHREITER 1998; MIKOLOV et al. 2015; BENGIO et al. 1994). Anhand eines Beispiels zeigt sich wie sich das Problem der verschwindenden Gradienten konkret auf das Textverständnis eines RNN auswirkt:

Beispiel A: Ein Tag hat 24 […].

Beispiel B: Ich wurde in Deutschland geboren und habe dann lange Zeit in Amerika gelebt, daher ist meine Muttersprache […].

Bei kurzen Sätzen, wie in Beispiel A, treten in der Regel keine Probleme auf, da die Informationen nicht über besonders viele Wörter hinweg transportiert werden müssen. So lässt sich annehmen, dass das vorherzusagende Wort *Stunden* ist. Beispiel B zeigt hingegen einen potentiellen Problemfall für RNN. Hier steht das Wort, welches maßgeblich das vorherzusagende Wort bestimmt, am Anfang des Satzes. Aufgrund des Problems der verschwindenden Gradienten wäre es einem RNN nur schwer möglich eine Verbindung zwischen den Wörtern *Deutschland geboren* und dem vorherzusagenden Wort *deutsch* herzustellen.

Dies ist primär auf die Pfadlänge zurückzuführen, über welche die Informationen transportiert werden müssen. Des Weiteren werden alle Wörter des Inputs berücksichtigt und mit Wörtern am Ende des Inputs verbunden. Dieses Verhalten kann RNN ineffizient machen, da auch irrelevante Informationen prozessiert werden. Darüber hinaus können RNN aufgrund der sequentiellen Prozessierung nicht parallelisiert werden. Dadurch kann die Verarbeitung sehr zeitaufwendig sein (GOLDBERG 2017b).

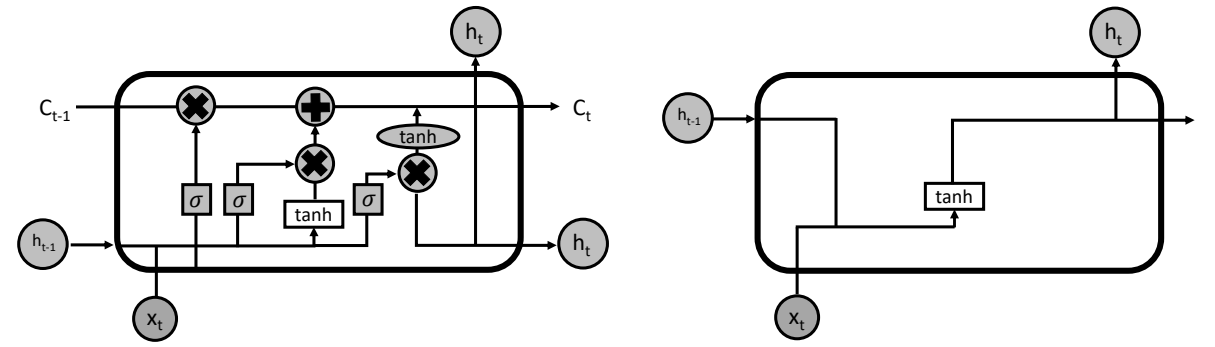
4.4 LSTM-Netzwerke

Um den bekannten Schwächen von RNN zu begegnen, wird an Modifikationen der RNN-Architektur geforscht. Neben *gated recurrent units* werden insbesondere sogenannte *long short-term memory* (LSTM)-Techniken eingesetzt (HOCHREITER und SCHMIDHUBER 1997). Diese erlangen Mitte der 2010er Jahre erhebliche Popularität, da mit rasant ansteigender Verfügbarkeit von Trainingsdaten und Rechnerleistung deutliche Leistungssteigerungen der LSTM-Modelle erzielt werden.

LSTM-Modelle werden in erster Linie entwickelt, um dem Problem der verschwindenden bzw. explodierenden Gradienten zu begegnen. Vereinfacht lässt sich sagen, dass LSTM-Modelle über ein langes Kurzzeitgedächtnis verfügen, sodass sie im Vergleich zu RNN einen umfassenderen Kontext bei der Textverarbeitung einbeziehen können. Abbildung 11 zeigt den internen Aufbau eines LSTM-Moduls auf der linken Seite im Vergleich zu dem eines RNN-Moduls auf der rechten Seite.

Wie alle RNN bestehen auch LSTM-Netzwerke aus einer Verkettung mehrerer miteinander verbundener Module. Während einfache RNN-Module eine einzelne neuronale Netzwerkschicht aufweisen, bestehen LSTM-Module aus insgesamt vier miteinander interagierenden Schichten, die es LSTM-Netzwerken ermöglichen auch längere Abhängigkeiten in sequentiellen Daten zu erkennen.

Abbildung 11: Architektur eines LSTM-Moduls (links) vs. eines RNN-Moduls (rechts).



Quelle: Eigene Darstellung nach Olah (2015).

Die zentrale Neuerung von LSTM- gegenüber RNN-Modulen ist der sogenannte *cell state*. Der cell state ist die zentrale Speicherinstanz von LSTM-Modulen und fungiert als eine Art Langzeitgedächtnis. Er sorgt dafür, dass bei der Anpassung der Gradienten während des Trainings relativ konstante Fehlerflüsse möglich sind (HOCHREITER und SCHMIDHUBER 1997).

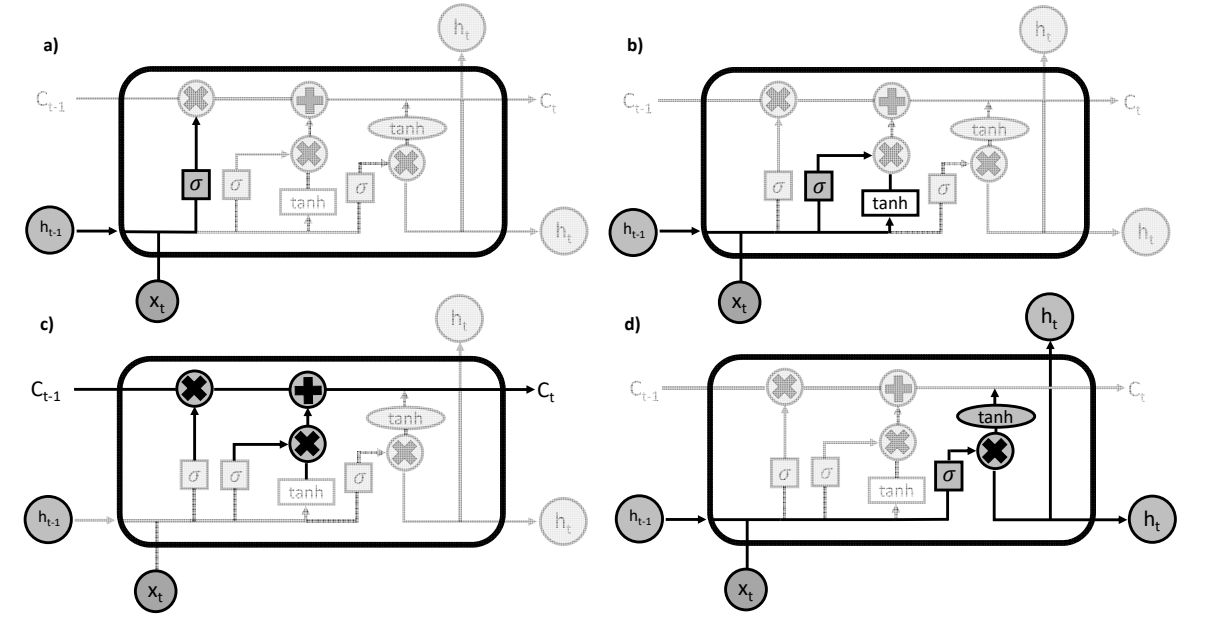
Im NLP-Kontext erhielten LSTM-Netzwerke durch diesen Aufbau die Möglichkeit spät auftretende Informationen einer Textsequenz mit früher aufgetretetenen Signalen zu verknüpfen. Weiterhin können während des Trainings irrelevante Informationen gelöscht werden, sodass die Berechnung nicht die Kapazitäten gängiger Rechner übersteigt. Abbildung 12 zeigt den Informationsfluss eines LSTM-Netzwerks.

Im ersten Schritt *a)* fließen sowohl der hidden state des vorherigen Moduls h_{t-1} als auch der aktuelle Input x_t als Inputs in das Modul hinein. In h_{t-1} sind sämtliche relevante Informationen über vorherige Datenpunkte der Textsequenz (z.B. Wörter eines Satzes) gespeichert. In x_t ist die numerische Repräsentation des aktuellen Inputs gespeichert (z.B. der Wortvektor). Die erste Verarbeitungsschicht ist eine Instanz, die entscheidet, wieviele der jeweiligen Informationen aus den beiden Inputs weiterverarbeitet werden sollen. Dieser Filterschritt wird mittels eine Sigmoid-Funktion operationalisiert. Das Ergebnis der Sigmoid-Funktion ist zwischen 0 und 1 normiert, sodass ein Wert von 0 bedeutet, dass sämtliche Informationen irrelevant sind und nicht weiter benötigt werden. Im Umkehrschluss sagt ein Wert von 1 aus, dass alle Informationen relevant sind und entsprechend beibehalten werden sollen.

Im nächsten zweistufigen Schritt *b)* wird mittels einer Sigmoid-Funktion entschieden, welche neuen Informationen in den neuen *cell state* C_t einfließen sollen. Anschließend werden mittels

einer hyperbolischen Tangentenfunktion die Kandidaten für den neuen *cell state* C_t ermittelt. Mit dem dritten Schritt *c)* erfolgt die Aktualisierung des cell state, indem die Informationen des alten *cell state* C_{t-1} mit den neuen Vektoren verrechnet werden.

Abbildung 12: Prozessverlauf innerhalb eines LSTM-Moduls.



Quelle: Eigene Darstellung nach Olah (2015).

Im letzten Schritt des Prozesses *d)* wird entschieden, welche Informationen des aktuellen *cell states* als Outputs an das nächste Modul der Kette weitergegeben werden sollen. Hierzu wird der aktuelle cell state mittels einer Sigmoid-Funktion gefiltert und anschließend durch eine hyperbolische Tangentenfunktionen aktualisiert (GOLDBERG 2017a, 2017b; OLAH 2015). Vereinfacht gesprochen erhalten LSTM-Netzwerke durch den internen Aufbau der einzelnen Module die Fähigkeit irrelevante Informationen zu „vergessen", besonders relevante Informationen über einen längeren Zeitraum zu speichern und weiterzugeben sowie vergangene Informationen mit aktuellen Inputs zu kombinieren.

Eine Abwandlung der LSTM-Netze stellen sogenannte Sequence-to-Sequence Modelle dar, wie sie von SUTSKEVER et al. (2014) vorgestellt wurden. Wie bereits aus dem Modellnamen hervorgeht, sind Sequence-to-Sequence Modelle besonders geeignet, um eine Inputsequenz in eine Outputsequenz zu übersetzen. Dies geschieht durch ein Kodierungsnetz, welches den Input in eine Vektordarstellung überführt. Diese Vektorrepräsentation wird über ein Dekodierungsnetz in die Outputsequenz überführt (SUTSKEVER et al. 2014). Bedingt durch den Modellaufbau sind Sequence-to-Sequence Modelle besonders für maschinelle Übersetzungssysteme relevant (WU et al. 2016).

4.5 Bi-LSTM-Netzwerke

Eine weitere Entwicklungsstufe von RNN stellen Bi-Directional LSTM (Bi-LSTM)-Modelle dar, die auf bidirektionalen RNN basieren (SCHUSTER und PALIWAL 1997). Mit der Entwicklung von Bi-LSTM-Modellen begegnet die Forschung einer zentralen Schwäche von unidirektionalen LSTM-Modellen (GRAVES und SCHMIDHUBER 2005). Diese können durch die sequentielle Verarbeitung zwar Informationen über längere Zeit hinweg transportieren. Allerdings geschieht dies ausschließlich in eine Richtung. Somit werden bei der Verwendung von LSTM-Modellen lediglich Informationen aus der Vergangenheit (beispielsweise. vorherige Wörter eines Satzes) genutzt, um die Zukunft (beispielsweise. das nächste Wort eines Satzes) vorherzusagen. Bi-LSTM verwenden zwei entgegengesetzte LSTM-Netze, die mit derselben Output-Schicht verbunden sind. Somit hat ein Bi-LSTM bei der Vorhersage eines Datenpunktes sowohl Informationen über vergangene als auch über zukünftige Sequenzen (GRAVES und SCHMIDHUBER 2005). Bezogen auf die Verarbeitung von Textsequenzen ließe sich also vereinfacht sagen, dass Bi-LSTM eine Textsequenz sowohl von vorne als auch von hinten lesen.

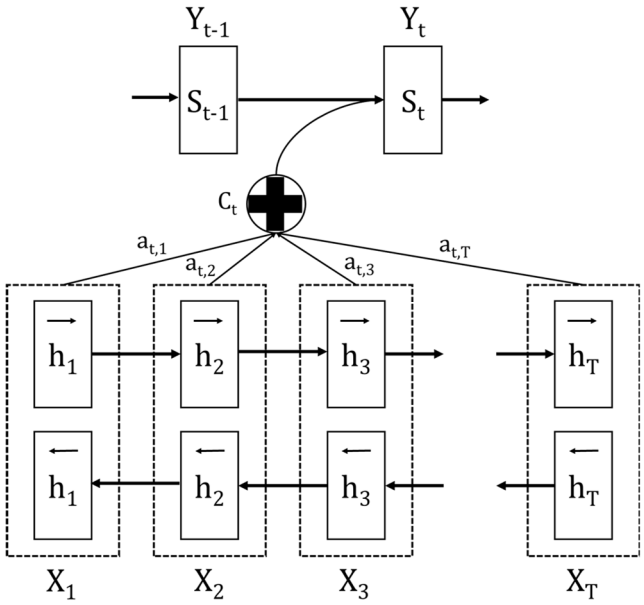
LSTM und Bi-LSTM sind trotz ihrer anspruchsvolleren Architektur weiterhin anfällig für das Problem des verschwindenden bzw. explodierenden Gradienten (SUTSKEVER et al. 2014). Durch die sequentielle Verarbeitung können Teile einer Textsequenz nicht gesondert betrachtet, sondern lediglich als komplette Vektordarstellung fixer Länge über eine Encoder-Decoder-Instanz verarbeitet werden. Neben der Problematik Langfristabhängigkeiten zu erlernen, haben Sequence-to-Sequence-Modelle das Problem, dass ihre Berechnung nicht parallel durchführbar ist. Somit ist die Prozessierungsdauer einer Textsequenz unmittelbar abhängig von der Länge der Textsequenz.

Eine bedeutende Weiterentwicklung von Bi-LSTM stellt die Implementierung von Attention-Mechansimen dar. Adaptierte Bi-LSTM zur Verarbeitung natürlicher Sprache werden erstmals von BAHDANAU et al. (2015) zur Textübersetzung entwickelt. In der Psychologie beschreibt der Begriff Attention bzw. Aufmerksamkeit die Fähigkeit des Menschen seine Konzentration auf bestimmte Inhalte zu lenken, während andere vollständig ignoriert werden (BREFCZYNSKI und DEYOE 1999). Dieses Konzept wird im Bereich des NLP auf die Art und Weise übertragen, wie neuronale Netze Texte verarbeiten sollen. Somit können Modelle, die Attention-Module einsetzen, für die Vorhersage eines Wortes frei wählen, welche anderen Wörter der Inputsequenz relevant sind und welche ignoriert werden können. Abbildung 13 zeigt die von BAHDANAU et al. (2015) vorgestellte Modellstruktur des Attention-Mechanisimus.

Für jedes Wort X_i der Inputsequenz X mit der Länge T wird im Encoderteil des Modells durch Anwendung eines Bi-LSTMs jeweils ein vorwärts- und ein rückwärtsgerichteter hidden state h_T berechnet. Somit werden in h_T Informationen vorheriger und folgender Wörter gespeichert, um den vollständigen Kontext eines Wortes X_T berücksichtigen zu können. Jeder hidden state h_t wird

nochmals mit einem Gewicht a_T versehen. Das Gewicht entspricht dem Alignment-Wert des jeweiligen hidden states. Der Alignment-Wert wird im Modell von BAHDANAU et al. (2015) mittels eines FNN mit einer Softmax-Aktivierungsfunktion bestimmt. Durch die Anwendung dieser Aktivierungsfunktion werden die Alignment-Werte so normiert, dass ihre Summe den Wert 1 ergibt. Durch diese Normierung werden die Alignment-Werte in Attention-Werte transformiert. Vereinfacht ausgedrückt misst der Attention-Wert, inwiefern das Eingabewort X_t zu dem Ausgabewort Y_t passt.

Abbildung 13: Aufbau des Attention-Mechanismus.



Quelle: Eigene Darstellung nach BAHDANAU et al. (2015: 3).

Somit bestimmt der Wert, welche Wörter der Inputsequenz von besonderer Relevanz für die Vorhersage des nächsten Wortes der Outputsequenz sind. Aus den mit den Attention-Werten gewichteten hidden states der Inputsequenz wird anschließend ein Kontextvektor berechnet. Zur finalen Vorhersage des nächsten Wortes der Outputsequenz Y_t dienen somit dreierlei Informationsquellen neben dem Kontextvektor C_t, der hidden state des vorherigen Wortes der Outputsequenz S_{t-1} sowie das vorherige prognostizierte Wort Y_{t-1} berücksichtigt.

4.6 Transformermodelle

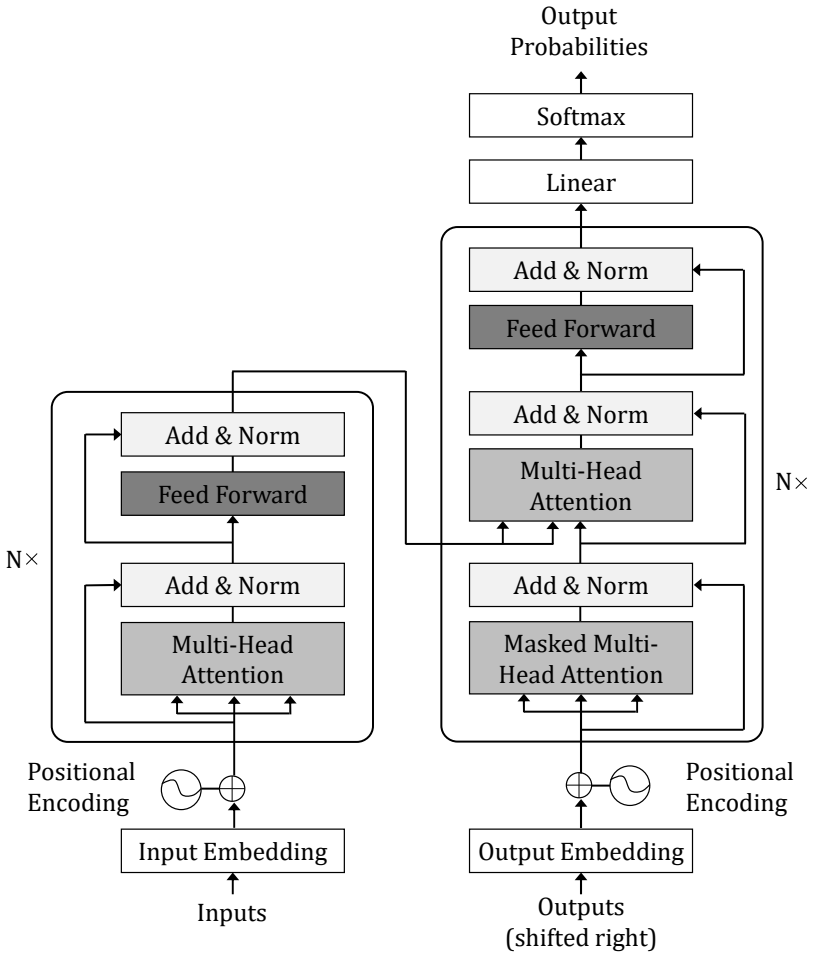
Wie aus den bisher beschriebenen Entwicklungen der NLP-Modelle zu entnehmen ist, stellen sequentielle Modelle in unterschiedlichen Ausbaustufen (RNN, LSTM, Bi-LSTM, Attention-Mechanismen) seit der Etablierung dichter Word-Embeddings den aktuellen Stand der Technik dar.

VASWANI et al. stellen 2017 eine gänzlich neue Modellarchitektur zur maschinellen Übersetzung von Text bzw. zur Sprachmodellierung vor. Ihr Beitrag „Attention is all you need“ markiert in der NLP-Literatur bis heute einen Meilenstein und schafft die Basis für die Entwicklung der Transfor-

mermodelle, die bis heute den aktuellen Stand der Technik des NLPs repräsentieren. Die Autor:innen entwickeln eine neue Architektur neuronaler Netze, die gänzlich auf sequentielle Strukturen im Encoderteil verzichtet und stattdessen drei verschiedene Adaptionen von Attention-Mechanismen einsetzt. Diese neue Modellarchitektur erzielt schnell bessere Ergebnisse als etablierte sequentielle Modelle und das bei geringerem Rechenaufwand.

Zwar wurden Transformermodelle ursprünglich zur Text- bzw. Sprachübersetzung entwickelt. Es stellt sich jedoch heraus, dass Weiterentwicklungen der Modelle ebenso zur Lösung unterschiedlichster NLP-Probleme eingesetzt werden können (DEVLIN et al. 2019; MCCANN et al. 2018). Entsprechend entwickeln sich die Transformermodelle schnell zur meist genutzten Modellarchitektur zur Lösung von NLP-Problemen (WOLF et al. 2020). Abbildung 14 zeigt den vollständigen Aufbau des vorgestellten Transformermodells.

Abbildung 14: Modellarchitektur des Transformermodells.



Quelle: VASWANI et al. (2017: 3).

Das Transformermodell besteht grundlegend aus zwei miteinander verbundenen Komponenten: dem Encoder auf der linken Seite und dem Decoder auf der rechten Seite des Modells. Der En-

coderteil verarbeitet den Input (z.B. ein zu übersetzender Satz), während der Decoderteil den Output (z.B. die Übersetzung des Inputs) erzeugt. Das von VASWANI et al. (2017) vorgestellte Modell verwendet insgesamt sechs gestapelte verbundene Encoder und Decoder (Nx=6). Die Encoder bestehen aus einer Multi-Head-Attention-Schicht, gefolgt von einer Add & Norm-Schicht sowie einer Feed-Forward-Schicht, der ebenfalls eine Add & Norm Schicht folgt.

Der Decoderteil umfasst eine Masked Multi-Head-Attention-Schicht, eine Multi-Head-Attention-Schicht sowie eine Feed-Forward-Schicht, die jeweils mit einer nachgelagerten Add & Norm-Schicht verbunden sind. Zwischen dem Encoder und der Multi-Head-Attention-Schicht besteht eine Verbindung, welche den Informationsfluss zwischen den beiden Komponenten sicherstellt. Oberhalb des Decoders sitzen noch eine lineare und Softmax-Schicht. Während RNNs aufgrund ihrer sequentiellen Architektur Inputsequenzen schrittweise verarbeiten, verläuft die Prozessierung einer Inputsequenz mit Transformermodellen in einem einzigen Schritt ab. Die Inputsequenz wird durch den Encoder prozessiert, während die bis dahin prozessierte Outputsequenz in dem Decoder verarbeitet wird. Die Informationen des Encoders und des Decoders werden kombiniert und aus dieser Kombination ergibt sich jeweils eine Vorhersage des nächsten Tokens in der Textsequenz.

4.6.1 Funktionsweise des Encoderblocks eines Transformermodells

Bevor die einzelnen Wörter in den Encoder bzw. den Decoder einfließen, werden sie durch ein vortrainiertes Word-Embedding zu Wortvektoren umgewandelt (vgl. Kapitel 4.2). Da Transformermodelle nicht sequentiell arbeiten, beinhalten die Word-Embeddings keinerlei Kontextinformationen zu anderen Wörtern innerhalb der Inputsequenz. Daher müssen die Wortvektoren im nächsten Schritt mit Positionsinformationen (Positional Encoding) angereichert werden. Somit fließen neben den numerischen Repräsentationen der einzelnen Wörter in Form der Word-Embeddings auch Informationen über ihre Position innerhalb der Inputsequenz in Form von Positionsvektoren in das Transformermodell.

Diese Informationen sind insbesondere für Wörter relevant, die in Abhängigkeit von ihrem Kontext unterschiedliche semantische Bedeutungen haben können. Neben den semantischen Informationen einzelner Wörter der Inputsequenz (in Form von Word-Embeddings) und den Kontextinformationen der einzelnen Wörter der Inputsequenz (in Form von Positionsvektoren) nutzen Transformermodelle ferner eine Gewichtung zur Bestimmung der Relevanz anderer Wörter innerhalb der Inputsequenz für die Prozessierung des aktuellen Wortes. Dieser Gewichtungs- und Selektionsprozess wird in der Fachliteratur als Attention-Mechanismus beschrieben und wurde wie bereits erläutert von BAHDANAU et al. (2015) vorgestellt. Attention-Mechanismen sorgen dafür, dass einzelne Sätze respektive Absätze in eine möglichst präzise numerische Repräsentation

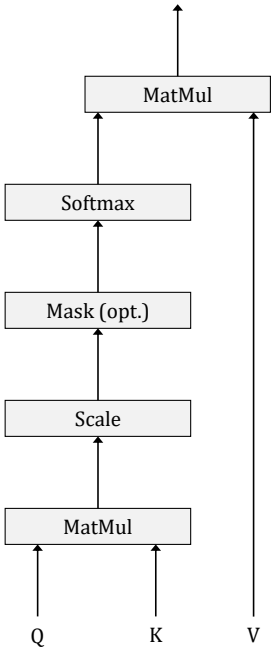
überführt werden können, die sowohl semantische als auch syntaktische Merkmale berücksichtigt.

4.6.2 Funktionsweise von Attention-Mechanismen in Transformermodellen

VASWANI et al. (2017) stellen mit dem Self-Attention-Mechanismus eine abgewandelte Version des Attention-Mechanismus von BAHDANAU et al. (2015) vor. Das Konzept der Self-Attention ist ein Schlüsselelement von Transformermodellen. Das Konzept wird im Folgenden näher beleuchtet und in Abbildung 15 schematisch dargestellt.

Die positionsindexierten Word-Embeddings werden im ersten Schritt des Attention-Prozesses mit drei Gewichtsmatritzen (W) multipliziert. Durch diese Multiplikation entsteht je Wortrepräsentation drei neue Vektoren: ein Query-Vektor (Q), ein Key-Vektor (K) und ein Value-Vektor (V).

Abbildung 15: Prozessablauf des Self-Attention-Mechanismus.



Quelle: Eigene Darstellung nach VASWANI et al. (2017: 4).

Der Query-Vektor ist der Vektor für das Wort, für das die Relevanz der anderen Wörter der Inputsequenz abgefragt werden soll. Die Key-Vektoren sind komplementär dazu Vektoren, der Wörter, die über den Query-Vektor abgefragt werden. Um die Relevanz der anderen Wörter der Inputsequenz für das aktuell betrachtete Wort bestimmen zu können, werden sogenannte Attention-Scores berechnet. Diese werden durch die Bildung des Skalarprodukts zwischen den Query-Vektoren und den Key-Vektoren gebildet. Die Attention-Scores sind somit ein Maß für die Ähnlichkeit der hinter den Vektoren stehenden Wörter. Ähnlichere Wörter erhalten höhere Werte, als unähnlichere Wortkombinationen. Zur Leistungsoptimierung findet dieser Schritt innerhalb des Transformermodells im Rahmen einer Matrizenmultiplikation (MatMul) und nicht anhand einzelner Vektoren statt.

Anschließend werden die berechneten Attention-Scores skaliert, wobei es sich hierbei lediglich um eine Dimensionsreduktion der Matrizen handelt, die zur Leistungssteigerung des Modells beiträgt. Falls kürzere Textsequenzen als die längstmögliche (512 Token) prozessiert werden sollen, folgt anschließend eine Maskierung – also ein Auffüllen mit Platzhaltern – der Inputsequenz.

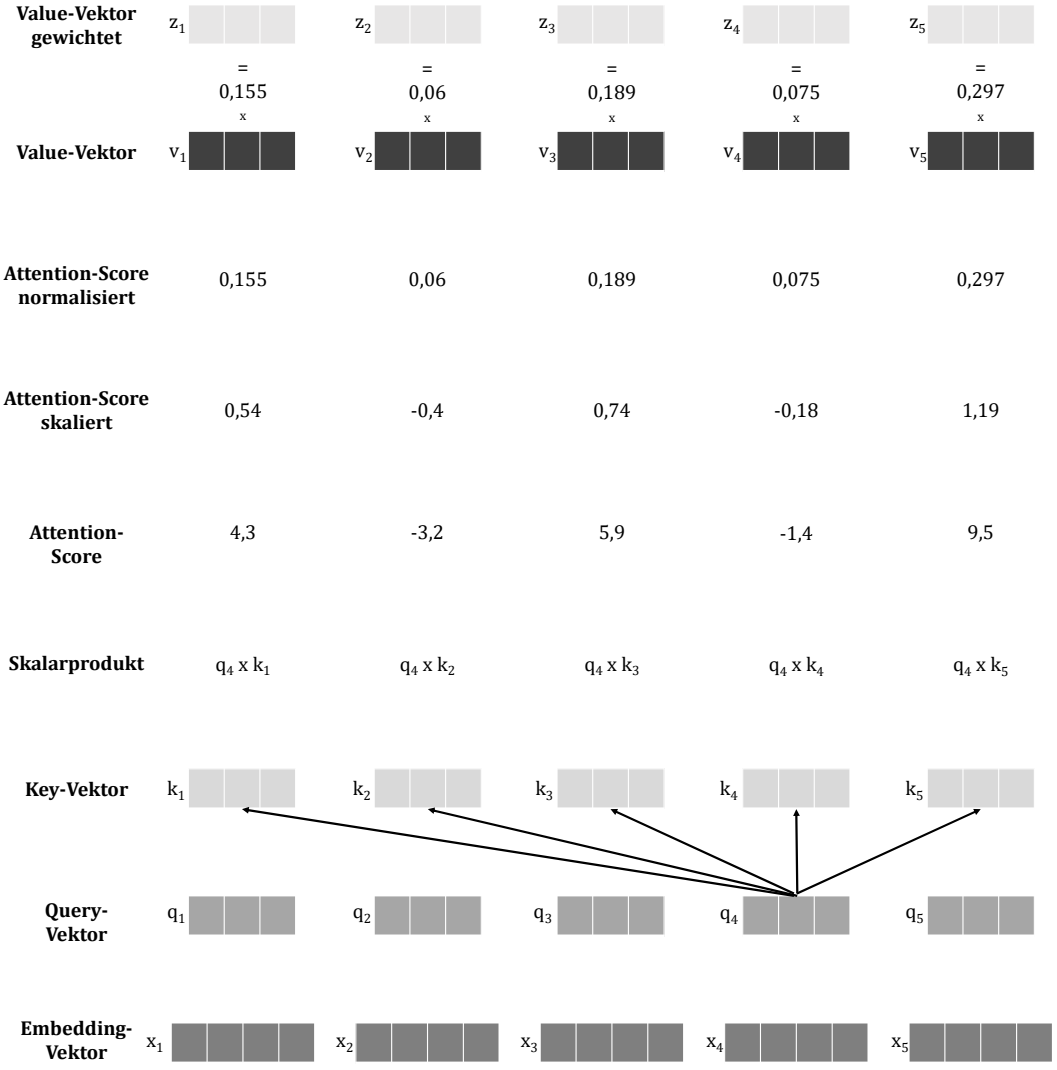
Die berechneten und gegebenfalls maskierten Attention-Scores werden anschließend durch Anwendung der Softmax-Funktion normalisiert. Nach der Normalisierung der Werte nehmen diese ausschließlich positive Ausprägungen an und ergeben in der Summe den Wert 1. Somit kann quantifiziert und verglichen werden, wie relevant die verschiedenen Wörter der Inputsequenz für das aktuell prozessierte Wörter sind. Je höher die normalisierten Attention-Scores ausfallen, desto relevanter ist ein Wort für das aktuell prozessierte Wort.

Im vorletzten Schritt des Self-Attention-Prozesses kommen die Value-Vektoren zum Einsatz. Diese repräsentieren ebenfalls die einzelnen Wörter der Inputsequenz und werden mit dem normalisierten Attention-Score der jeweiligen Wörter multipliziert, sodass gewichtete Value-Vektoren entstehen. Vereinfacht gesprochen geben die Value-Vektoren für jedes Wort an, wie wichtig jedes andere Wort der Inputsequenz für die semantische Einbettung des betrachteten Wortes ist.

Im letzten Schritt des Self-Attention-Mechanismus werden die gewichteten Value-Vektoren aufsummiert. Es entsteht somit ein neuer Vektor, welcher als kontextualisierte Repräsentation des aktuell prozessierten Wortes beschrieben werden kann. Dieser kontextualisierte Vektor enthält neben den a priori bekannten Positions- und Bedeutungsinformationen ebenso Informationen über die Relevanz anderer Wörter der Inputsequenz für das mit dem Vektor verknüpfte Wort. Folglich ist der entstandene Wortvektor individuell an die vorliegende Inputsequenz angepasst. Abbildung 16 fasst die beschriebenen Rechenschritte für das vierte Wort einer Inputsequenz zusammen.

Ausgangspunkt stellen die mit Positionsinformationen angereicherten Word-Embeddings dar. Die Query-, Key- und Valuevektoren werden durch Multiplikation mit während des Trainings erzeugten Gewichtsmatrizen berechnet. Anschließend wird aus dem Query-Vektor und den Key-Vektoren das Skalarprodukt berechnet, welches die Ähnlichkeit des Query-Vektors mit dem jeweiligen Key-Vector quantifiziert. Der entstandene Attention-Score wird skaliert, indem er durch den Wert 8 geteilt wird. Dieser Wert entspricht der Quadratwurzel der Dimensionalität der Key-Vektoren. Diese Dimensionsreduktion wird primär zu Zwecken der Laufzeitoptimierung angewandt und hat keine inhaltlichen Auswirkungen. Die skalierten Attention-Scores werden mittels der Softmax-Funktion normalisiert und anschließend mit den Value-Vektoren multipliziert.

Abbildung 16: Berechnung der Attention-Scores.



Quelle: Eigene Darstellung nach KRÜGER (2021: 301).

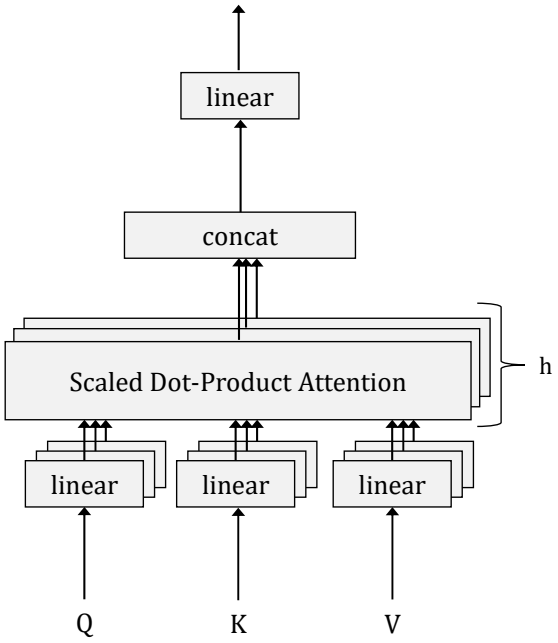
4.6.3 Multi-Head-Attention-Mechanismen in Transformermodellen

Im Encoderblock des Transformermodells kommt ein Mechanismus namens Multi-Head-Attention zum Einsatz. Dies ist in Abbildung 14 (Kapitel 4.6) dargestellt. Dabei handelt es sich um eine Parallelisierung des Self-Attention-Mechanismus, sodass jedem der parallel laufenden Attention-Prozesse ein proprietärer Attention-Head zugeordnet ist. Abbildung 17 illustriert das Konzept der Multi-Head-Attention.

Als Inputs werden die bereits beschriebenen Query-, Key- und Valuevektoren in *h* Teile geteilt und in einer linearen Schicht mit den Gewichtsmatrizen multipliziert. Im vorgestellten Modell von VASWANI et al. (2017) verläuft dieser parallelisierte Prozess insgesamt achtmal (*h*=8). Da sich die aufgespaltenen Vektoren jeweils voneinander unterscheiden, entstehen jeweils eigene Vektor-

räume und schlussendlich unterschiedliche Attention-Scores. Die finalen Vektoren der acht Attention-Heads werden in einer weiteren Schicht verkettet und in einer linearen Schicht durch Multiplikation mit einer weiteren Gewichtsmatrix in den finalen Vektor überführt.

Abbildung 17: Multi-Head-Attention-Mechanismus.



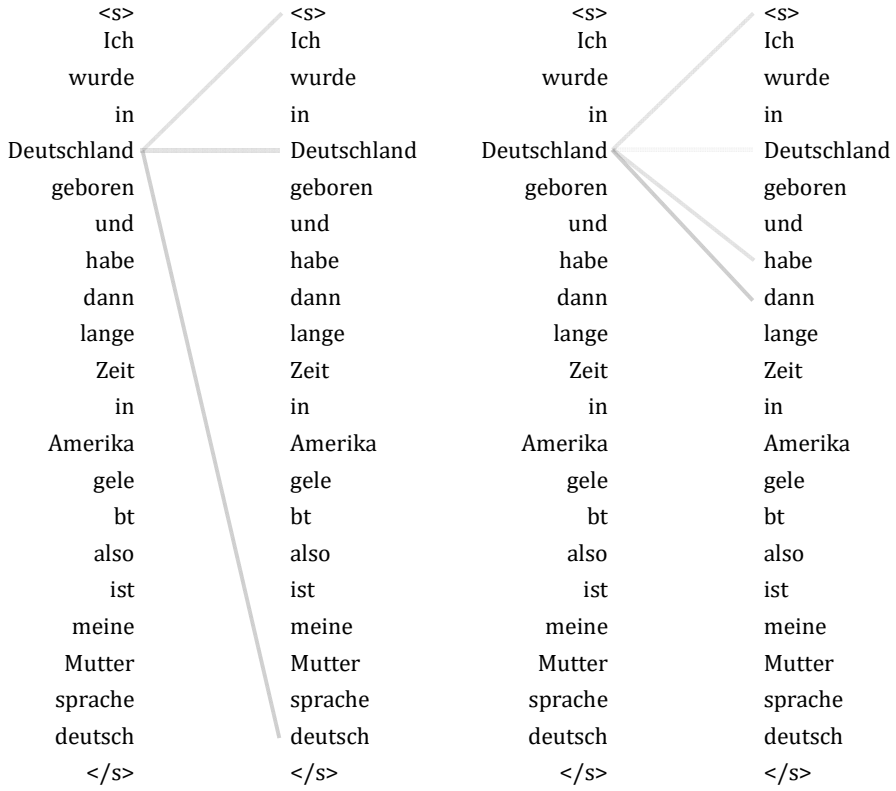
Quelle: Eigene Darstellung nach VASWANI et al. (2017: 4).

Multi-Head-Attention wird primär aus zweierlei Gründen angewendet. Einerseits sorgt die parallele Verarbeitung aufgeteilter Vektoren für Effizienzgewinne in der Verarbeitungsgeschwindigkeit. Andererseits werden durch die parallele Berechnung unterschiedlicher Attention-Scores mehr und unterschiedliche semantische und syntaktische Verflechtungen der einzelnen Wörter der Inputsequenz aufgedeckt (VASWANI et al. 2017). Im letzten Teil des Encoderblocks werden die Outputs des Multi-Head-Attention-Mechanismus in einem FNN verarbeitet, welches die kontextualisierten Vektoren nochmals optimiert.

Anschließend an die Multi-Head-Attention-Schicht und den FNN laufen, wie in Abbildung 14 zu sehen, weitere Verbindungen in eine Add & Norm-Schicht. Bei diesen Verbindungen handelt es sich um Residualverbindungen, welche unverarbeitete Kopien der jeweiligen Matrizen an den nachgelagerten Schichten vorbeiführen. Diese werden in tiefen neuronalen Netzen eingesetzt, um Effizienzgewinne beim Training dieser auszulösen (KOEHN 2020). Die unverarbeiteten sowie die verarbeiteten Matrizen werden jeweils in der Add & Norm-Schicht zusammengeführt und normalisiert. Dieser Vorgang geschieht ebenfalls lediglich zur Optimierung des Trainingsprozesses (BA et al. 2016). Hat die zu verarbeitende Inputsequenz alle sechs Encoderblöcke durchlaufen, wird die entstandene numerische Repräsentation der Inputsequenz an den Decoder überführt.

Abbildung 18 veranschaulicht beispielhaft die Aufmerksamkeitsverteilung zweier Attentionheads und Layer. Für die Berechnung der Aufmerksamkeitsverteilung wurde das vortrainierte Transformermodell *xlm-roberta-base* verwendet. Dieses wurde von den Autor:innen auf 2,5 Terrabyte Textdaten aus dem CC vortrainiert und ist in über 100 Sprachen verfügbar (LIU et al. 2019). Die Attentionscores wurden mit dem Python-Paket *BERTViz* berechnet (VIG 2019).

Abbildung 18: Beispiel für normalisierte Attention-Scores.



Quelle: Eigene Darstellung nach VIG (2019).

Wie bereits erläutert arbeitet sowohl das Transformermodell als Ganzes als auch der Attention-Prozess stark parallel. Daher zeigt Abbildung 18 lediglich die Aufmerksamkeitsverteilung eines Bruchteils des Gesamtmodells und die Aufmerksamkeitsverteilung kann in einem anderen Layer bzw. Attention-Head vollkommen anders ausfallen. Transformermodelle zerlegen zusammengesetzte Wörter wie beispielsweise *Muttersprache* (*Mutter* und *Sprache*) nochmals in Subwörter bzw. character n-grams (vgl. Kapitel 4.2.3). Somit muss nicht für jedes einzelne Wort eine eigene Repräsentation erlernt werden. Die Bedeutung zusammengesetzter Wörter wird durch die Kombination der bekannten Subwörter erlernt.

Die Visualisierung der Attention-Scores veranschaulicht zwar die grundlegende Arbeitsweise von Transformermodellen. Nichtsdestotrotz zeigen einige Studien, dass die Betrachtung von Attention-Scores keine Erklärung der Wortvorhersage darstellt (VASHISHTH et al. 2019; JAIN und WALLACE 2019; WIEGREFFE und PINTER 2019).

Im Beispiel auf der linken Seite der Abbildung liegen für das betrachtete Wort *Deutschland* besonders große Aufmerksamkeitsgewichte auf den Wörtern *deutsch* und *Deutschland* sowie auf den Satzbeginntoken (*<s>*). Somit legt das Modell in diesem Fall besondere Aufmerksamkeit auf das Adjektiv *deutsch*. Es wird also deutlich, dass Transformermodelle in der Lage sind komplexe Beziehungen zwischen Wörtern auch über größere Abstände zu erlernen. In diesem Beispiel hilft die Verknüpfung zwischen *Deutschland* und *deutsch,* um die Bedeutung der gesamten Inputsequenz korrekt greifen zu können. Insbesondere für die Vorhersage des letzten Tokens *deutsch* stellt die Verbindung zu dem Token *Deutschland* eine entscheidende Informationsgrundlage dar. CLARK et al. (2019) zeigen auf, dass Transformermodelle häufig Aufmerksamkeit auf die Satzbeginn- und Satzendtoken legen. Die Autor:innen vermuten, dass dieses Verhalten eintritt, wenn in diesem Schritt des Attentionmechanismus keine weitere Beachtung auf andere Wörter gelegt werden soll (CLARK et al. 2019).

Für das rechtsstehende Beispiel legt das Modell für das betrachtete Wort *Deutschland* besonders viel Aufmerksamkeit auf die folgenden Wörter *habe* und *dann* sowie *Deutschland* und das Satzbeginntoken (<s>). Nach diesem Schema werden in verschiedenen parallel verlaufenden Prozessen unterschiedliche Aufmerksamkeitsgewichte für die Kontextualisierung der einzelnen Wörter einer Inputsequenz berechnet. Durch die mehrfache Berechnung können die unterschiedlichsten Wechselwirkungen zwischen den Wörtern und Subwörtern der Inputsequenz berücksichtigt und somit eine möglichst vollständige Repräsentation der Semantik und Syntax erzeugt werden.

4.6.4 Funktionsweise des Decoderblocks in Transformermodellen

Der finale Output des Encoderblocks - also die numerische Repräsentation der Inputsequenz in Form von kontextualisierten Vektoren – wird anschließend in den Decoderblock des Transformermodells übergeben. Während der Encodierungsprozess wie bereits erläutert stark parallelisiert verläuft, arbeitet der Decoder analog zu RNN sequenziell. Basierend auf den encodierten Repräsentationen der Inputsequenz übersetzt der Decoder diese Wort für Wort in die Zielsprache. Dabei wird für die Vorhersage des nächsten Wortes neben der encodierten Inputsequenz auch die bis dahin decodierte Sequenz betrachtet.

Wie Abbildung 14 zeigt, gleichen sich viele Elemente des Decoder- und des Encoderblocks. So arbeitet der Decoderblock ebenfalls mit vortrainierten Word-Embeddings, welche mit Positionsinformationen (Positional Encoding) angereichert werden. Die hieraus entstehenden Query-, Key- und Valuevektoren werden anschließend von einem Masked Multi-Head-Attention-Mechanismus weiterverarbeitet. Dieser Mechanismus ist in der Trainingsphase von Transformermodellen relevant. Transformermodelle werden zur Textübersetzung mit großen bilingualen Textkorpora vortrainiert. Aufgabe des Transformermodells ist es während des Trainings, basierend auf der encodierten Inputsequenz und den bisher decodierten Wörtern, das nächste Wort vorherzusagen.

Durch diese umfassenden Informationen sowie die parellele Prozessierung könnten Transformermodelle während des Trainings theoretisch das vorherzusagende Wort aus den Trainingsdaten kopieren und würden in diesem Fall keine generalisierbaren Fähigkeiten erlernen. Daher werden für das Training alle folgenden Wörter der Outputsequenz maskiert, also verdeckt. Somit läuft der Multi-Head-Attention-Mechanismus im Decoderteil analog zu dem des Encoderteils ab. Der einzige Unterschied besteht darin, dass die skalierten Attention-Scores zukünftiger Wörter der Inputsequenz auf - ∞ gesetzt sind. Somit nehmen diese nach der Softmax-Normalisierung die Werte 0 an und werden damit bei der Kontextualisierung der Word-Embeddings nicht berücksichtigt.

Im nächsten Schritt kommt ein weiterer Multi-Head-Attention-Mechanismus zum Einsatz. Auch hier existiert im Vergleich zum Ablauf im Encoderblock ein Unterschied. Während die Query-vektoren aus der bisher prozessierten Sequenz des Decoders stammen, kommen die Key- und Valuevektoren aus dem Encoderteil des Transformermodells. Durch diese Encoder-Decoder-Attention sowie auch Cross-Attention (BAHDANAU et al. 2015; WU et al. 2016) wird die Relevanz der Wörter der Inputsequenz (in der Ausgangssprache) für Vorhersage des nächsten Wortes der Outputsequenz (in der Zielsprache) berechnet. Für die Textübersetzung ist diese Verknüpfung von Informationen aus der Inputsequenz mit Informationen der bisher erstellten Outputsequenz ein zentraler Schritt, um das nächste Wort der Outputsequenz vorherzusagen.

Die weitere Verarbeitung innerhalb der Multi-Head-Attention gleicht der des Encoderteils. Gleiches gilt für die zwischengeschalteten Add & Norm-Schichten. Die letzte lineare Schicht des Decoderblocks transformiert den Decoderoutput in einen Vektor, dessen Länge dem Umfang des Vokabulars der Trainingsdaten entspricht. Der Vektor wird abschließend in einer Softmax-Schicht normalisiert, sodass letztlich eine Wahrscheinlichkeitsverteilung des Vokabulars entsteht. Jedes Wort, das in den Trainingsdaten enthalten ist, wird mit einer Wahrscheinlichkeit annotiert. Je höher die Wahrscheinlichkeit, desto besser passt das Wort syntaktisch und semantisch in die Outputsequenz. Das Wort mit der höchsten Wahrscheinlichkeit wird ausgewählt und fließt als Zusatzinformation in den Decoderblock zur Vorhersage des nächsten Wortes.

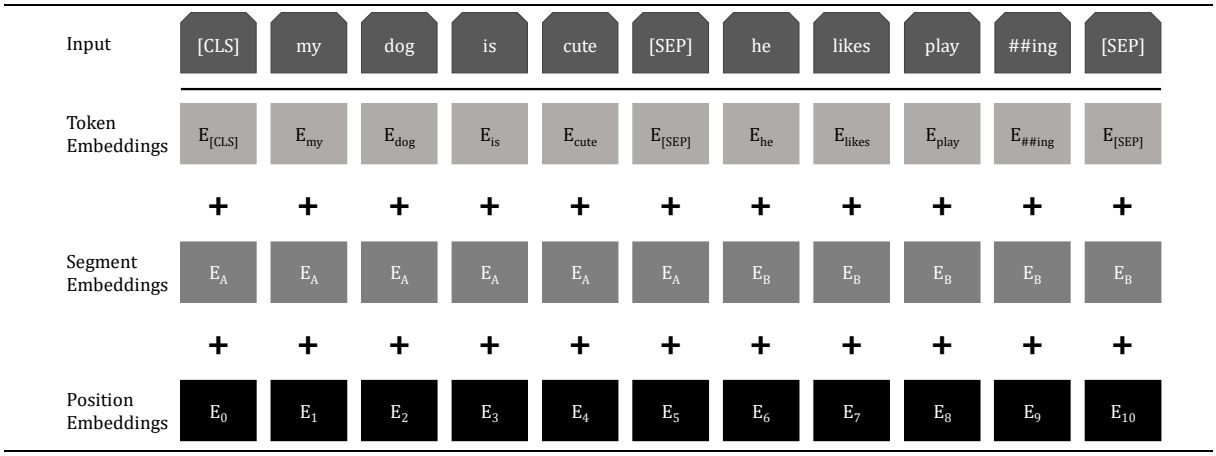
4.7 Bidirectional Encoder Representations from Transformers (BERT)

Als Weiterentwicklung des ursprünglichen Transformermodells stellen DEVLIN et al. (2019) das BERT-Modell vor. Die Transformermodelle von VASWANI et al. (2017) werden vorrangig zur Textübersetzung bzw. zur Sprachmodellierung entwickelt. Die Kernaufgabe solcher Sprachmodelle ist es, innerhalb eines Satzes jeweils das nächste Wort vorzusagen und somit Text zu produzieren. Dementsprechend verläuft das Training von Transformermodellen innerhalb eines Satzes unidirektional (von links nach rechts) ab, da auf Basis vorheriger Wörter das nächste Wort vorhergesagt werden soll. Das BERT-Modell greift grundlegend die Modellarchitektur von Transformermodellen auf, verfolgt jedoch einen bidirektionalen Trainingsansatz. Während des Trainings

wird, anders als bei unidirektionalen Transformermodellen, nicht das jeweils nächste, sondern ein zufälliges Wort des Satzes maskiert.

Anschließend wird das Modell trainiert, um in den Trainingssätzen das jeweilig maskierte Wort zu prognostizieren. Für die Vorhersage fließen sämtliche Informationen des Satzes in das Modell mit ein, um ein bidirektional vortrainiertes Sprachmodell zu erhalten. Neben der Fähigkeit maskierte Wörter in Sätzen vorhersagen zu können, verfolgt das Training noch ein zweites Ziel, welches besonders wichtig für die Verarbeitung vollständiger Textsequenzen und Dokumente ist. BERT erlernt die Fähigkeit vorherzusagen, ob zwei Sätze aufeinander folgen oder nicht. Somit erhält das Sprachmodell die Fähigkeit, neben den Abhängigkeiten von Wörtern innerhalb eines Satzes auch die Abhängigkeiten zwischen mehreren Sätzen zu verstehen. Abbildung 19 stellt die Input-Embeddings eines BERT-Modells dar.

Abbildung 19: Input-Embeddings des BERT-Modells.



Quelle: Eigene Darstellung nach DEVLIN et al. (2019: 5).

. Die Token-Embeddings umfassen die Wortvektoren der einzelnen Tokens. Dabei können Wörter, wie in unidirektionalen Transformermodellen auch, nochmals in character n-grams (vgl. Kapitel 4.2.3) zerlegt werden. Eine Zerlegung eines Wortes in mehrere Wortteile wird mit den „##“ Symbolen signalisiert. Im Beispiel von Abbildung 19 wird das Wort „playing“ in die N-Gramme „play“ und „ing“ aufgeteilt. Somit können aus der Kombination der Wortvektoren für das Wort „play“ und dem N-Gramm „ing“ ein neues Wort gebildet werden. Durch diese Eigenschaft gelingt es BERT Wörter, die nicht in den Trainingsdaten vorkommen, zu verstehen und den Umfang des Vokabulars zu reduzieren. Weiterhin nutzt das BERT-Modell Segment-Embeddings, die signalisieren, welchem Satz die jeweiligen Tokens angehören. Drittens werden Position-Embeddings genutzt, um Informationen über die Reihenfolge der Tokens in das Modell einfließen lassen zu können. Wie aus Abbildung 19 hervorgeht, sind die Input-Embeddings eines BERT-Modells die Summe der Token-Embeddings, der Segment-Embeddings und der Position-Embeddings.

Das von DEVLIN et al. (2019) vorgestellte Modell wurde mittels unüberwachtem Training auf Basis des 800 Millionen Wörter umfassenden BooksCorpus und des 2,5 Milliarden Wörter umfassenden englischsprachigen Wikipedia vortrainiert (DEVLIN et al. 2019). Somit erhält BERT ein sehr profundes, grundlegendes Verständnis von Sprache, sodass das Modell auf unterschiedlichste Problemstellungen relativ zeitunaufwendig feinabgestimmt werden kann. Die Autor:innen zeigen, dass BERT in verschiedensten NLP-Aufgaben durchschnittlich um 7 % besser abschneidet, als das bis dahin führenden Modell OPENAI GPT (RADFORD et al. 2018). Sie zeigen weiterhin auf, dass die beiden Veränderungen im Trainingsablauf deutliche Auswirkungen auf die Performanz des Modells haben. Beispielsweise schneidet dasselbe Modell beim Training von Frage-Antwort-Systemen durch unidirektionales Training und das Auslassen der Vorhersage des Folgesatzes über 10 % schlechter ab als das bidirektionale Modell.

Auf Basis der Transformerarchitektur entwickeln sich stetig neue und performantere Modelle, wie beispielsweise das sogenannte Text-to-Text Transfer Transformer Modell (T5) (RAFFEL et al. 2019) oder Generative Pre-trained Transformer 3 (BROWN et al. 2020). Insbesondere diese modernen Adaptionen des Transformermodells erreichen in Benchmarkstudien außergewöhnliche Ergebnisse. Zur Evaluation von NLP-Modellen wird häufig eine kombinierte Metrik verwendet, die die Leistungsfähigkeit unterschiedlicher NLP-Aufgaben in einer Zahl ausdrückt. Solch eine kombinierte Metrik namens GLUE (General Language Understanding Evaluation) stellen 2019 WANG et al. vor. Allerdings erreichen modernste NLP-Modelle extrem schnell Höchstwerte, sodass im selben Jahr eine neue Metrik mit komplexeren Aufgaben vorgestellt wird (SUPERGLUE) (WANG et al. 2019a). Ein adaptiertes BERT-Modell übertrifft im Januar 2021 erstmals den menschlichen Referenzwert von 89,8 Punkten um 0,1 Punkte (HE et al. 2021). Im Februar 2022 stellen ZOPH et al. (2022) ein adaptiertes Transformermodell vor, welches im SUPERGLUE-Benchmark einen Wert von 91,2 Punkten erreicht.

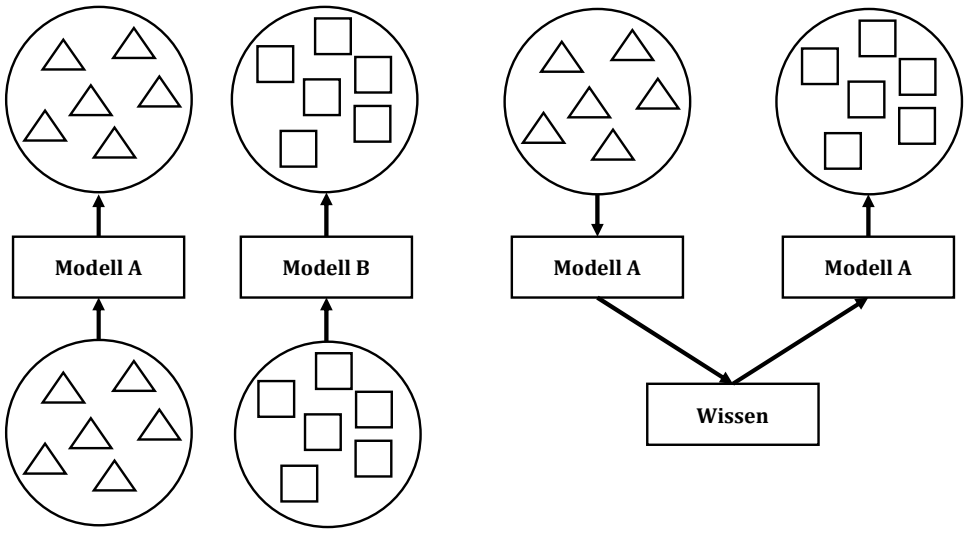
4.8 Transfer Learning

Eine bedeutende Eigenschaft von Transformermodellen ist, dass sie für das sogenannte Transfer Learning genutzt werden können (WOLF et al. 2020). Transfer Learning ermöglicht es, große Sprachmodelle, die mit umfassenden und generischen Textkorpora auf professionellsten Rechnersystemen trainiert wurden, auf individuelle Problemstellungen zu adaptieren. Durch das Vortrainieren großer Sprachmodelle mit großen Datenmengen erlangen diese generische und universell einsetzbare Fähigkeiten und Verständnisse von Sprache (RAFFEL et al. 2019; WOLF et al. 2020).

Dieses allgemeine Sprachverständnis kann anschließend genutzt werden, um nachgelagerte NLP-Probleme zu bearbeiten. Grundlegend erlangen Transformermodelle ihr Sprachverständnis auf

ähnliche Art wie Menschen. So gehen Wissenschaftler:innen davon aus, dass Kinder die Bedeutung neuer Wörter nicht ausschließlich durch aktives Lernen verstehen, sondern ein Verständnis für Wortbedeutungen auch durch Lesen geschaffen wird (JURAFSKY und MARTIN 2019). Auf Basis des Kontexts, in dem ein unbekanntes Wort auftaucht, kann der Mensch Rückschlüsse auf die semantische Bedeutung des Wortes ziehen. Transfer Learning und Transformermodelle bedeutet für die NLP-Forschung einen Paradigmenwechsel. Abbildung 20 verdeutlicht den Unterschied zwischen konventionellen überwachten Lernverfahren und dem neuen Ansatz des Transfer Learnings.

Abbildung 20: Supervised Learning (links) vs. Transfer Learning (rechts).



Quelle: RUDER (2019: 43) (verändert).

Typischerweise benötigen Forschende große Datenmengen manuell annotierter Daten, die für das sogenannte überwachte (supervised) Lernen genutzt wurden. Dieser Prozess ist auf der linken Seite von Abbildung 20 dargestellt. Auf Basis dieser Datensätze können Modelle trainiert werden, die Regelmäßigkeiten und Strukturen in den annotierten Daten erlernen und dieses Wissen auf bisher ungesehene Daten übertragen können. Während des Trainings wird die Vorhersage des Modells mit den manuell annotierten Daten verglichen und somit die Leistungsfähigkeit des Modells evaluiert. Je nach Aufgabe werden jedoch enorme Mengen manuell annotierter Datenpunkte für das Training benötigt, sodass ein großer Vorbereitungsaufwand entstehen kann. Für das Transfer Learning bedarf es deutlich weniger manuell annotierte Trainingsdaten als bei klassischen überwachten Lernverfahren.

Die Transformerarchitektur setzt unüberwachtes (unsupervised) Lernen ein, um Sprachmodelle vorzutrainieren. Mit unsupervised Learning geht der bedeutende Vorteil einher, dass nicht-annotierter Text durch das Internet in Massen verfügbar ist. Um Transformermodelle vortrainieren zu

können, müssen lediglich einzelne Tokens in den Trainingsdaten maskiert werden. Da dies automatisiert umgesetzt werden kann, können mit unüberwachtem Lernen immer größere und damit auch immer leistungsfähigere Modelle vortrainiert werden (GOLDBERG 2017a; SHAZEER et al. 2017; JOZEFOWICZ et al. 2016b; SHAZEER et al. 2018; KESKAR et al. 2019). Durch die Betrachtung gigantischer Textmengen und der Skalierbarkeit neuronaler Netze erlernen diese Sprachmodelle ein grundlegendes Verständnis von Sprache. Exemplarisch zeigen BAEVSKI et al. (2019), dass mehr Trainingsdaten für bessere Vorhersageleistungen bei der Sprachmodellierung sorgen. Somit können diese vortrainierten Modelle anschließend auf ein spezielles NLP-Problem hin adaptiert werden.

Um das Sprachverständnis vortrainierter Modelle für nachgelagerte NLP-Aufgaben zu nutzen, bestehen grundsätzlich zwei Möglichkeiten. Das Sprachmodell kann zur Feature-Extraktion verwendet werden, indem die numerischen Repräsentationen der Textsequenzen an ein eigenständiges Modell z.B. zur Textklassifikation überführt werden. Dabei findet keine Anpassung des Sprachmodells statt, sondern es werden lediglich die hochdimensionalen, vortrainierten Features aus dem Sprachmodell genutzt. Bei der Feinabstimmung (engl. fine-tuning) findet eine Aktualisierung der vortrainierten Repräsentationen statt, sodass das Modell an die vorliegenden Daten angepasst wird (RUDER 2019). Dazu wird ein bestehendes Modell, wie auf der rechten Seite von Abbildung 20 zu sehen, mit den neuen Daten (weiter)trainiert und ggf. marginale Änderungen an den letzten Schichten des Modells vorgenommen, z.B. durch das Hinzufügen oder Adaptieren aufgabenspezifischer Schichten (RUDER et al. 2019a; PETERS et al. 2019). RUDER et al. (2019a) zeigen, dass feinabgestimmte Modelle besonders dann besser abschneiden, wenn sich Ausgangs- und Zielaufgabe ähneln.

Der Einsatz vortrainierter Modelle zur Prozessierung spezieller NLP-Probleme zeigt dabei in der Praxis deutlich bessere Ergebnisse im Vergleich zu kleineren Modellen, die von Grunde auf trainiert wurden (RAFFEL et al. 2019; CLARK et al. 2018; PETERS et al. 2018). HOWARD und RUDER (2018) zeigen beispielsweise, dass im Nachgang feinabgestimmte vortrainierte Transformermodelle strikt bessere Ergebnisse erzielen als von Grund auf trainierte Modelle. Das ist sogar dann der Fall, wenn ein frisch initialisiertes Modell mit 100 mal mehr Trainingsdaten trainiert wird als das vortrainierte, feinabgestimmte Modell (HOWARD und RUDER 2018).

Somit bedeutet das Konzept des Transfer Learnings eine Revolution der NLP-Verfahren. Während korpuslinguistische Verfahren wie Bag-of-Words oder statische Word-Embeddings Wortbedeutungen lediglich auf Basis vorliegender Korpora lernen, nutzt Transfer Learning das volle Potential der Massen an digital verfügbarem Text und erleichtert gleichzeitig massiv das Training von nachgelagerten NLP-Anwendungen. Darüber hinaus lösen vortrainierte Transformermodelle un-

ter Einbezug syntaktischer und semantischer Merkmale viele klassische Probleme der Vorverarbeitung von Texten. Beispielsweise stellen Synonyme, Polysemie, Flexion oder Ironie korpuslinguistische Verfahren vor Probleme, da diese Wörter zwar unter Einbezug des Kontexts, aber dennoch als statistische Merkmale verarbeitet werden, sodass einem Wort lediglich eine Bedeutung zugeordnet werden kann.

Mittlerweile umfasst das Modellrepertoire über 66.000 unterschiedlich vortrainierte Sprachmodelle in über 170 Sprachen (HUGGING FACE 2022a). Einige dieser Modelle haben einen thematischen Fokus, beispielsweise für spezielle thematische Subdisziplinen, wie Biomedizin (LEE et al. 2020), Psychologie (VAJRE et al. 2021) oder Rechtswissenschaften (ZHENG et al. 2021). Zudem existieren Modelle zur Verarbeitung von wissenschaftlicher Literatur (BELTAGY et al. 2019), Texten aus den sozialen Medien (LOUREIRO et al. 2022) oder zur Modellierung von Programmcode (FENG et al. 2020).

Weiterhin sind die Modelle teilweise für spezielle Textverarbeitungsaufgaben vortrainiert, wie beispielsweise Textklassifikation, Textzusammenfassung, Textübersetzung oder zur Fragenbeantwortung. Neueste Entwicklungen im Bereich des Transfer Learnings versuchen den Trainings- bzw. Adaptierungsaufwand vortrainierter Modelle zur Lösung individueller NLP-Probleme zu minimieren. Daher forschen aktuelle Studien an sogenannten Few-Shot-, One-Shot- oder gar Zero-Shot-Verfahren. Während Few-Shot-Verfahren noch zwischen 10 und 100 annotierte Trainingsbeispiele benötigen, genügt One-Shot-Verfahren ein Trainingsbeispiel, um das Modell an spezifische Aufgaben anzupassen. Zero-Shot-Ansätze verfolgen das Ziel, vollkommen ohne manuell annotierte Trainingsdaten spezifische NLP-Aufgaben zu lösen. Statt annotierter Trainingsbeispiele erhält das Modell eine Beschreibung der Aufgabe und kann diese ohne weitere Adaptierung umsetzen (BROWN et al. 2020).

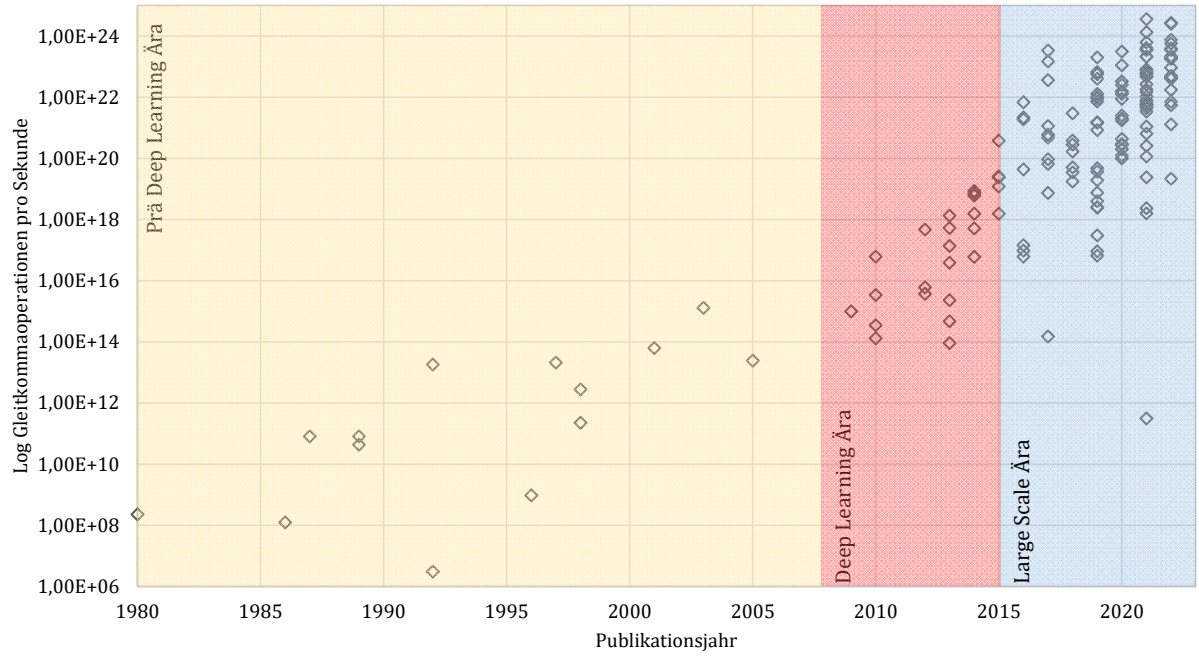
4.9 Zusammenfassung der Entwicklung des NLP

Zusammenfassend lässt sich festhalten, dass das Forschungsgebiet des NLP insbesondere innerhalb der letzten Dekade eine enorme Dynamik entwickelt hat. Trotz dieser markanten Veränderungen von Verfahren und Modellen lässt sich seit Anbeginn der NLP-Forschung ein übergeordnetes gemeinsames Ziel festhalten: Wörter und Textsequenzen sollen möglichst adäquat in numerische Formate überführt werden, um diese maschinell weiterverarbeiten zu können. Auslöser dieser Entwicklung sind dabei die methodischen Innovationen der neuronalen Netze und deren Anwendung auf Textdaten. Während bis in die 2000er Jahre hinein einfache BOW-Verfahren rudimentäre Klassifikations- und Clustermöglichkeiten bieten, stellt die Entwicklung der Word-Embeddings zu Beginn der 2010er Jahre einen bis heute zentralen Meilenstein in der NLP-Forschung dar. Sie ermöglichen erstmals den Einbezug des Kontexts bei der mathematischen Interpretation

von Wörtern und zeigen somit, dass Computer in der Lage sind semantische Beziehungen aus Texten abzuleiten.

Aufsetzende Arbeiten verfolgen die Zielsetzung mittels unterschiedlicher Ausbaustufen neuronaler Netze (z.B. RNN, LSTM, Bi-LSTM) einen immer umfassenderen und individuelleren Kontext bei der numerischen Einbettung von Textsequenzen zu berücksichtigen. Die enormen Fortschritte komplexer ML-Modelle sind in erster Linie mit steigenden Rechnerleistungen zu begründen. SEVILLA et al. (2022) gliedern die Entwicklungen des ML in drei Epochen. Abbildung 21 zeigt die Entwicklung der Rechenleistung ausgewählter NLP-Modelle anhand der Anzahl der Gleitkommaoperationen pro Sekunde. Dieses Maß repräsentiert die Leistungsfähigkeit von Computerprozessoren. Wie aus Abbildung 21 hervorgeht, hat die Rechnerleistung in der Prä Deep Learning Ära nur langsam zugenommen (18 Monate Verdopplungszeit). Das Aufkommen der Deep Learning Ära ist unmittelbar mit deutlich schneller zunehmenden Rechnerleistungen verknüpft. So zeigen SEVILLA et al. (2022), dass sich während dieser Phase die Rechenleistung von ML-Systemen alle sechs Monate verdoppelte.

Abbildung 21: Entwicklung der Rechenleistung ausgewählter NLP-Modelle.



Quelle: Eigene Darstellung nach SEVILLA et al. (2022: 3).

In der Large Scale Ära zwischen 2015 und 2022 flacht die Wachstumsrate etwas ab, dennoch verdoppelt sich die Rechenleistung in dieser Zeit alle 10 Monate. In diesem Zeitraum erfährt die NLP-Forschung nochmals einen großen Sprung durch die Entwicklung der Transformermodelle. Die komplexe Modellarchitektur ermöglicht einerseits eine enorm spezifische numerische Einbettung von Textsequenzen. Andererseits eröffnet der Einsatz von unüberwachten Lernverfahren für das

Training die Möglichkeit eine enorme Menge an digital verfügbarem Text in die Sprachmodelle einfließen zu lassen.

Darüber hinaus zeigt sich, dass das Training größerer Modelle auf Basis größerer Textmengen in vielen Fällen in besseren Ergebnissen resultiert (RAFFEL et al. 2019; KESKAR et al. 2019; SHAZEER et al. 2017). Somit können relativ einfach enorm große und entsprechend leistungsfähige Sprachmodelle trainiert werden, die in kürzester Zeit in unterschiedlichsten Textverarbeitungsaufgaben menschliche Qualität erreichen.

Das allgemeine Sprachverständnis großer Sprachmodelle nutzt das sogenannte Transfer Learning, um mit vergleichsweise geringem Adaptierungsaufwand individuelle NLP-Aufgaben lösen zu können. Mittlerweile geht die Anwendung von NLP-Modellen über die reine Textverarbeitung hinaus. Beispielsweise sind sogenannte Stable Diffusion Modelle in der Lage eine Textsequenz in ein Bild zu übersetzen (RAMESH et al. 2021; TANG et al. 2022).

Für die sozialwissenschaftliche Forschung eröffnen diese rapiden Entwicklungen in der automatisierten Textverarbeitung die Potentiale von Big Data und digitalem Text zu nutzen. Textdaten bieten enorm detaillierte und vielfältige Einblicke in die soziale Wirklichkeit. Sie umfassen neben sachlichen Informationen ebenso Wertungen, Ansichten oder Wahrnehmungen. Diese Multidimensionalität der Datenquelle, gepaart mit den weitreichenden Umfängen, in denen Textdaten vorliegen, kann es sozialwissenschaftlicher Forschung gelingen, neue Untersuchungsperspektiven einzunehmen, bestehende Forschungsfragen neu denken zu können und neue bearbeiten zu können.

Wie NLP und Web Mining perspektivisch in wirtschaftsgeographische Forschung eingebunden werden können, soll der empirische Teil dieser Arbeit illustrieren. Das folgende Kapitel gibt einen Überblick über das Forschungsdesign.

5 Vorstellung des Forschungsdesigns und der Fallstudien

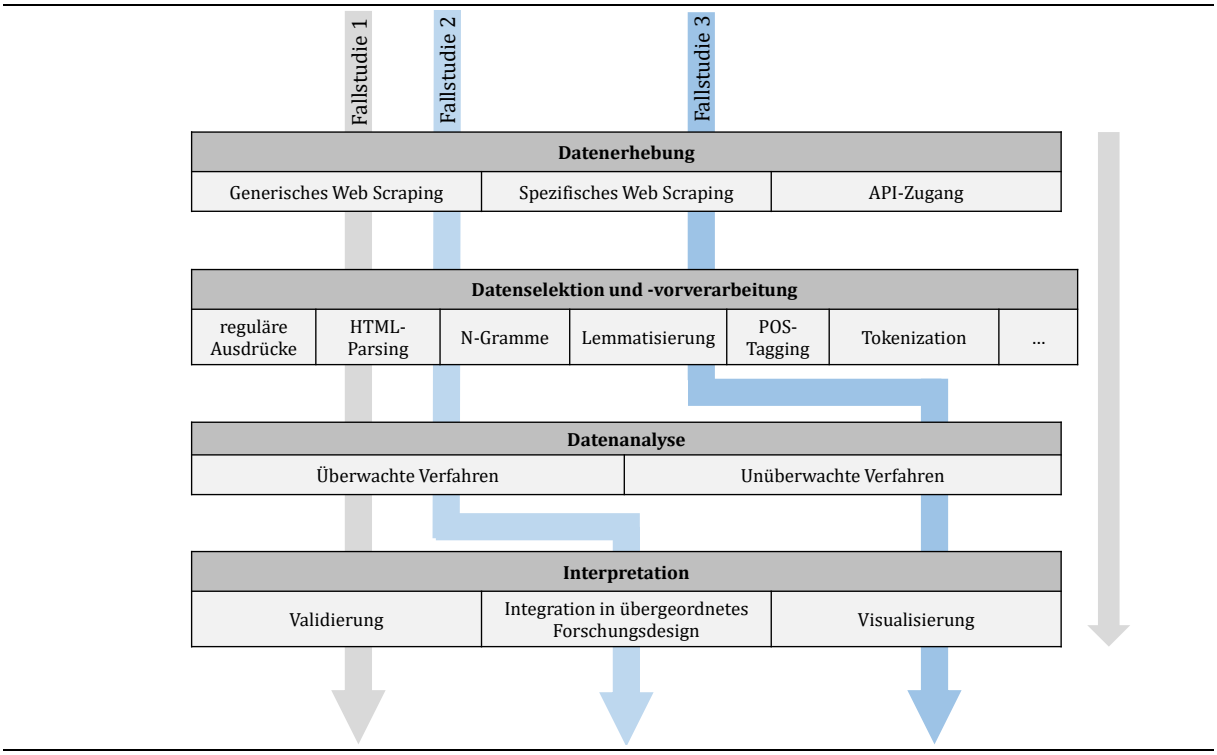
Nachdem in den vorherigen Kapiteln die konzeptionellen Grundlagen zu den Themenbereichen NLP und Web Mining vorgestellt wurden, sollen in den folgenden Kapiteln anhand von Fallstudien konkrete Anwendungsmöglichkeiten der dargestellten Methodiken für die (Wirtschafts-)geographie skizziert werden. Über die reine Deskription potentieller Anwendungsgebiete hinausgehend werden explizit die notwendigen Schritte beschrieben, um Webmassendaten als Datenquelle für geographische Fragestellungen nutzbar zu machen.

Das Analysedesign dieser Arbeit ist daher als anwendungs- sowie methodenzentrierte Exploration zu verstehen. Um der Bandbreite potentieller Anwendungsbereiche gerecht zu werden, umfassen die Fallstudien sowohl Längs- als auch Querschnittsbetrachtungen, nehmen unterschiedliche Untersuchungsgegenstände in den Fokus, verarbeiten Texte unterschiedlicher Sprachen und demonstrieren den Einsatz verschiedener NLP-Techniken. An die Fallstudien schließt sich eine abschließende Synthese der Empirie an, welche die Ergebnisse der empirischen Untersuchungen unter Rückbezug auf die bereits erläuterten konzeptionellen Grundlagen diskutiert und interpretiert.

Wie bereits in Kapitel 3 beschrieben, kann der Web Mining-Prozess als vierstufiger Prozess dargestellt werden, welcher von der Datenerhebung über die Datenselektion und –vorverarbeitung bis hin zur Analyse und Interpretation reicht. Innerhalb der jeweiligen Prozessstufen stehen abhängig von Forschungsfrage und Datenzugang verschiedene Verfahren zur Verfügung, die nahezu beliebig kombiniert werden können. Dieser Logik folgend kombinieren die Fallstudien des empirischen Teils dieser Arbeit unterschiedliche Bausteine des Web Mining-Prozesses. Abbildung 22 zeigt die Verfahrensabfolgen der drei Fallstudien auf. Die Datenerhebungsmethode des API-Zugangs wird im Rahmen dieser Arbeit nicht vorgestellt, da diese in der Regel strukturierte Daten bereitstellt, sodass der Zugang außerhalb des Untersuchungsfokus liegt. Auf Stufe der Datenselektion und- vorverarbeitung können nochmals die verschiedenen Verfahren kombiniert werden.

Der Fokus der ersten Fallstudie liegt auf dem Datenzugang und der Datenerhebung und adressiert daher die ersten beiden Forschungsfragen. Fallstudie 2 dient schwerpunktmäßig zur Beantwortung der dritten und vierten Forschungsfrage. Die dritte Fallstudie zielt auf die Beantwortung der fünften Forschungsfrage ab. Aus der Synthese der Beantwortungen der Forschungsfragen sowie den empirischen Ergebnissen der Fallstudien soll abschließend die letzte Forschungsfrage beantwortet werden.

Abbildung 22: Analysedesign der Empirie.



Quelle: Eigene Darstellung.

5.1 Fallstudie 1: Web-Mining deutscher Unternehmenswebseiten

Die erste Fallstudie zielt inhaltlich auf die Webseitenidentifizierung und Geokodierung deutscher Handelsregisterunternehmen ab. Als Datengrundlage dient das offene Webrepositorium CC. Das Projekt stellt mehrfach pro Jahr Auszüge eines umfassenden Webcrawlings frei zur Verfügung und stellt somit eine offen zugängliche Domaindatenbank dar. Dieses wird genutzt, um mittels generischen Web Scrapings über 9 Millionen Domains abzufragen und die Unternehmenswebseiten zu identifizieren.

Methodisch liegt der Fokus neben dem umfangreichen Web Scraping von Webmassendaten insbesondere auf der Entwicklung und Implementierung eines überwachten Textklassifikationsverfahrens auf Wortebene. Dieses wird genutzt, um Impressumseinträge von Webseiten syntaktisch zu kategorisieren. Diese klassifizierten Impressumseinträge werden anschließend genutzt, um die Ortsangaben zu geokodieren. Inhaltlich exploriert diese Fallstudie die Qualität des CC für geographische Webforschung. Dies geschieht einerseits durch eine umfassende deskriptive Betrachtung des Datensatzes. Andererseits werden die Ergebnisse des Web Minings mit amtlicher Statistik räumlich validiert und interpretiert.

5.2 Fallstudie 2: Identifizierung und Standortanalyse deutscher KI-Unternehmen

Fallstudie 2 setzt auf die Ergebnisse von Fallstudie 1 auf und illustriert, wie Webseiteninhalte deutscher Unternehmenswebseiten bezogen, aufbereitet und gefiltert werden können. Es wird

ein Verfahren zur systematischen Extraktion des zentralen Webseitentexts vorgestellt. Außerdem kombiniert der Filterprozess eine Stichwortsuche mit einem überwachten Textklassifikationsverfahren auf Absatzebene. Dieses basiert auf einem vortrainierten Transformermodell (vgl. Kapitel 4.7) und nutzt das Konzept des Transfer Learnings, um das generische Sprachmodell auf die vorliegende Problemstellung hin zu adaptieren. Anzumerken ist hierbei, dass angewandte Lernverfahren in der Lage sind, auch mehrsprachige Texte zu verarbeiten, ohne, dass ein gesondertes Klassifikationsmodell trainiert werden muss.

Inhaltlich setzt sich Fallstudie 2 mit der Standortanalyse deutscher Unternehmen auseinander, die mit KI arbeiten. Somit können die webgenerierten Daten genutzt werden, um diese in ein ökonometrisches Modell zu integrieren. Damit zeigt Fallstudie 2 auf, wie die deskriptiven Klassifikationen quantitativer Textanalyse zur Erklärung regionaler Entwicklungen eingesetzt werden können.

5.3 Fallstudie 3: Dynamisches Topic Modeling wirtschaftsgeographischer Literatur

Fallstudie 3 stellt dar, wie große Textmengen mittels NLP strukturiert und analysiert werden können. Anhand einschlägiger, wirtschaftsgeographischer Fachliteratur werden die inhaltlichen Schwerpunkte des wissenschaftlichen Diskurses der letzten 30 Jahre untersucht. Methodisch veranschaulicht Fallstudie 3 den Datenzugang des spezifischen Web Scrapings und zeigt auf, wie unüberwachte Lernverfahren Texte strukturieren können. Fallstudie 3 betrachtet die Abstracts wissenschaftlicher Papiere und illustiert somit die Textanalyse auf Dokumentenebene. Da Fallstudie 3 ausgewählte (wirtschafts-)geographische Literatur der letzten 30 Jahre betrachtet, skizziert sie ebenso, wie Längsschnittanalysen auf Basis von Textdaten durchgeführt werden können.

6 Fallstudie 1: Web Mining deutscher Unternehmenswebseiten

In dieser ersten empirischen Fallstudie soll aufgezeigt werden, wie Web Mining und NLP angewandt werden können, um eine flächendeckende, koordinatenscharfe Adressdatenbank deutscher Unternehmen zu generieren. Hierzu wird zunächst die Datengrundlage CC deskriptiv beleuchtet und die grundlegende Funktionsweise des verwendeten Web Scraping-Frameworks beschrieben. Anschließend beschreibt diese Fallstudie, wie aus unstrukturierten Textdaten mithilfe eines Named Entity Recognition-Modells (NER-Modell) Adressdaten gewonnen werden können. Diese Adressdaten werden genutzt, um Unternehmen in dem generierten Datensatz zu identifizieren und zu geokodieren. Abschließend beleuchtet die Fallstudie die räumlichen Verteilungsmuster der extrahierten Unternehmensdomains und vergleicht die Ergebnisse mit offiziellen Statistiken.

6.1 Problemstellung und Hintergrund

Insbesondere in der Wirtschaftsgeographie sind räumlich granulare und inhaltlich detaillierte Informationen von essentieller Bedeutung, um die komplexen Wechselwirkungen zwischen wirtschaftlicher Entwicklung und räumlichen Gegebenheiten verstehen zu können. Um die in Kapitel 2.2 beschriebenen Potentiale von Textdaten für die wirtschaftsgeographische Forschung nutzen zu können, stellen Unternehmenswebseiten eine zentrale Ressource für sämtliche aufsetzende Forschungsvorhaben dar. Trotz der vielversprechenden Potentiale, die Web Mining und NLP bereithalten, sind diese Daten bis dato für die Wissenschaft nicht frei zugänglich. Existierende Arbeiten zur systematischen und umfassenden Analyse von Unternehmenswebseiten basieren auf kommerziellen Datenbanken, welche einen freien und offenen Wissenschaftsbetrieb behindern. Dabei ist das Internet theoretisch nahezu barrierefrei abrufbar, da kostenfrei und effizient nach unterschiedlichsten Inhalten gesucht werden kann. Dennoch ist der Zugang zu Webseiten mit einigen methodischen Hürden verbunden. Umso wünschenswerter ist die Etablierung einer Methodik, die einen kostenfreien Zugang zu Unternehmenswebseiten für wissenschaftliche Forschung ermöglicht.

Vor dem Hintergrund der zunehmenden Relevanz des Internets für F&E haben sich verschiedene Projekte entwickelt, welche umfassende Webkorpora bereitstellen. Diese werden in diversen Forschungsvorhaben als Datengrundlage verwendet, um beispielsweise NLP-Modelle zu trainieren oder Modelle zur Informationsextraktion zu entwickeln (LIU et al. 2019; BROWN et al. 2020). Entsprechend haben sich in der jüngeren Vergangenheit vermehrt Webkorpora herausgebildet, welche hochqualitativen Text enthalten, um performantere Modelle trainieren zu können. Beispielhaft zu nennen sind die Korpora CCNet (WENZEK et al. 2019), MassiveText (RAE et al. 2021) oder C4 (RAFFEL et al. 2019).

Die Datengrundlage dieser Projekte ist das offene Webarchiv CommonCrawl (COMMONCRAWL 2022). Das CC ist die größte frei zugängliche Quelle für Webseiten und deren Metadaten. Das Projekt wird von der Non-Profit-Organisation CC betrieben, die seit 2008 eine frei zugängliche Teilkopie des Internets für Forschungszwecke und andere Analysen erstellt. Insgesamt umfasst das Repositorium über 220 Milliarden Webseiten und deren Inhalte. Das Crawling erfolgt nicht zufällig, sondern priorisiert Domains mit einer höheren harmonisierten Netzwerkzentralität innerhalb des hausinternen Webgraphen. Damit verfolgt CC das Ziel Spam zu reduzieren und repräsentative Stichproben von Webseiten bereitzustellen (NAGEL 2021). Wie aus der bisherigen Darstellung der Thematik hervorgeht, werden die in Webkorpora enthaltenen Texte primär für das Training von NLP-Modellen verwendet. Abseits der Computerliguistik bestehen nur wenige Studien, die auf dem CC aufbauen. Inwiefern das CC geeignet ist, um systematisch Webseitenbetreiber:innen zu identifizieren und geographisch zu verorten, ist bis dato ebenso noch ungeklärt. Einen Überblick über die Datenformate und Inhalte des CC gibt das folgende Kapitel.

6.2 Datengrundlage

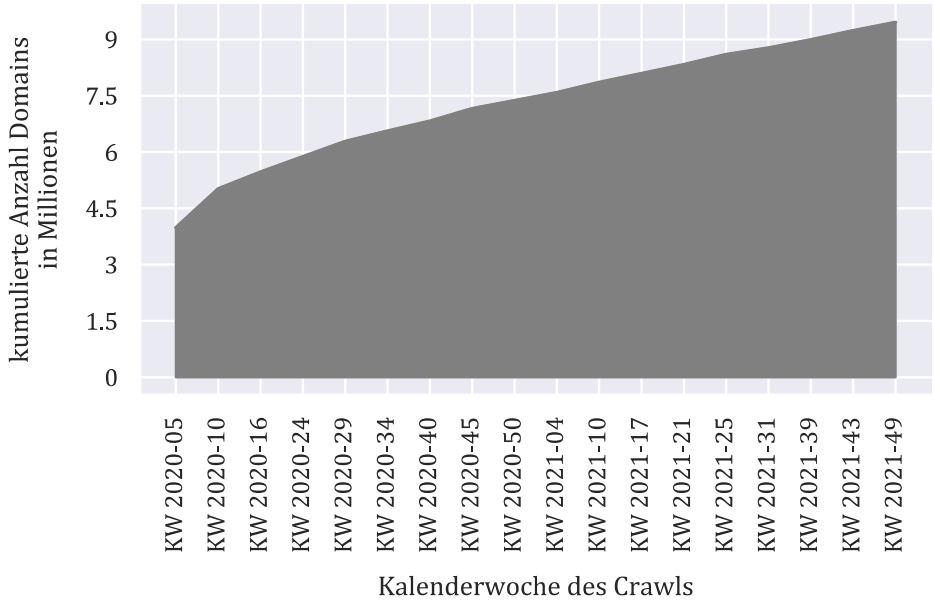
Das CC stellt Webseiten und deren Metadaten in monatlichen Auszügen und in drei unterschiedlichen Datenformaten zur Verfügung. Das umfangreichste Datenformat ist das Web Archive-Format (WARC). Dieses beinhaltet neben den vollständigen HTML-Codes der jeweiligen Webseiten Informationen über den Crawlingprozess sowie Metadaten zu den Webseiten. Das WAT-Format (WAT) beinhaltet ebenfalls die vollständigen HTML-Codes, sowie Metadaten zu den Webseiten. Informationen über den Crawlingprozess sind dort nicht enthalten. Drittens stellt das WET-Format (WET) eine schlankere Alternative zu WARC und WAT dar. WET beinhaltet lediglich den Klartext der Webseiten und wird daher häufig als Datengrundlage für die NLP-Forschung verwendet (WENZEK et al. 2019).

Neben den drei genannten Datenformaten stellt das CC einen URL-Index zur Verfügung. Dieser stellt eine Übersicht der in den monatlichen Auszügen enthaltenen URLs dar. Weiterhin sind die enthaltenen URLs mit Informationen zu Sprache des Webseitentexts und TLD attributiert. Im Vergleich zu den drei Standardformaten WARC, WAT und WET ist die Dateigröße des URL-Indexes um ein Vielfaches kleiner und erlaubt damit ein vergleichsweise schnelles Herunterladen und Verarbeiten der Daten. So beträgt die durchschnittliche Dateigröße des URL-Index rund 0,26 Terabyte pro Crawl, während das WET-Format mit knapp zehn Terabyte, das WAT-Format mit rund 23 Terabyte sowie das WARC-Format mit über 90 Terabyte pro Crawl jeweils um ein Vielfaches größer sind.

Speziell für deutschsprachige Webseiten existieren weder seitens des CC noch in der Fachliteratur Untersuchungen, inwieweit Webseiten untererfasst – also nicht im CC enthalten - sein könnten,

sodass zunächst eine kritische Analyse der enthaltenen Webseiten sinnvoll ist. Als Datengrundlage dieser Arbeit dienen die CC-Datensätze der Jahre 2020-2021. Innerhalb dieser zwei Jahre wurden in unregelmäßigen Abständen insgesamt 18 Datensätze seitens des CC publiziert. Insgesamt beinhalten diese 18 Datensätze 9.455.551 einzigartige Domains. Ein Crawl umfasst dabei zwischen 3,9 Millionen und 4,4 Millionen einzigartige deutschsprachige Domains. Um eine möglichst vollständige Liste deutschsprachiger Domains zu erhalten, wurden aus den betrachteten Crawls alle einzigartigen Domains selektiert.

Abbildung 23: Kumulierte Verteilung einzigartiger, deutschsprachiger Domains.

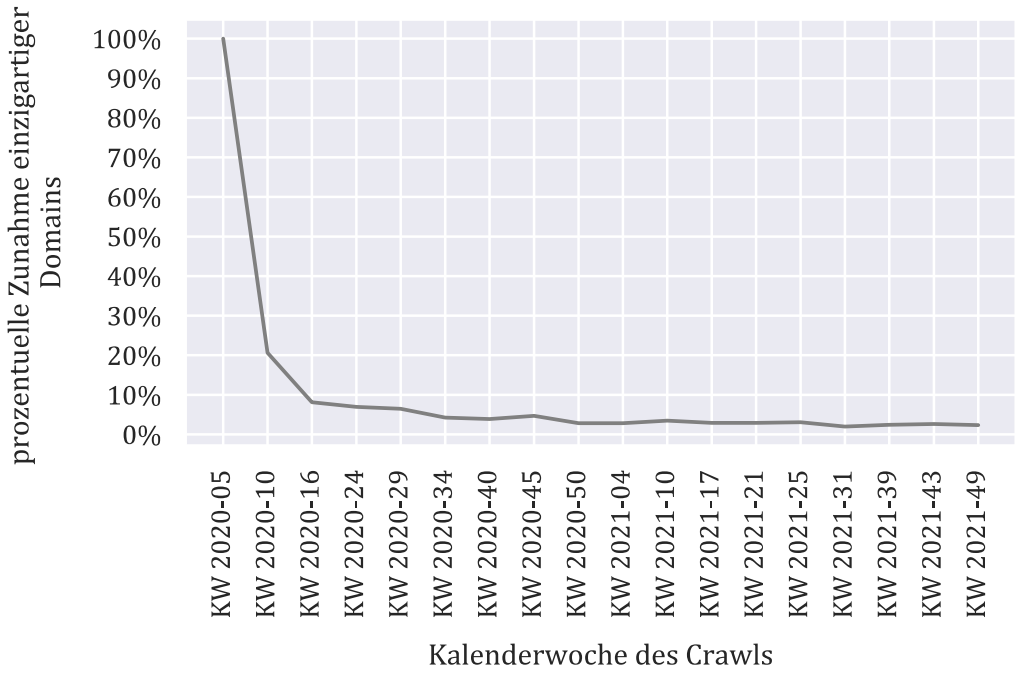


Quelle: Eigene Darstellung.

Abbildung 23 zeigt die Verteilung der identifizierten deutschsprachigen Domains über die 18 Crawls. Der erste betrachtete Crawl enthält rund 3,9 Millionen einzigartige Domains. Damit sind bereits 42,15 % aller untersuchten Domains in dem ersten analysierten Datensatz enthalten. Wie aus Abbildung 24 hervorgeht, nimmt der relative Anstieg neuer Domains bereits nach zwei betrachteten Datensätzen deutlich ab. Durch die Hinzunahme des Datensatzes *KW 2020-10* wächst die Anzahl einzigartiger Domains noch um 20,6 %. Ab Crawl *KW 2020-50* beträgt der Anteil neuer Domains in den jeweiligen Crawls nur noch zwischen 2 und 3,4 %. In absoluten Zahlen kommen ab Crawl *KW 2020-50* zwischen rund 176.000 und 273.000 neue Domains pro Crawl hinzu.

Folglich lässt sich festhalten, dass nur wenige Datensätze ausreichen, um einen Großteil der im CC gespeicherten Domains zu erhalten. Zwar könnte es durch die Hinzunahme weiterer Datensätze noch zu stärkeren Anstiegen der prozentualen Zunahme einzigartiger Domains kommen. Allerdings ist nach Betrachtung der vorliegenden 18 Datensätze zu vermuten, dass nur geringe und damit vernachlässigbare Zuwächse zu erwarten sind.

Abbildung 24: Zunahme einzigartiger Domains pro Crawl.



Quelle: Eigene Darstellung.

Da Webseitentexte in mehreren Sprachen formuliert sein können, gibt das CC bis zu drei unterschiedliche Sprachen pro URL an. Die Sprachen der Webseiten werden von CC während des Crawlings erhoben. Dabei entspricht die Reihenfolge der Sprachenangaben den Anteilen der jeweiligen Sprachen auf den Webseiten. An dieser Stelle wurden Webseiten extrahiert, die mindestens in Teilen deutsche Sprache enthalten. Der so entstandende Datensatz umfasst Domains in insgesamt 7.914 Sprachkombinationen. Die Betrachtung von Tabelle 4 zeigt, dass die Sprachkombinationen extrem schief verteilt sind.

Tabelle 4: Übersicht der zehn häufigsten Sprachkombinationen im betrachteten Datensatz.

Sprache(n)	absolute Häufigkeit	prozentualer Anteil
deu	4.794.206	50,07 %
deu,eng	1.619.788	17,13 %
eng,deu	1.037.476	10,97 %
zho,deu	97.751	1,03 %
rus,deu	82.397	0,87 %
fra,deu	49.122	0,52 %
zho,eng,deu	48.820	0,50 %
fra,eng,deu	47.412	0,50 %
rus,eng,deu	46.549	0,49 %
eng,fra,deu	45.066	0,48 %

Quelle: Eigene Darstellung und Berechnung.

Über die Hälfte der enthaltenen Domains sind rein in deutscher Sprache formuliert. Weitere knapp 28 % enthalten sowohl englische als auch deutsche Inhalte. Alle weiteren Sprachkombinationen sind anteilig nur sehr selten vertreten. Somit lässt sich festhalten, dass die gewählte Datengrundlage zur Untersuchung deutschsprachiger Webseiten gut geeignet ist. Eine gleichmäßigere Verteilung unterschiedlicher Sprachkombinationen hätte bei der Analyse aufgrund der Heterogenität der Sprachen zu erheblichem Mehraufwand geführt.

Weiterhin lässt sich der Datensatz hinsichtlich der Verteilung der TLD analysieren. Insgesamt umfasst der Datensatz 2.304 einzigartige TLD. Die Verteilungen der TLD sind dabei sehr einseitig, sodass die zehn häufigsten TLD bereits 84,55 % Domains abdecken. Wie aus Tabelle 5 hervorgeht, haben über 41 % der analysierten Domains die TLD „.de“. Weitere häufig auftretende TLD sind neben den allgemeinen TLD „.com“, „.net“ und „.org“ die länderspezifischen TLD der Schweiz, Österreichs, Russlands und Italiens. Insgesamt waren 2022 rund 17,2 Millionen Domains mit der Endung „.de“ registriert (VERISIGN, INC. 2022). Somit sind über 20 % aller „.de“-Domains im vorliegenden deutschsprachigen Subset enthalten. Neben der sprachlichen Eignung der enthaltenen Webseiten ist folglich auch davon auszugehen, dass die Verteilung der TLD eine geeignete Datengrundlage zur Untersuchung deutscher Webseiten darstellt.

Tabelle 5: Übersicht der zehn häufigsten TLD im betrachteten Datensatz.

TLD	Absolute Häufigkeit	Relative Häufigkeit
.de	3.933.071	41,60 %
.com	2.163.188	22,88 %
.ch	474.572	5,02 %
.at	419.498	4,43 %
.net	248.204	2,62 %
.org	205.717	2,17 %
.eu	167.513	1,77 %
.ru	140.622	1,48 %
.info	110.601	1,17 %
.nl	105.124	1,11 %

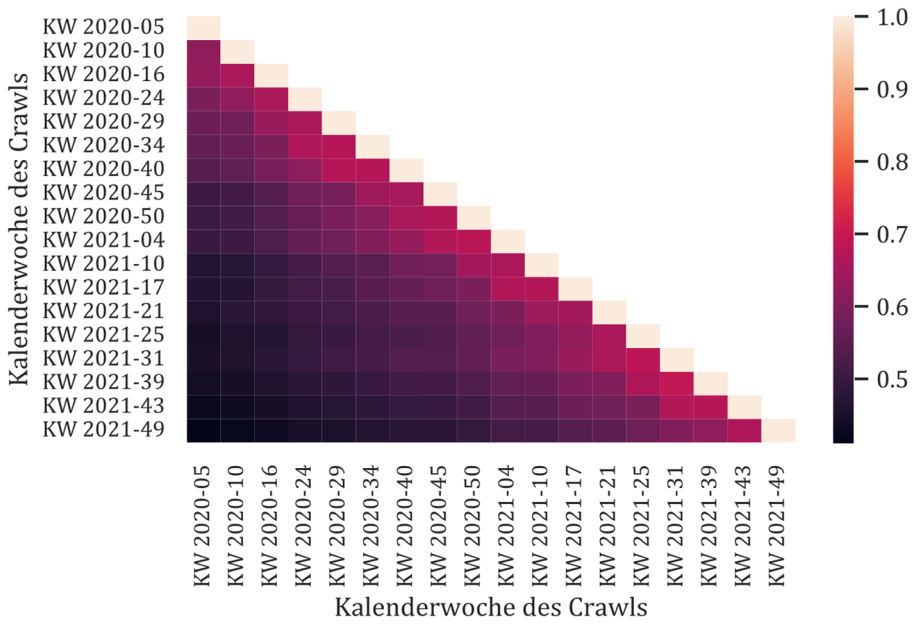
Quelle: Eigene Darstellung.

Aufgrund der schiefen Verteilungen der Sprachkombinationen und TLD in den betrachteten Crawls ist von einer relativ starken Überlappung der Crawls untereinander auszugehen. Abbildung 25 veranschaulicht diese Überlappung anhand des Jaccard-Koeffizienten. Hierzu wurden die enthaltenen Domains der jeweiligen Crawls miteinander verglichen.

Der Jaccard-Koeffizient bescheinigt den untersuchten Datensätzen eine moderate bis starke Überlappung von Werten zwischen 0,4 und 0,7. Weiterhin wird ersichtlich, dass das Datum des Crawls Einfluss auf die Stärke der Überlappung nimmt. Während zeitlich weiter auseinanderliegende

Crawls eher geringere Überschneidungen aufweisen, fallen die Werte für kalendarisch enger beieinander liegende Crawls systematisch höher aus. Dieses Muster kann durchaus auf die Dynamik des Internets als Ganzes zurückzuführen sein, sodass einige Domains innerhalb weniger Monate wieder gelöscht werden und dafür monatlich viele neue Domains aktiv werden. Dennoch unterschreitet sich der Jaccard-Koeffizient nie Werte von 0,4, sodass von einem stabilen Domainstamm auszugehen ist, welcher sich auch im Zeitverlauf nicht verändert.

Abbildung 25: Jaccard-Koeffizienten der betrachteten Crawls.



Quelle: Eigene Darstellung.

6.3 Web Scraping

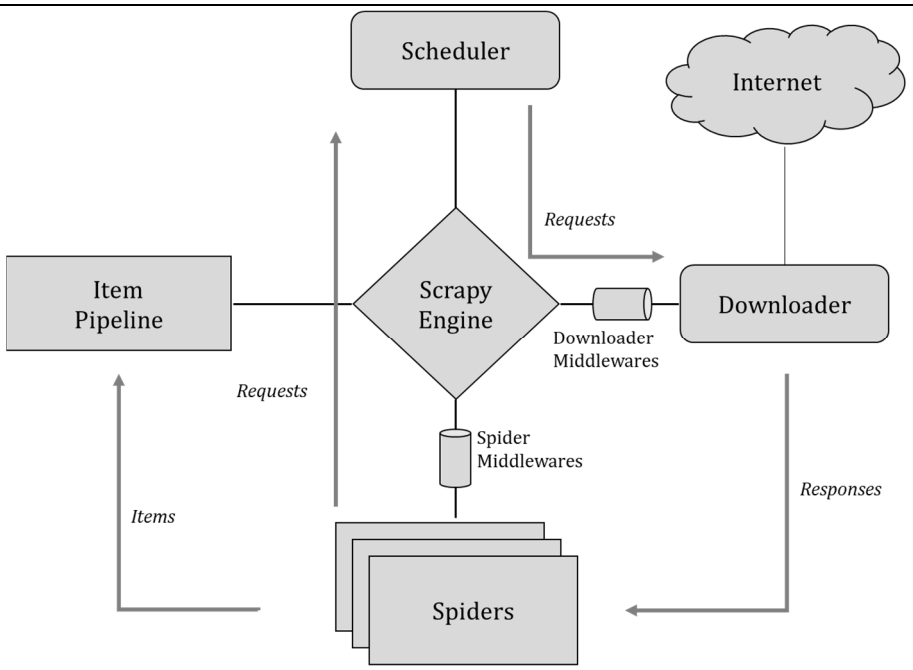
Nachdem in den vorherigen Kapiteln das CC-Projekt vorgestellt und die enthaltenen Domains deskriptiv analysiert wurden, sollen in diesem Kapitel der Aufbau sowie die Funktionsweise eines Web Scrapers erläutert werden. Wie bereits in Kapitel 3.1.1 erläutert stehen methodisch je nach Art des Scrapingvorhabens verschiedene Programmbibliotheken zur Verfügung. Da es sich beim Web Scraping mehrerer Millionen Webseiten um ein generisches Vorhaben handelt, welches nicht explizit auf dynamische Webseiteninhalte abzielt, wurde ein Verfahren gewählt, welches den HTML-Code der Webseiten abruft und ausliest.

Das größte, vielseitigste und performanteste Web Scraping-Framework Scrapy wurde 2008 entwickelt und ist vor allem für umfassende Scrapingvorhaben gut geeignet (SCRAPY COMMUNITY 2022b). Je nach Komplexität des verwendeten Scraping-Frameworks können mit Scrapy mehrere Tausend Webseiten pro Minute abgerufen werden. Darüber hinaus ist Scrapy über individuell

programmierbare Klassen und Algorithmen quasi beliebig individualisierbar. Scrapy enthält standardmäßig allein für die Downloader-Middleware 16 Komponenten, die individuell parametrisierbar sind. Mittels dieser Middleware lässt sich global das Anfrage- und Abrufverhalten von Scrapy einstellen. Beispielsweise wird definiert wie Scrapy mit Cookies und Weiterleitungen umgehen soll oder welche Wartezeiten zwischen den Abrufen eingehalten werden sollen (SCRAPY COMMUNITY 2022a).

Abbildung 26 zeigt schematisch den grundlegenden Aufbau des Frameworks sowie den Datenfluss zwischen den einzelnen Komponenten. Das Herzstück des Scrapy-Frameworks bildet die Scrapy Engine. Diese kontrolliert den Datenfluss zwischen allen Systemkomponenten und aktiviert bzw. deaktiviert einzelne Komponenten abhängig vom Datenfluss. Der Scheduler erhält die Anfragen von der Engine, reiht diese ein und gibt bei Bedarf die Anfragen an die Engine zurück. Der Downloader ruft die angefragten Webseiten aus dem Internet ab und sendet die Antwort an die Engine. Spiders sind individuell programmierbare Klassen, welche die benutzerdefinierte Verarbeitung und Extraktion von Inhalten der Webseiten (Items) definieren. Sie können dabei weitere Anfragen (Requests), z.B. über das Folgen von Hyperlinks, produzieren und diese an die Engine weitergeben.

Abbildung 26: Aufbau des Scrapy Frameworks.



Quelle: Eigene Darstellung, nach Scrapy (2021).

Außerdem können pro Antwort fest definierte Merkmale (Items) extrahiert werden. Die Item-Pipeline verarbeitet die Items, die von den Spiders extrahiert wurden, und speichert diese als Ergebnis des Scrapingprozesses ab. Die Downloader Middleware sitzt zwischen Engine und Downloader und kann Anfragen, die von der Engine an den Downloader gesendet werden, individuell

prozessieren. Die Spider Middleware sitzt zwischen Engine und den Spiders und kann sowohl den Spider-Input (Antworten), als auch die Spider-Outputs (Items und Anfragen) prozessieren.

Der Datenfluss zwischen den einzelnen Komponenten des Frameworks kann wie folgt beschrieben werden:

1. Die Engine öffnet eine neue Domain.
2. Die Engine erhält die erste URL von der Spider und übersendet diese an den Scheduler.
3. Die Engine gibt dem Scheduler ein Signal die nächste URL anzufragen.
4. Der Scheduler übergibt die nächste URL über den Engine an den Downloader. Dabei passiert die URL die Downloader Middleware (Anfrage-Richtung), die ggf. die Anfrage prozessiert.
5. Der Downloader ruft die URL ab, generiert eine Antwort und sendet diese über die Downloader Middleware (Antwort-Richtung) an die Engine.
6. Die Engine erhält die Antwort des Schedulers und sendet diese über die Spider Middleware (Input-Richtung) an die Spider.
7. Die Spider prozessiert die Antwort, generiert Items und neue Anfragen und sendet diese an die Engine.
8. Die Engine sendet Items an die Item-Pipeline und neue Anfragen an den Scheduler.
9. Der Prozess wird ab 2. solange wiederholt bis der Scheduler keine Anfragen mehr enthält. Die Engine schließt die Domain und beginnt mit 1.

Scrapy bietet neben der Möglichkeit, Spiders von Grund auf zu programmieren, vier generische Spiderklassen an, in welche bereits grundlegende Mechanismen integriert sind, die einen fortgeschrittenen Ausgangspunkt für weitere individuelle Anpassungen darstellen. Die Crawlspider enthält, wie bereits aus dem Namen hervorgeht, Mechanismen für das Crawling von Webseiten. Sie beinhaltet mit dem Linkextractor eine programmierbare Klasse, die nach zu definierenden Regeln Hyperlinks aus Webseiten extrahiert. Diese Regeln können beispielsweise in Form von HTML-Code, regulären Ausdrücken oder Black- bzw. Whitelists definiert sein. Die Crawlspider ruft anschließend die extrahierten Links ab, selektiert gewünschte Inhalte und extrahiert über den Linkextractor wieder neue Links.

Die XMLFeedSpider ist speziell entwickelt, um Inhalte aus XML-Feeds zu extrahieren. Sie durchsucht Webfeeds nach bestimmten XML-Knoten und extrahiert die dahinterliegenden Informationen. Damit ist sie der CSVFeedSpider sehr ähnlich. Der einzige Unterschied zwischen den Klassen besteht darin, dass CSVFeedSpider durch die Reihen eines Feeds iteriert, während die XMLFeedSpider über die XML-Knoten der HTML-Seite iteriert. Letztlich bietet die SitemapSpider die Möglichkeit die Sitemaps von Webseiten systematisch zu durchsuchen (SCRAPY COMMUNITY 2022a).

6.4 Methodische Vorgehensweise

Das Scrapy Framework wurde mit dem Ziel, aus den 9.544.441 identifizierten deutschsprachigen Domains die Unternehmenswebseiten zu extrahieren, entsprechend angepasst. Als Grundlage diente eine generische *Spider.* Um festzustellen, ob eine besuchte Domain zu einem Unternehmen gehört, folgt die im Rahmen dieser Arbeit entwickelte Spider einer einfachen Heuristik. Das Telemediengesetzes schreibt im Rahmen der allgemeinen Informationspflichten vor, dass Unternehmen in Deutschland Name, Anschrift und die Rechtsform „leicht erkennbar, unmittelbar erreichbar und ständig verfügbar" zu halten haben (§ 5 Absatz 1 Satz 1 Telemediengesetz).

Entsprechend ist davon auszugehen, dass auf den Webseiten deutscher Unternehmen ein Impressum zu finden ist. Diese rechtliche Grundlage wurde genutzt, um die Spider intelligent programmieren zu können. Somit wurde im Programmcode der Spider hinterlegt, dass auf den jeweiligen Domains nur Links gefolgt werden soll, die eines der Stichworte: „impressum", „imprint", „corporate" oder „legal" enthalten. In einigen Fällen enthält der Link der zum Impressum einer Webseite führt, keines der genannten Stichworte, sondern lediglich den Text, der auf der Webseite angezeigt wird. Folglich wurde auch dieser Fall im Programmcode berücksichtigt, sodass die Spider auch Links folgt, deren Text auf der Webseite eines der Stichworte enthält.

Somit werden sämtliche Domains, die über kein Impressum verfügen, nicht weiter beachtet. Die identifizierten Webseiten mit Impressumsangaben wurden dann automatisch heruntergeladen und in einen Algorithmus zur Textbereinigung überführt. Der Algorithmus identifizierte entsprechend den Absatz mit den Adressinformationen, indem nach Absätzen mit einer Postleitzahl gesucht wurde. Die Postleitzahl eignet sich besonders gut, um Absätze mit Adressinformationen zu extrahieren, da die Struktur einfach anhand syntaktischer Regeln erkannt werden kann. Beispielsweise sind Postleitzahlen in Deutschland immer ein fünfstelliger Zahlencode. Hierzu eignen sich reguläre Ausdrücke (engl. regular expressions) sehr gut, die in der Lage sind, Zeichenketten anhand ihrer Syntax zu analysieren und zu filtern (FRIEDL 2006). Ebenso lässt sich die Umsatzsteuer-Identikifikationsnummer mittels eines regulären Ausdrucks leicht aus dem Impressum extrahieren. Diese muss, falls vorhanden, laut Telemediengesetz ebenfalls im Impressum eingetragen sein (§ 5 Absatz 1 Satz 6 Telemediengesetz). Neben Unternehmen haben in Deutschland auch juristische Personen eine Umsatzsteuer-Identifikationsnummer, wenn diese für „innergemeinschaftliche Erwerbe" benötigt wird (§ 27a Absatz 1 Umsatzsteuergesetz). Ferner sind im Handelsregister, Vereinsregister, Partnerschaftsregister oder Genossenschaftsregister eingetragene Diensteanbieter verpflichtet das entsprechende Register sowie die Registernummer anzugeben (§ 5 Absatz 4 Satz 1 Telemediengesetz). Auch die Registernummer lässt sich aufgrund ihrer Struktur zuverlässig mit regulären Ausdrücken aus den Impressen extrahieren.

Enthält eine Impressumsseite mehrere Adressen, wählt der Algorithmus die erste (oberste) aus. Diese Auswahl resultiert aus der Annahme, dass in den Impressen zuerst Webseitenbetreiber:innen genannt werden und andere involvierte Unternehmen (z.B. Webagenturen, Rechtsanwälte) ebenfalls weiter unten auf der Webseite genannt sind. Die identifizierten Absätze wurden mit einem individuell vortrainierten NER-Modell klassifiziert. Der beschriebene Scrapingprozess kann parallelisiert ausgeführt werden, sodass pro Minute ca. 500 Seiten abgerufen, prozessiert und abgespeichert werden[1]. Entsprechend können selbst mehrere Millionen Webseiten innerhalb weniger Tage verarbeitet werden.

6.4.1 Named Entity Recognition (NER)

NER-Modelle weisen zur Erkennung von benannten Entitäten jedem Wort in einem Satz ein Label aus einer vordefinierten Menge von Labels zu. Eine einzelne Entität kann sich über mehrere Wörter erstrecken (LAMPLE et al. 2016, 2016). Typischerweise unterscheiden bestehende NER-Modelle zwischen Entitäten wie Personen, Organisationen oder Orten. Damit nehmen viele der bestehenden Modelle lediglich eine grobe Differenzierung der Entitäten vor. Beispielsweise fallen unter die Entität Ort sowohl Ländernamen als auch Ortsbezeichnungen oder Adressen. Für präzisiere und individuelle Zuordnungen von Entitäten ist die Entwicklung eines angepassten Modells notwendig, welches die Daten für aufsetzende Folgeverwertungen strukturiert. In der Geographie verwendete NER-Modelle werden auch Geoparser genannt (MEDAD et al. 2020). Das Geoparsing erfolgt dabei in zwei Schritten. Zunächst werden geographische Entitäten in unstrukturiertem Fließtext annotiert (Geotagging) und diese anschließend Koordinaten zugeordnet (Geocoding).

Das im Rahmen dieser Arbeit genutzte NER-Modell wurde trainiert, um die Entitäten Name, Straße, Ort und Postleitzahl aus den Impressumsangaben von Webseiten zu extrahieren. Das Tagging-Schema wurde so gestaltet, dass neben den für die Geokodierung der Adresse notwendigen Informationen auch der Name und die Rechtsform des Webseiten-Betreibers identifiziert wurde. Für das Training wurden reale Adressdaten von Unternehmenswebseiten verwendet. Die Trainingsdaten wurden gemäß dem IOB-Format annotiert (LAMPLE et al. 2016). Dementsprechend werden die Trainingsdaten wortweise annotiert, wobei Mehrwortentitäten ebenfalls über das Tagging-Schema annotiert werden können. Hierzu wird über den Präfix „I“ signalisiert, dass es sich bei dem betreffendem Wort über einen Teil einer Mehrworteinheit handelt. Während der Präfix „B“ dem Modell signalisiert, dass das betreffende Wort den Beginn einer neuen Entität darstellt.

[1] Für das Scraping wurde ein Office-Rechner verwendet mit 16 Gigabyte Arbeitsspeicher und einer Intel Core i7-2600 CPU.

Auf eine Vorverarbeitung der Textdaten wurde verzichtet, da insbesondere Satzzeichen wie Punkte, Kommata oder Bindestriche die Charakteristik der einzelnen Adressentitäten prägen. Für Modellentwicklung, Training und Evaluation wurde die Spacy-Bibliothek genutzt (MONTANI et al. 2022). Das Training des Modells wurde mittels Transfer Learning durchgeführt. Das vortrainierte Sprachmodell ist das Modell *de_core_news_lg* (SPACY 2022). Dieses wurde anhand von Textdaten aus Wikipedia (NOTHMAN et al. 2013) sowie aus dem deutschsprachigen TIGER-Korpus (BRANTS et al. 2004) vortrainiert. Das Modell ist für die CPU-Nutzung optimiert, sodass es effizient in das Web Scraping Framework eingebunden werden kann. Tabelle 6 zeigt einen Überblick über das verwendete Tagging-Schema.

Tabelle 6: Tagging-Schema des NER-Modells.

Tags	Beschreibung
B-NAME	Beginn eines Namens
I-NAME	Teil eines Namens
B-STR	Beginn eines Straßennamens
I-STR	Teil eines Straßennamens
B-PLZ	Beginn einer Postleitzahl
I-PLZ	Teil einer Postleitzahl
B-ORT	Beginn eines Ortsnamens
I-ORT	Teil eines Ortsnamens
O	Keine benannte Entität

Quelle: Eigene Darstellung.

Da die extrahierten Adressen bereits sehr stark vorstrukturiert sind und ein vortrainiertes Sprachmodell als Trainingsgrundlage verwendet wurde, reichten 843 Trainingsbeispiele aus, um eine zufriedenstellende Tagging-Genauigkeit zu erreichen. Tabelle 8 gibt einen Überblick über die Leistungskennzahlen des Modells. Precision, Recall und F1 sind Metriken zur Evaluation der Vorhersageleistung von Klassifikationsmodellen. Die Precision oder Spezifität ist der Anteil richtig-positiver Vorhersagen an der Summe richtig-positiver und falsch-positiver Vorhersagen. Somit gibt die Precision die Wahrscheinlichkeit an, mit welcher ein positiver Fall tatsächlich als ein solcher klassifiziert wird. Der Recall oder die Sensitivität ist der Anteil der richtig-positiven Vorhersagen an der Summe der richtig-positiven und falsch-negativen Vorhersagen. Analog zur Precision gibt der Recall die Wahrscheinlichkeit an, mit der ein negativer Fall auch als negativ eingestuft wird. Die F1-Metrik ist der harmonisierte Mittelwert aus Precision und Recall. Sie wird berechnet, indem das Produkt aus Precision und Recall durch deren Summe geteilt und anschließend mit zwei multipliziert wird. Die F1-Metrik belohnt daher Modelle mit ähnlichem Recall und ähnlicher Precision. Recall und Precision stehen in direkter Beziehung zueinander, da eine Erhöhung des Recalls immer mit einer Verringerung der Precision einhergeht und umgekehrt (GÉRON 2022).

Wie aus Tabelle 7 hervorgeht, ordnet das Modell den Impressumsangaben in fast neun von zehn Fällen erfolgreich die richtige Bezeichnung zu. Precision und Recall liegen in allen Fällen relativ nah beieinander, sodass von einer hohen Klassifikationsgüte sowohl hinsichtlich negativer als auch positiver Fälle ausgegangen werden kann. Darüber hinaus wird ebenfalls deutlich, dass die Vorhersagegenauigkeit je nach Entität variiert.

Tabelle 7: Performance-Metriken des NER-Modells.

Entität	F1	Precision	Recall
Insgesamt	88,81 %	90,37 %	87,26 %
Name	78,52 %	75,01 %	76,72 %
Straße	83,89 %	86, 15 %	81,75 %
Postleitzahl	98,59 %	99,29 %	97,91 %
Ort	96,50 %	97,87 %	95, 17 %

Quelle: Eigene Darstellung und Berechnung.

Während die F1-Metrik für die Entitäten Postleitzahl und Ort über 95 % liegen, beträgt diese für die Entität Straße circa 80 %. Der F1-Wert ist für die Entität Name etwas geringer. Hier erreicht das Modell nur eine Genauigkeit von knapp 80 %. Diese Unterschiede sind darauf zurückzuführen, dass Firmennamen im Vergleich zu den anderen Entitäten sehr heterogen und weniger stark strukturiert sind. Allerdings ordnete das Modell den Webseiten in deutlich mehr als 80 % der Fälle einen Namen zu. Häufig wurden Wörter, die vor oder nach dem korrekten Namen standen, ebenfalls als Teil des Namens klassifiziert. Diese Artefakte verringern zwar die Genauigkeit der Namensextraktion. Sie schränken die weitere Analyse jedoch nur geringfügig ein, da dennoch ausgelesen werden kann, ob der extrahierte Name die Rechtsform eines Unternehmens beinhaltet.

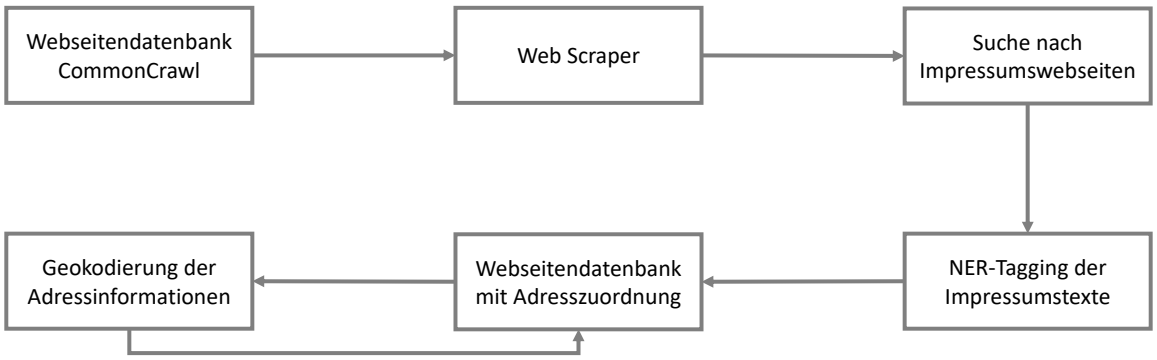
6.4.2 Geokodierung

Die zugeordneten Adressdaten der Impressen wurden im nächsten Schritt geokodiert. Ähnlich dem Scrapingprozess stellt auch die Geokodierung mehrerer Millionen Adressdaten eine Herausforderung dar. Da API-Lösungen in Abrufzahl und –geschwindigkeit limitiert sind, stellen solche out-of-the-box Lösungen für das Vorhaben keine praktikable Alternative dar. Vor diesem Hintergrund wurde das Open Source Gekodierungssystem Nominatim genutzt, welches auf OpenStreetMap (OSM) basiert (CLEMENS 2015). Da OSM auf der freiwilligen Kartierung von Straßen, Gebäuden und Plätzen basiert, besteht keine einheitliche Datenqualität bzw. –vollständigkeit. Validierungsstudien zeigen, dass die Datenqualität in dichter besiedelten Gebieten höher ist als in ländlichen Regionen (HELBICH 2012) und von Land zu Land variiert (MA et al. 2015). Aufgrund dieser Heterogenität besteht keine flächendeckende Qualitätsmetrik zur Bewertung der Vollständigkeit bzw. Genauigkeit von OSM-Daten. Dennoch kommen verschiedene Fallstudien zu dem Ergebnis,

dass OSM insgesamt eine zufriedenstellende Abdeckung bietet, unterschiedliche Anwendungsoptionen ermöglicht und insgesamt eine solide und vor allem kostenfreie Datengrundlage für GIS-Analysen darstellt (HAKLAY 2010; ZIELSTRA und ZIPF 2010; MOONEY und MINGHINI 2017). Für umfangreiche Geokodierungsvorhaben kann Nominatim lokal installiert werden, steht in verschiedenen Programmiersprachen zur Verfügung und basiert auf einer PostGIS-fähigen PostgresSQL-Datenbank, sodass diese effizient über einen Server angefragt werden kann (CLEMENS 2015). Ein einzelner Nominatimserver kann somit bis zu 30 Millionen Anfragen pro Tag verarbeiten (NOMINATIM 2022).

Die extrahierten und annotierten Adressdaten der Webseiten konnten demgemäß über eine lokale Installation des Nominatimservers effizient georeferenziert werden. Zur Geokodierung müssen zunächst die OSM-Daten für die gewünschte Region hinterlegt werden. Um lediglich mit in Deutschland ansässigen Unternehmen weiterarbeiten zu können, wurden die OSM-Daten für Deutschland hinterlegt. Somit werden Adressen außerhalb Deutschlands keine Koordinaten zugewiesen, sodass diese Fälle im Nachgang aus dem Datensatz entfernt werden konnten. Abbildung 27 fasst den Ablauf der methodischen Vorgehensweise zusammen.

Abbildung 27: Ablauf der methodischen Vorgehensweise.



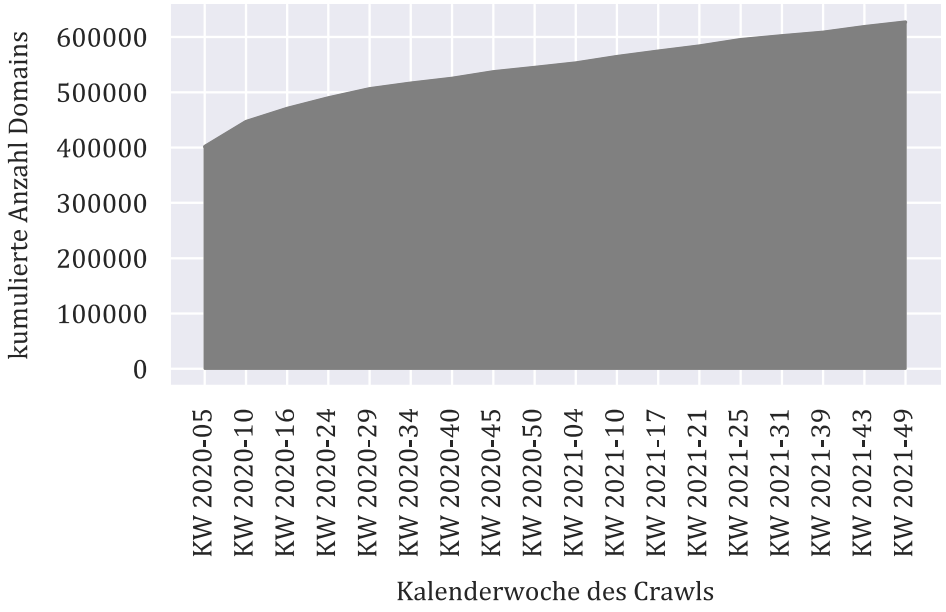
Quelle: Eigene Darstellung.

Ausgangspunkt bildete die vorgestellte Webseitendatenbank (CC). Auf Basis des Scrapy Frameworks wurde ein Web Scraper eingesetzt, welcher automatisiert nach Impressumswebseiten aller im CC enthaltenen Domains sucht. Falls eine solche Webseite gefunden wurde, rief der Scraper diese ab und mittels des NER-Taggers wurden die Adressdaten des Impressums automatisch annotiert. Die annotierten Adressdaten konnten somit zur Geokodierung verwendet werden. Auf Basis der identifizierten Unternehmensnamen und den extrahierten Handelsregisternummern konnten die Unternehmen abschließend aus dem Gesamtdatensatz gefiltert werden. Diese stellen somit die Grundlage für die weitere Analyse im folgenden Unterkapitel dar.

6.5 Analyse der identifizierten Unternehmenswebseiten

Insgesamt konnten aus den 9.544.441 untersuchten Domains 3.789.161 einzigartige Webseiten mit Impressumsangaben extrahiert werden. Um die in Deutschland wirtschaftsaktiven Entitäten zu erhalten, wurden anschließend alle Fälle entfernt, die weder eine Umsatzsteuer-Identifikationsnummer, eine Handelsregisternummer noch eine in Deutschland gültige Rechtsform beinhalten. Eine Handelsregisternummer wiesen 618.049 Fälle auf, während 761.555 Fälle eine Umsatzsteuer-Identifikationsnummer haben. Um die Unternehmen für die weitere Analyse zu selektieren, wurden die Unternehmen mit einer Eintragung ins Handelsregister ausgewählt und mit den Unternehmen, die in ihrem Impressum eine unternehmensbezogene Rechtsform angeben, ergänzt. Der finale Datensatz umfasst damit 627.141 Unternehmen.

Abbildung 28: Verteilung der Unternehmensdomains über die monatlichen Crawls.



Quelle: Eigene Darstellung.

Analog zu der Betrachtung des CC zeigt Abbildung 28 die Verteilung der identifizierten Unternehmendomains über die einzelnen Crawls. Bei der Betrachtung wird die noch markantere Ballung von Unternehmenswebseiten im ersten Crawl deutlich. Über 62 % der identifizierten Unternehmensdomains sind im ersten Crawl der Kalenderwoche *2020-05* enthalten. Bereits die ersten sechs Crawls enthalten über 80 % der extrahierten Unternehmensdomains, sodass die weiteren zwölf Crawls nur noch lediglich knapp 20 % weitere Domains enthalten. Auch hinsichtlich der Sprachen der Domains zeigt sich ein klares Bild: Die überwiegende Zahl der identifizierten Unternehmensdomains ist deutsch (71,7 %). Etwas mehr als jede fünfte Unternehmensdomain enthält deutsche und englische (21,1 %) bzw. englische und deutsche (4 %) Inhalte. Insgesamt lassen sich somit knapp 97 % der identifizierten Unternehmensdomains einer dieser drei Sprachkombinationen zuordnen.

Bevor der generierte Datensatz tiefer analysiert wird, gilt es zu klären, inwiefern dieser in der Lage ist, die tatsächliche Unternehmenslandschaft Deutschlands zu repräsentieren. Wie Tabelle 8 zu entnehmen ist, betreiben insgesamt 62 % der Unternehmen in Deutschland eine eigene Webseite (STATISTISCHES BUNDESAMT 2021b). Abhängig von Unternehmensgröße und –branche bestehen jedoch deutliche Disparitäten hinsichtlich des Webseitenbetriebs. Während große Unternehmen ab 250 Beschäftigten nahezu flächendeckend eine eigene Webseite unterhalten, liegt die Abdeckung bei Unternehmen mit weniger als zehn Beschäftigten teilweise deutlich unter 50 %. Mehr als neun von zehn Unternehmen mit zehn oder mehr Beschäftigten betreiben eine eigene Webseite. Markante Differenzen hinsichtlich der Webseitenabdeckung bestehen außerdem zwischen den einzelnen Wirtschaftszweigen. Unternehmen im Bereich der Informations- und Kommunikationstechnologie (IKT) haben besonders häufig eine eigene Webseite. Auch Unternehmen des verarbeitenden Gewerbes, des Handels und der Reparatur von Datenverarbeitungs- und Telekommunikationsgeräten haben in mehr als 70 % der Fälle eine eigene Webseite. Wirtschaftszweige mit geringer Abdeckung sind Verkehr, Lagerei, Post-, Kurier- und Expressdienste, Grundstücks- und Wohnungswesen sowie Unternehmen des Wirtschaftszweigs Energie- und Wasserversorgung, Abwasser- und Abfallentsorgung und Beseitigung von Umweltverschmutzungen.

Tabelle 8: Unternehmen mit Webseite nach Wirtschaftszweigen und Beschäftigten.

Wirtschaftszweig	**Unternehmen mit ... bis ... Beschäftigten**				
	Gesamt	**1-9**	**10-49**	**50-249**	**249+**
	Anteil in % an allen Unternehmen				
Insgesamt	62	59	87	93	97
Verarbeitendes Gewerbe	77	70	91	96	97
Energie- und Wasserversorgung, Abwasser- und Abfallentsorgung und Beseitigung von Umweltverschmutzungen	45	42	90	98	99
Baugewerbe	53	48	86	96	98
Handel; Instandhaltung und Reparatur von Kraftfahrzeugen	71	68	87	94	100
Verkehr, Lagerei, Post-, Kurier- und Expressdienste	41	32	66	85	92
Gastgewerbe	63	58	87	88	99
Information und Kommunikation	84	82	98	98	97
Grundstücks- und Wohnungswesen	42	41	89	100	90
Erbringung von freiberuflichen, wissenschaftlichen und technischen Dienstleistungen	68	66	94	95	96
Erbringung von sonstigen wirtschaftlichen Dienstleistungen	64	60	82	86	97
Reparatur von Datenverarbeitungs- und Telekommunikationsgeräten	75	73	98	*	*

*Quelle: STATISTISCHES BUNDESAMT (2021b); * entspricht keiner Angabe, da Zahlenwert nicht sicher genug.*

Tabelle 8 macht deutlich, dass ein Web Mining-Ansatz zur Identifizierung deutscher Unternehmen nur eine eingeschränkte Stichprobe aus der Grundgesamtheit aller deutschen Unternehmen darstellt. Insgesamt waren im Jahr 2020 rund 3,3 Millionen Unternehmen in Deutschland ansässig. Ein Großteil der Unternehmen stellen dabei Einzelunternehmer:innen (knapp 2 Millionen Einheiten) dar. Die restliche Unternehmenslandschaft setzt sich darüber hinaus aus Personengesellschaften (ca. 400.000 Einheiten), Kapitalgesellschaften (ca. 760.000 Einheiten) und sonstigen Rechtsformen (ca. 200.000 Einheiten) zusammen (STATISTISCHES BUNDESAMT 2021a).

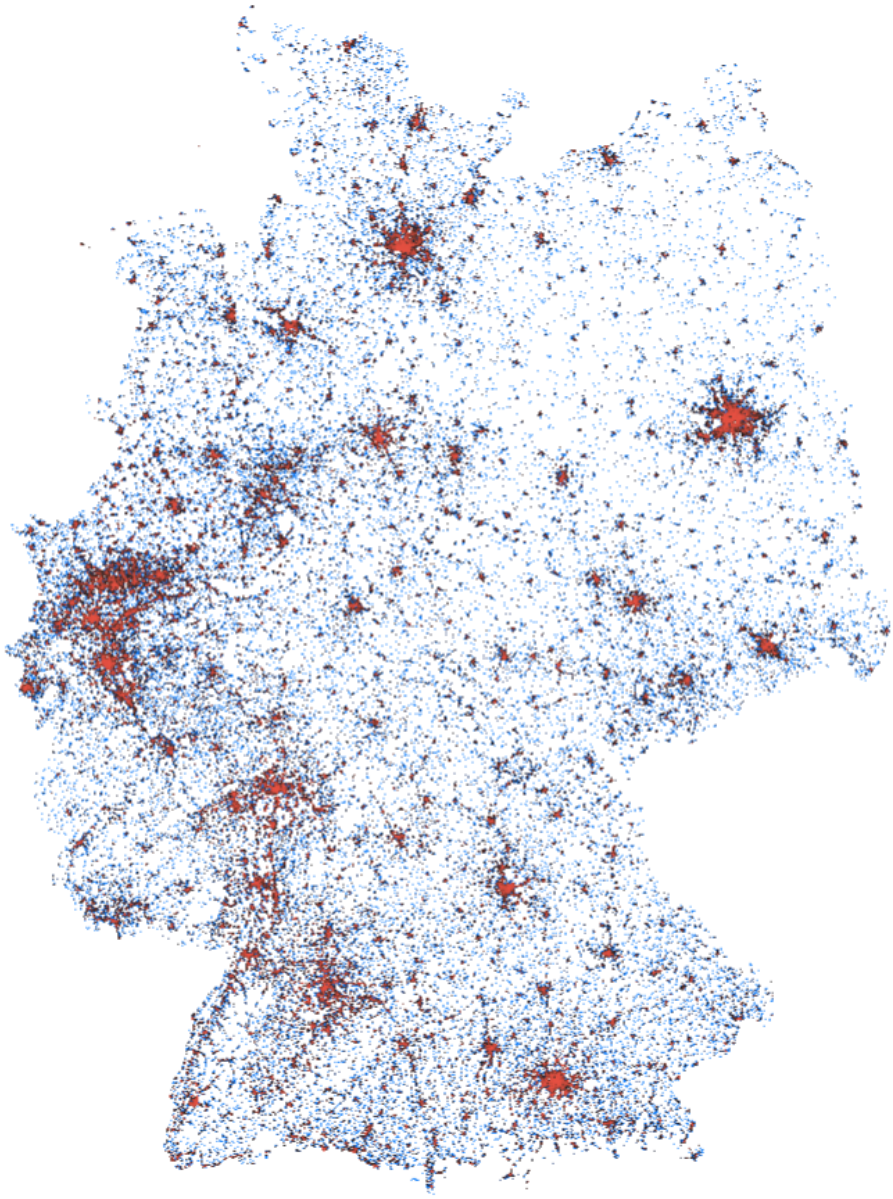
Die Gruppe der Einzelunternehmer:innen lässt sich nochmals untergliedern in in das Handelsregister eingetragene Kaufleute und Einzelunternehmer:innen, die nicht in das Handelregister eingetragen sind. Nicht in das Handelregister eingetragenen Einzelunternehmer:innen sind entsprechend weder Rechtsformen noch eine Handelsregisternummern zugeordnet, sodass diese Gruppe der Unternehmen nicht im Rahmen dieser Analyse untersucht werden kann. Somit lässt sich die tatsächliche Grundgesamtheit der Handelsregisterunternehmen mit eigener Webseite nur grob approximieren, da unbekannt ist, wieviele Einzelunternehmer:innen einen Handelsregistereintrag haben. Die theoretische Grundgesamtheit setzt sich somit aus den Personengesellschaften, den Kapitalgesellschaften und Organisationen sonstiger Rechtsformen sowie den eingetragenen Einzelunternehmer:innen zusammen.

Darüber hinaus kann die Anzahl der identifizierten Unternehmensdomains mit anderen Web Mining-Studien deutscher Unternehmen verglichen werden. Die Datengrundlage dieser Arbeiten stellt jeweils das Mannheimer Unternehmenspanel dar, welches als umfangreichste Unternehmensdatenbank außerhalb der amtlichen Statistik gilt (ZENTRUM FÜR EUROPÄISCHE WIRTSCHAFTSFORSCHUNG 2022). 2021 waren 1.155.867 Unternehmen des Unternehmenspanels mit Webadressen attributiert (ABBASIHAROFTEH et al. 2021), sodass zunächst von einer relativ großen Unterdeckung des vorliegenden Datensatzes auszugehen ist. Die durchgeführten empirischen und georeferenzierten Untersuchungen, die auf Basis des Mannheimer Unternehmenspanels durchgeführt wurden, verwenden allerdings 633.523 (ABBASIHAROFTEH et al. 2021), 684.873 (KRÜGER et al. 2020) sowie 685.057 Unternehmen (KINNE und LENZ 2021) als Datengrundlage. Als Gründe für diese massive Reduzierung der tatsächlich untersuchten Unternehmensdomains führen die Autor:innen potentiell fehlerhafte Weiterleitungen und fehlerhafte Downloads an (KINNE und AXENBECK 2020). Somit ist davon auszugehen, dass die absolute Anzahl deutscher Unternehmensdomains, welche als Grundlage für raumbezogene, Web Mining-Analysen herangezogen werden kann, zwischen 600.000 und 700.000 Unternehmen liegt. Daher kann für die weitere Analyse von einer ausreichenden Repräsentativität der im Rahmen dieser Arbeit identifizierten Unternehmen ausgegangen werden.

Weiterhin lässt sich prüfen, inwiefern die geographische Verteilung der identifzierten Unternehmen mit eigener Webseite mit der tatsächlichen geographischen Verortung sämtlicher Betriebe in Deutschland übereinstimmt. Hierzu wurden die identifizierten Unternehmen auf Gemeindeebene aggregiert und mit offiziellen Statistiken der Bundesagentur für Arbeit verglichen (STATISTIK DER BUNDESAGENTUR FÜR ARBEIT 2021). In den Sekundärdaten sind Zahlenwerte kleiner drei aufgrund der Geheimhaltungspflicht von Sozialdaten anonymisiert. Im vorliegenden Datensatz betrifft dies 1.945 Fälle. Um diese dennoch in der Analyse berücksichtigen zu können, wurden die Werte mit der Zahl der identifizierten Unternehmensdomains verglichen. Anonymisierte Werte in den Sekundärdaten wurden, falls verfügbar, dann den webgenerierten Zahlen gleichgesetzt. Wenn in beiden Datensätzen keine Zahlenwerte existierten, wurde von fehlenden Werten ausgegangen. Der Korrelationskoeffzient nach Pearson beträgt 0,993 und bescheinigt den Daten einen höchst signifikanten ($p< 0.001$) und nahezu perfekten positiven Zusammenhang. Entsprechend kann davon ausgegangen werden, dass die identifizierten Unternehmen zumindest in der geographischen Verteilung ein realistisches Abbild der realen Unternehmenslandschaft in Deutschland darstellen.

Abbildung 29 zeigt die geographische, koordinatenscharfe Verteilung der identifizierten Unternehmen. Deutliche Ballungen sind insbesondere in den Großstädten Berlin, Hamburg, München, Frankfurt, Dresden, Hannover, Bremen und Stuttgart zu beobachten. Weiterhin wird die polyzentrische Siedlungsstruktur im Ruhrgebiet, im Rhein-Main-Gebiet und in der Rhein-Neckar-Region deutlich. Wie aus der Betrachtung von Abbildung 28 hervorgeht, bestehen markante Disparitäten in der geographischen Verteilung der Unternehmen. Im Norden, Osten und im Zentrum Deutschlands sind deutlich weniger Unternehmen zu erkennen als im Süden und Westen.

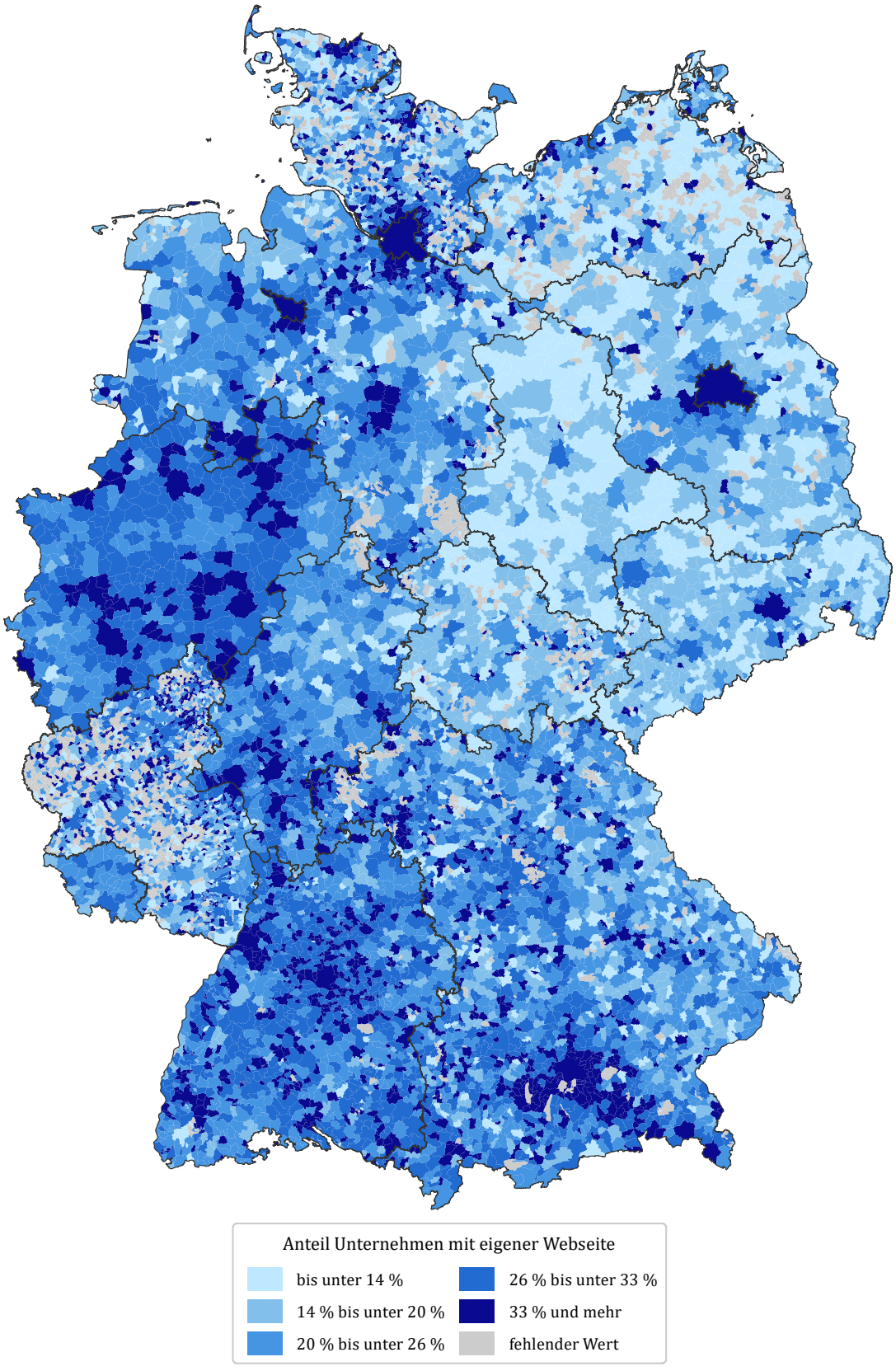
Abbildung 29: Verteilung der identifizierten Unternehmen.



Quelle: Eigene Darstellung.

Neben der unternehmensscharfen geographischen Betrachtung stellt sich die Frage, inwieweit es regionale Unterschiede hinsichtlich des Anteils der Unternehmen mit einer eigenen Website an allen Unternehmen pro Gemeinde gibt. Abbildung 30 zeigt die Webseiten-Abdeckung aggregiert auf Gemeindeebene. Der Mittelwert der Verteilung liegt bei 23,8 % und der Median bei 23 %, sodass von einer gleichmäßigen Verteilung der Werte ausgegangen werden kann. Die Standardabweichung liegt bei 0,11 und signalisiert ebenfalls eine relativ ausgeglichene Werteverteilung. In 1432 Gemeinden wurden keine Unternehmensdomains verortet. Dabei handelt es sich um kleine Gemeinden mit weniger als 300 Einwohner:innen sowie unbewohnte gemeindefreie Gebiete.

Abbildung 30: Anteil von Unternehmen mit Webseite auf Gemeindeebene.



Quelle: Eigene Darstellung.

Markantere Unterschiede fallen bei der Betrachtung der geographischen Verteilungsmuster auf. Besonders auffällig sind die Unterschiede zwischen Ost- und Westdeutschland. Außerdem weisen die Großstädte Berlin, München, Stuttgart und Hamburg sowie die Großstädte des Ruhrgebiets relativ hohe Werte auf. Die räumlichen Disparitäten gelten demnach nicht nur in den absoluten Verteilungen aus Abbildung 29, sondern auch in der relativen Betrachung im Verhältnis zu der Gesamtzahl aller Unternehmen pro Gemeinde, dargestellt in Abbildung 30. Somit lässt sich vermuten, dass signifikante Unterschiede hinsichtlich Struktur und Lage zwischen den unterschiedlichen Raumtypen bestehen.

Um die regionalen Unterschiede hinsichtlich Struktur und Lage der Gemeinden statistisch prüfen zu können, wurden die Gemeinden nach den vom Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) ausgewiesenen Gemeindetypen klassifiziert (BBSR 2020). Anschließend wurde eine Welch-ANOVA durchgeführt. Diese zeigte signifikante Mittelwertunterschiede zwischen den jeweiligen Gemeindetypen (p-Wert < 0,001). Um feststellen zu können, zwischen welchen Gemeindetypen signifikante Unterschiede in der Webseitenabdeckung bestehen, wurde im Anschluss ein Games-Howell post-hoc Test durchgeführt. Tabelle 9 zeigt die Ergebnisse des Tests.

Tabelle 9: Mittelwertunterschiede zwischen Strukturtypen.

Raumtyp A	Raumtyp B	Mittelwert Raumtyp A	Mittelwert Raumtyp B	p-Wert
Ländlich	Teilweise städtisch	24,15 %	24,91 %	0,121
Ländlich	Überwiegend städtisch	24,15 %	27,08 %	0,001***
Teilweise städtisch	Überwiegend städtisch	24,91 %	27,08 %	0,001***

Quelle: Eigene Darstellung.

Während sich zwischen den Raumtypen *ländlich* und *teilweise städtisch* keine signifikanten Mittelwertunterschiede beobachten lassen, sind statistisch signifikante Unterschiede hinsichtlich der Webseitenabdeckung zwischen den *überwiegend städtischen* Gemeinden und den anderen Raumtypen festzustellen. Es lässt sich daher festhalten, dass insbesondere in überwiegend städtischen Gebieten der Anteil von Unternehmen mit eigener Webseite signifikant höher ist als in anderen Gebietstypen.

Neben ihrer Struktur lassen sich Gemeinden auch hinsichtlich ihrer Lage analysieren. Hierfür wurden die Gemeinden nach den vom BBSR vorgegebenen Lagetypen segregiert. Analog zur Untersuchung des Strukturtyps wurde auch für die unterschiedlichen Lagetypen eine Welch-Anova durchgeführt, um die Mittelwertunterschiede in der Webseitenabdeckung statistisch zu prüfen. Die Ergebnisse der Varianzanalyse zeigten erneut signifikante Mittelwertunterschiede zwischen

den Lagetypen (p-Wert < 0,001). Tabelle 11 zeigt analog zu Tabelle 10 die Ergebnisse des im Anschluss durchgeführten Games-Howell post-hoc Tests.

Wie Tabelle 10 zu entnehmen ist, sind signifikante Mittelwertunterschiede zwischen allen vier Lagetypen zu beobachten. Die sehr zentral gelegenen Gemeinden nehmen dabei die höchsten Mittelwerte hinsichtlich der Webseitenabdeckung an, während in sehr peripheren Gemeinden lediglich knapp jedes fünfte Unternehmen eine eigene Webseite betreibt.

Tabelle 10: Mittelwertunterschiede zwischen Raumtypen.

Raumtyp A	Raumtyp B	Mittelwert Raumtyp A	Mittelwert Raumtyp B	p-Wert
peripher	sehr peripher	24,10 %	20,15 %	0,001***
peripher	sehr zentral	24,10 %	28,60 %	0,001***
peripher	zentral	24,10 %	26,61 %	0,001***
sehr peripher	sehr zentral	20,15 %	28,60 %	0,001***
sehr peripher	zentral	20,15 %	26,61 %	0,001***
sehr zentral	zentral	28,60 %	26,61 %	0,001***

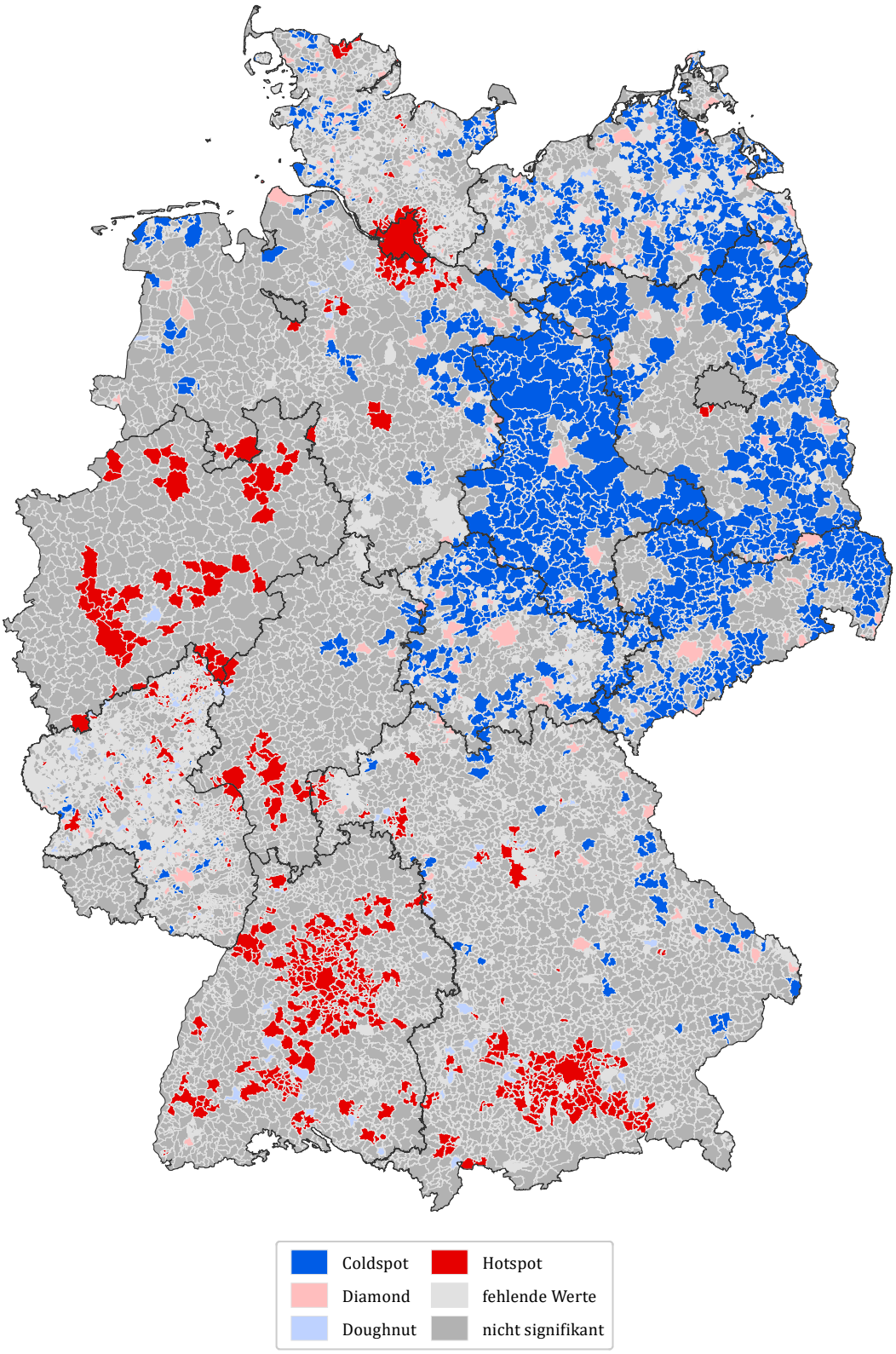
Quelle: Eigene Darstellung.

Zusammenfassend lässt sich also festhalten, dass zwischen den Struktur- und Lagetypen der Gemeinden signifikante Unterschiede in der Webseitenabdeckung von Unternehmen bestehen. Insgesamt haben Unternehmen in städtischeren bzw. zentraleren Gemeinden häufiger eine eigene Webseite als in periphereren bzw. ländlicheren Gebieten. Diese Schlussfolgerung ist dabei zumindest in Teilen mit branchen- und betriebsgrößenspezifischen Faktoren zu erklären.

Über die Untersuchung der Webseitenabdeckung nach Gemeindetyp können auch räumliche Verteilungsunterschiede in der Abdeckung untersucht werden. Durch die Berechnung der Moran's I Statistik kann getestet werden, ob im Untersuchungsgebiet eine räumliche Autokorrelation vorliegt (MORAN 1950). Hierbei konnte eine signifikante (p-Wert < 0,001) und leicht positive Autokorrelation (Moran's I: 0,1413) beobachtet werden. Entsprechend ist die Webseitenabdeckung der Unternehmen einer Gemeinde im Untersuchungsgebiet nicht zufällig räumlich verteilt, sondern wird signfikant von den Werten benachbarter Gemeinden beeinflusst. Auf Basis des Local Moran kann pro Region im Untersuchungsgebiet ein gesonderter Wert ermittelt werden, der Richtung und Stärke der räumlichen Autokorrelation beschreibt. Somit lässt sich quantifizieren, inwiefern Grad und Ausprägung der räumlichen Autokorrelation über das Untersuchungsgebiet hinweg variieren. Methodisch wird dabei für jeden dieser Werte die Standardabweichung ermittelt, welche mit einer Teststatistik verknüpft ist. Statistisch signifikante Beziehungen benachbarter Regionen können somit identifiziert und kartographisch visualisiert werden.

Abbildung 31 zeigt die sogenannten *Local indicators for spatial association* (ANSELIN 1995). Lokale Hotspots - also Nachbarschaften mit hoher Webseitenabdeckung – sind insbesondere im Großraum München, in der Rhein-Neckar-Region, im Rhein-Main-Gebiet, in Nordrhein-Westfalen, Hannover sowie im Hamburger Großraum zu beobachten. Signifikant autokorrelierte Regionen mit niedriger Abdeckung - sogenannte Cold-Spots - von Unternehmenswebseiten sind schwerpunktmäßig im Osten Deutschlands feststellbar. Speziell die Regionen an den östlichen Landesgrenzen sowie an der ehemaligen innerdeutschen Grenze sind signifikant miteinander korreliert und die Werte der Attributvariablen weisen geringe Werte auf. Signifikante Ausreißer sind im Untersuchungsgebiet nur vereinzelt zu beobachten. Doughnuts – Gemeinden mit niedrigen Werten in Nachbarschaft mit Gemeinden mit hohen Werten – sind hauptsächlich kleinere Städte und Gemeinden unter 10.000 Einwohner:innen. Die meisten dieser Gemeinden liegen in Baden-Württemberg im Umkreis von kreisfreien Städten wie z.B. Leimen, Sulz am Neckar oder Weinsberg. Bei Diamonds handelt es sich um Gemeinden, die selbst über eine hohe Webseitenabdeckung verfügen, jedoch signifikant negativ mit ihren Nachbargemeinden korreliert sind. Die größten Gemeinden dieser Kategorie sind im Untersuchungsgebiet Großstädte in Ostdeutschland, wie beispielsweise Magdeburg, Halle an der Saale, Erfurt oder Chemnitz. Bei den sonstigen Diamanten handelt es sich vor allem um sehr kleine Gemeinden mit weniger als 1.000 Einwohner:innen. Insgesamt lässt sich somit festhalten, dass lokale Hotspots eher in starken wirtschaftlichen Regionen zu finden sind, während Coldspots in zusammenhängenden strukturschwachen Gebieten erkennbar sind.

Abbildung 31: Lokale Indikatoren räumlicher Autokorrelation.



Quelle: Eigene Darstellung.

7 Fallstudie 2: Identifizierung und Standortanalyse von KI-Unternehmen

Nachdem die vorherige Fallstudie den komplexen Datenerhebungsprozess von Webmassendaten skizziert und die geographische Repräsentativität der Daten evaluiert hat, soll die folgende Fallstudie Möglichkeiten zur inhaltlichen Auseinandersetzung mit Textdaten aufzeigen. Als Beispiel wird dabei die Identifizierung und Standortanalyse deutscher KI-Unternehmen genutzt. Hierzu werden zunächst klassische wirtschaftsgeographische Erklärungsansätze zur Standortwahl von Unternehmen beleuchtet. Anschließend wird die methodische Vorgehensweise zur Identifizierung von deutschen KI-Unternehmen auf Basis von Webtexten vorgestellt. Aufbauend auf dieser Identifizierung folgt eine Standortanalyse, welche die Standortmuster auf Kreisebene tiefer beleuchtet. Um der räumlichen Granularität der gewonnenen Daten Rechnung zu tragen, folgt eine Mikrobetrachtung zweier KI-Cluster.

7.1 Problemstellung und Hintergrund

Technologische Innovationen werden seit der Industrialisierung als wichtiger Treiber von Wirtschaftswachstum und Unternehmenserfolg beschrieben. IKT-Technologien gelten als besonders wertschöpfend, sowohl direkt als auch indirekt in Form von Spillovern und Prozessoptimierungen (MORETTI 2012; AUDRETSCH und FELDMAN 1996). Somit leisten diese Technologien einen besonderen Beitrag zur langfristigen Entwicklung von Regionen und Volkswirtschaften (ROMER 1990; LUCAS 1988; COOKE et al. 2004).

Innerhalb der jüngsten technologischen Innovationen wird KI nochmals eine besondere Relevanz für die zukünftige Wirtschaftsentwicklung zugesprochen, da sie durch die Generierung von Effizienzgewinnen generisch einsetzbar ist und daher als Basisinnovation gesehen werden kann (PARKES und WELLMAN 2015; BIANCHINI et al. 2020; KLINGER et al. 2018; TRAJTENBERG 2019). Angefangen im Transportwesen über Robotik, das Gesundheitswesen, den Bildungssektor, die freie Wirtschaft bis hin zur öffentlichen Verwaltung finden KI-Technologien Anwendung und werden daher von Wissenschaftler:innen, Politiker:innen und Unternehmer:innen als besonders richtungsweisende Technologie wahrgenommen (STONE et al. 2016; AGHION et al. 2017). Die Integration von Basisinnovationen beinhaltet somit das Potential ganze Branchen im Sinne der schumpeterschen schöpferischen Zerstörung zu transformieren (SCHUMPETER 1939) und darüber hinaus Arbeitsmärkte nachhaltig zu verändern (BRYNJOLFSSON und MITCHELL 2017). Ferner geht aktuelle Forschung davon aus, dass KI-Methoden die Entdeckung weiterer Innovationen beschleunigen und als eine Art Innovationskatalysator gesehen werden können (COCKBURN et al. 2019; AGRAWAL et al. 2019).

Wirtschaftsgeographische Auseinandersetzungen mit der Entstehung von Basisinnovationen orientieren sich häufig an der Produktlebenszyklus-Hypothese (VERNON 1992) sowie der Theorie der langen Wellen (DICKEN 1998). Dabei variieren die Standortfaktoren, die eine Entwicklung der Technologie begünstigen, über Zeit. Bei der Einführung neuer Technologien spielt das räumlich verortete und verwandte Wissen eine entscheidende Rolle, da dieses rekombiniert bzw. umfunktioniert und somit die neue Technologie leichter integriert werden kann (FRENKEN et al. 2007; HIDALGO und HAUSMANN 2009).

Insbesondere der Zugang zu Wissen und Humankapital – also in Personen gebundenenes Wissen – wird im wirtschaftsgeographischen Diskurs als ausschlaggebender Faktor für technologische Innovationen gesehen. Wissen stellt ein abstraktes Konstrukt dar, welches in unterschiedlichen Formen produziert und transferiert werden kann. Nach DAVENPORT und PRUSAK (1998) ist Wissen eine Mischung aus Erfahrungen, Werten, kontextualisierten Informationen und Expertenwissen, das innerhalb einer Person entsteht und auch dort angewendet wird. Innerhalb von Organisationen und Unternehmen liegt Wissen daher nicht nur in Form von Dokumenten vor, sondern auch in organisationsbezogenen Routinen, Praktiken, Prozessen und Normen (DAVENPORT und PRUSAK 1998).

Somit kann Wissen neben expliziten, kodifizierten Formen auch implizite Gestalt annehmen. POLANYI (2012) beschreibt implizites Wissen (tacit knowledge) als nicht verbalisierbares Wissen. Es ist daher unmittelbar an Personen gebunden und entsprechend nur durch persönliche Kontakte und informelles Lernen transferierbar (HOWELLS 2002). Folglich ist der Wissenstransfer von implizitem Wissen nicht in beliebigen Kontexten möglich. Persönliche, kulturelle und sprachliche Barrieren können daher eine räumlich weitreichende Verteilung von implizitem Wissen behindern (LUNDVALL 2012).

Vor diesem Hintergrund haben sich in der Wirtschaftsgeographie in den letzten Dekaden verschiedene Theorien entwickelt, welche die Rolle von räumlicher Nähe und die Interaktion unterschiedlicher Akteure im Raum für Wissensproduktion und –transfer beschreiben. Beispielsweise zeigen verschiedene Ausgestaltungen der Clustertheorie (PORTER 1990; AUDRETSCH und FELDMAN 1996; GLAESER et al. 2010), dass Industriecluster zu höherer regionaler Wirtschaftsleistung, mehr Arbeitsplätzen, höheren Patentierungsraten und zu vermehrten Gründungen neuer Unternehmen führen können (DELGADO et al. 2016).

Ein Großteil der bestehenden Clusterdefinitionen geht auf die Arbeit von MARSHALL (1890) zurück, der Wissensspillover, Arbeitsmarktpooling sowie Beziehungen zwischen Kund:innen und Lieferant:innen als zentrale Lokalisationsvorteile von Industriedistrikten beschreibt. MARSHALLS Agglomerationsvorteile bilden somit auch für die bekannteste Deutung des Clusterbegriffs die definitorische Grundlage. Diese wurde von PORTER (1990) vorgestellt und beschäftigt sich vor allem

mit der räumlichen Ausgestaltung von Clustern. Nach PORTER (1990) sind Cluster geprägt durch die räumliche Konzentration von Unternehmen eines Technologiefeldes und komplementär-unterstützenden Organisationen (z.B. Universitäten, Forschungseinrichtungen, Verbände, Dienstleister, Verkehrsträger), die zur Wettbewerbsfähigkeit des Clusters beitragen. Damit bieten Technologiecluster den notwendigen Nährboden für die Einführung neuer Technologien wie KI.

KERR und ROBERT-NICOUD (2020) argumentieren, dass diese Faktoren auch in modernen Technologieclustern eine zentrale Rolle spielen. Die Autor:innen führen an, dass insbesondere die IKT-Branche technisch affine Akademiker:innen benötigt, um moderne Digitaltechnologien entwickeln und einsetzen zu können. Gleichzeitig bietet die Agglomeration von Unternehmen einer Branche auch für die Arbeitskräfte Vorteile. Erstens sind sie abgesichert gegen wirtschaftliche Krisen einzelner Unternehmen, zweitens sorgt die erhöhte Unternehmensdichte für einen tieferen Arbeitsmarkt und drittens spezialisierte Weiterbildungsangebote. HELSLEY und STRANGE (2002) beobachten daher eine hohe Übereinstimmung zwischen dem Arbeitskräfteangebot und den nachgefragten Qualifikationen innerhalb von Technologieclustern. Darüber hinaus betont die Literatur die Relevanz sogenannter Ankerfirmen für die Entwicklung von Clustern (FELDMAN 2003; AGRAWAL und COCKBURN 2003). Sie tragen dazu bei, dass sich vor Ort ein spezialisierter Arbeitsmarkt mit ausgebildetem Personal entwickelt, sich verwandte Industrien ansiedeln und bilden damit einen Ausgangspunkt für Wissenspillover (FELDMAN 2003). Andererseits siedeln sich rund um diese Ankerfirmen kleinere Firmen an, welche als Dienstleister für die großen Firmen fungieren (AGRAWAL et al. 2014). Gleichzeitig formen große Unternehmen das Wissensprofil einer Region und können somit einen wichtigen Ausgangspunkt für die weitere technologische Entwicklung von Regionen darstellen (NEFFKE et al. 2011; HIDALGO et al. 2018).

Die Literatur zu regionalen Innovationssystemen greift die grundlegenden Elemente der Clustertheorie auf, betont jedoch stärker die Relevanz institutioneller regionaler Aspekte wie die staatliche Technologie- und Innovationspolitik oder Kooperationen zwischen Wirtschaft und Wissenschaft. Akteursseitig betrachten regionale Innovationssysteme somit ebenfalls neben Unternehmen komplementäre und unterstützende Einrichtungen wie Forschungsinstitute, Universitäten oder Technologietransferorganisationen. Diesem Verständnis nach bestehen regionale Innovationssysteme sowohl aus Komponenten der Wissensgenerierung und –diffusion als auch aus Elementen der Wissensanwendung und –verwertung (COOKE 2001).

Ein zentraler Aspekt regionaler Innovationssysteme sind demnach interaktive Lernprozesse, die einen Wissens- und Technologietransfer ermöglichen. Ausgelöst werden diese einerseits durch Kooperationen sowie gemeinschaftliche F &E, andererseits durch informellen Austausch und die Kopräsenz verschiedener Akteure. Akteure, die in das regionale Innovationssystem eingebunden sind, profitieren vom vorhandenden impliziten Wissen und dem sogenannten *local buzz* (BATHELT

et al. 2004; BATHELT und TURI 2011). Durch dieses „lokale Rauschen" entsteht ein Transfer von implizitem Wissen durch persönliche Kontakte, welcher einen relevanten Wissensvorteil für Gründungen und Innovationen darstellt. Im Umkehrschluss profitiert auch das Innovationssytem durch die Neugründungen und Innovationen. Neben lokalen Austausch- und Lernprozessen betonen BATHELT et al. (2004) die Relevanz von *global pipelines*, die zusätzliches Wissen in die Region transportieren und dort verbreiten.

Wissenschaftliche Beiträge zur Lokalisation und Analyse von Unternehmen in bestimmten Technologiefeldern greifen in der Regel auf Patentdaten zurück, um Innovationsaktivitäten verstehen und räumlich analysieren zu können (BALDINI et al. 2007; HALL 2022). Beispielsweise nutzen COCKBURN et al. (2019) sowie FUJII und MANAGI (2018) IPC-Codes, um KI-Innovatoren zu identifizieren. Patentdaten offerieren zwar Einblicke in unternehmensbezogene Innovationsaktivitäten. Allerdings beziehen sie sich lediglich auf patentierbare Innovationen (GONZALEZ 2006). Insbesondere für die Identifizierung von KI-Unternehmen können Patentdaten lediglich ein unvollständiges Bild zeichnen. Speziell im Softwarekontext werden nur die wenigsten Innovationen tatsächlich patentiert und ein Großteil der Unternehmen greifen bei ihren KI-Anwendungen auf das geistige Eigentum anderer zurück. RAMMER et al. (2022) zeigen beispielsweise, dass lediglich jedes dritte Unternehmen, welches KI anwendet selbst KI-bezogene Patente hält. Weiterhin ist KI als Basisinnovation zu verstehen, sodass KI-gestützte Innovationen in unterschiedlichsten Patentklassen auftreten können.

Klassische Sekundärstatistik kann bei der Identifizierung von Unternehmen, die KI einsetzen, ebenfalls nicht weiterhelfen. Die Daten sind sowohl räumlich als auch sektoral hoch aggregiert, sodass eine granulare Zuordnung weder räumlich noch technologisch möglich ist. Ensprechend wird der Bedarf an quantitativen unternehmensbezogenen Informationen zur KI-Nutzung, die eine wissenschaftliche Analyse zu Innovations- und Diffusionsprozessen ermöglichen, immer größer (RAJ und SEAMANS 2019). Auch politisch besteht erhöhter Informationsbedarf hinsichtlich der Ausbreitung von KI, da bis dato Ungewissheit herrscht, inwiefern sich die Implementierung von KI auf die Produktivität von Unternehmen auswirkt, welche Unternehmen KI verstärkt einsetzen, welche Auswirkungen für den Arbeitsmarkt erwartbar sind oder ob die KI-Implementierung zu einer veränderten Strategieausrichtung führt (RAJ und SEAMANS 2019).

Um diese Fragen präzise beantworten zu können, bedarf es technologisch und geographisch fein aufgelöster Daten. Da etablierte Indikatorik wie beschrieben nur unzureichend bei einer quantitativen Identifizierung von KI-Unternehmen helfen kann, nutzt diese Fallstudie Unternehmenswebseiten als alternative Datenequelle, um die KI-Nutzung von Unternehmen kartieren und analysieren zu können.

7.2 Methodische Vorgehensweise

Als Datengrundlage dient dabei der in Kapitel 6 erstellte Datensatz deutscher Unternehmenswebseiten. Da einige Unternehmen extrem umfassende Webauftritte mit mehreren Tausend Webseiten pflegen, ist es notwendig bereits beim Download der Webseiten sehr selektiv vorzugehen. Bis dato besteht hinsichtlich der Selektion von Webseiten zur Textverarbeitung keine etablierte Methodik, vielmehr ist der Auswahlprozess stark von dem Untersuchungsgegenstand abhängig, sodass fragestellungsabhängig individuelle Verfahren entwickelt werden. Zur Textanalyse von Unternehmenswebseiten schlagen KINNE und LENZ (2021) vor, die Webseiten mit den 25 kürzesten URLs zu selektieren. Dieser Vorschlag ist das Ergebnis einer umfassenden Studie von KINNE und AXENBECK (2020), die die URL-Anzahl deutscher Unternehmenswebseiten untersucht. Die Studie belegt, dass die Medianananzahl von URLs pro Unternehmensdomain bei rund 15 liegt, sodass bei einem Limit von 50 URLs pro Unternehmensdomain für zwei Drittel aller Unternehmensdomains alle URLs vollständig heruntergeladen sind (KINNE und AXENBECK 2020). Die Wahl der kürzesten URLs begründen KINNE und LENZ (2021) damit, dass allgemeinere Informationen zu Unternehmen auf den vorderen Webseiten einer Domain zu finden sind.

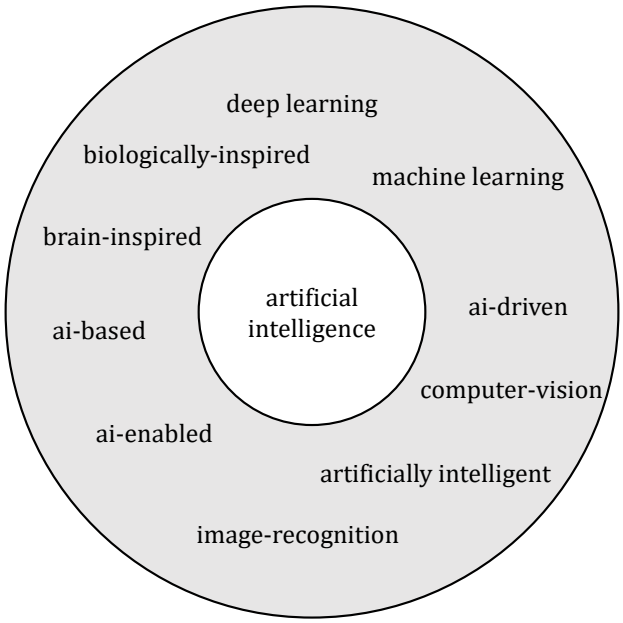
Zur Textgenerierung der vorliegenden Untersuchung wurde eine ähnliche Heuristik zur Webseitenselektion verwendet. Der Scraper startet auf der Startseite des jeweiligen Unternehmens und lädt die Inhalte aller internen URLs herunter, die auf der Startseite gelistet sind. Der Scraper ignoriert dabei automatisch URLs, die einen Dateidownload - z.B. von PDF-Dateien, Videos oder Fotos – auslösen würden. Diese Inhalte werden somit im Rahmen dieser Arbeit nicht betrachtet. Insgesamt wurden 12.981.819 Webseiten heruntergeladen. Dies entspricht einem Mittelwert von 20,7 Webseiten pro Unternehmensdomain. Die Medianananzahl liegt deutlich unterhalb des Mittelwerts bei 12,7 Webseiten pro Domain. Damit bestätigen die Ergebnisse die Pilotstudie von KINNE und AXENBECK (2020) und verdeutlichen markante Unterschiede hinsichtlich der Webseitengrößen von Unternehmen.

Bereits während des Downloads wurde neben des HTML-Inhalts auch der zentrale Webseitentext jeder URL abgespeichert. Zur Identifikation und Extraktion des zentralen Webseitentexts wurde das Python-Paket *Trafilatura* verwendet (BARBARESI 2021). Trafilatura verwendet einen eigenen regelbasierten Extraktionsalgorithmus, der zentrale HTML-Elemente des Webseitencodes identifiziert und die dahinterliegenden Textelemente ausliest. Somit werden irrelevante und potentiell irreführende Textbausteine, wie beispielsweise Werbebanner, Kopf- und Fußzeilen oder Steuerelemente, herausgefiltert. Im Vergleich zu anderen Textextraktionsalgorithmen besticht Trafilatura neben einer zuverlässigen Klassifikationsgüte durch vielfach kürzere Verarbeitungslaufzeiten (BARBARESI 2021).

Da Trafilatura zwar sehr zuverlässig die zentralen Textelemente einer Webseite identifizieren, jedoch keine inhaltliche Bewertung des Texts vornehmen kann, wurde ein weiterer Filterschritt eingesetzt, um inhaltlich grundsätzlich relevante Webseiten extrahieren zu können. Hierzu wurde eine Liste mit Stichwörtern angelegt, welche grob das Technologiefeld KI umreißen. Da die Auswahl der Stichwörter ein wichtiger erster Filterschritt ist, wurde diese nicht subjektiv getroffen, sondern intersubjektiv nachvollziehbar aus vortrainierten Word-Embeddings abgeleitet. Wie bereits in Kapitel 4 erläutert sind Word-Embeddings in der Lage semantische Ähnlichkeiten von Wörtern zu quantifizieren. Trotz der geschilderten Nachteile statischer Word-Embeddings bei der Sprachmodellierung sind Word-Embeddings dennoch geeignet, um verwandte Wörter zu explorieren. Ausgehend von dem Begriff „artificial intelligence“ wurde in einem englischsprachigen Word-Embedding nach verwandten Begriffen gesucht. Das verwendete Word-Embedding wurde in FastText auf Textdaten aus Wikipedia und CommonCrawl vortrainiert (GRAVE et al. 2018).

Abbildung 32 visualisiert die semantische Nachbarschaft des Begriffs „artificial intelligence“. Hierbei fällt auf, dass vier benachbarte Begriffe direkt den Begriff „artificial intelligent“ bzw. die Abkürzung „ai“ enthalten. Weitere technische Begriffe, welche zur Identifizierung von KI-Unternehmen verwendet werden können, sind „deep learning“, „machine learning“, „computer-vision“ und „image-recognition“. Die weniger technischen sondern eher deskriptiven Begriffe „biologically-inspired“ oder „brain-inspired“ sind für die Identifizierung von Softwareunternehmen eher ungeeignet, da sie nicht exklusiv im KI-Kontext eingesetzt werden könnten.

Abbildung 32: Semantische Nachbarbegriffe des Begriffs "artificial intelligence".



Quelle: Eigene Darstellung.

Zur Generierung weiterer potentieller Suchbegriffe wurde ebenfalls die unmittelbare semantische Nachbarschaft der in Abbildung 32 dargestellten Wörter betrachtet. Durch dieses Vorgehen lässt sich schnell, aber dennoch reproduzierbar, ein semantischer Vektorraum absuchen und Begriffskandidaten definieren. Die finale Auswahl der Begriffe erfolgte durch eine manuelle Selektion. Diese Begriffe wurden anschließend in die deutsche Sprache übersetzt, um Webseiten sowohl in englischer als auch in deutscher Sprache durchsuchen zu können.

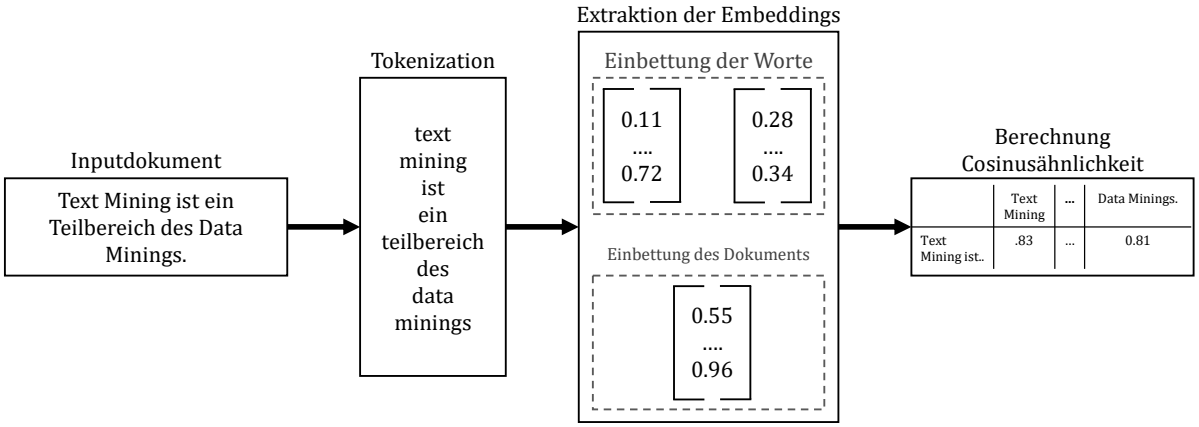
Wie Abbildung 32 zeigt, handelt es sich bei vielen extrahierten Begriffen um N-Gramme. Um die Webseitentexte systematisch nach den definierten Stichwörtern durchsuchen zu können, wurde der Webseitentext in ein standardisiertes Format überführt. Die Texte der einzelnen Webseiten wurden zunächst in Absätze aufgeteilt. Dies dient einerseits dazu, die Textmenge der einzelnen betrachteten Entitäten zu verringern. Standardmäßig arbeiten bidirektionale Transformermodelle mit einer maximalen Textlänge von 512 Token (DEVLIN et al. 2019). Längere Texte werden abgeschnitten und bei der Prozessierung nicht weiter berücksichtigt. Somit ermöglicht die Klassifikation auf Absatzebene die Berücksichtigung des gesamten vorliegenden Webseitentexts. Andererseits können auf mehreren Absätzen einer Webseite verschiedene Themen behandelt werden. Durch die Betrachtung der Texte auf Absatzebene ist ein stärkerer Sinnzusammenhang innerhalb einzelner Absätze zu erwarten, als dies bei Betrachtung ganzer Webseiten der Fall wäre. Entsprechend ist davon auszugehen, dass Absätze eine eindeutigere semantische Zuordnung zulassen, als vollständige Webseiten.

Im nächsten Schritt wurden mittels eines unüberwachten Stichwortextraktionsverfahrens (Keyword-Extraction) für jeden Absatz des Webseitenkorpus repräsentative Stichwörter extrahiert. Keyword-Extraction-Verfahren zielen darauf ab, aus einem Textdokument Stichwörter zu extrahieren, die alle wichtigen Inhalte des Textdokuments möglichst vollständig und inhaltlich kohärent darstellen. Somit können die Ergebnisse einer Keyword-Extraction als Kurzzusammenfassung des Ausgangsdokuments betrachtet werden (FIROOZEH et al. 2020).

Zur Keyword-Extraction wurde das Python-Paket *KeyBERT* verwendet (GROOTENDORST 2021). Abbildung 33 zeigt den Ablauf der Keyword-Extraction mit KeyBERT. Zunächst werden alle Wörter des Inputdokuments in Tokens überführt. Anschließend wandelt der Algorithmus sowohl die einzelnen Tokens als auch das gesamte Inputdokument in einen numerischen Vektor um. Zur Einbettung des Dokuments nutzt KeyBERT die Sentence-Transformersarchitektur (REIMERS und GUREVYCH 2019). Diese ist in der Lage ganze Textsequenzen in einen numerischen Vektor umzuformen. Im letzten Schritt wird die Cosinus-Ähnlichkeit zwischen den Dokument-Embeddings und den Word-Embeddings berechnet, um die Wörter/N-Gramme zu bestimmen, welche dem Inhalt des Dokuments am besten entsprechen (GROOTENDORST 2021).

Die identifizierten Stichwörter wurden anschließend durch Lemmatisierung in ihre Grundform transformiert (BALAKRISHNAN und LLOYD-YEMOH 2014). Zur Lemmatisierung deutscher und englischer Texte wurde die Programmbibliothek Spacy verwendet (MONTANI et al. 2022). Nach abschlossener Lemmatisierung können die den Webseitentexten assoziierten Stichwörter nach den ausgewählten Suchbegriffen zur Identifizierung von KI-Unternehmen durchsucht werden.

Abbildung 33: Funktionsweise der Keyword-Extraction.



Quelle: Eigene Darstellung nach GROOTENDORST (2021).

Tabelle 11 zeigt die finale Begriffsliste, die für die Stichwortsuche verwendet wurde. Neben der Keyword-Extraction und der Lemmatisierung wurden sämtliche Wörter in Kleinbuchstaben transformiert sowie Sonderzeichen und Zahlen entfernt.

Tabelle 11: Stichwortlisten zur Identifizierung von KI-Unternehmen.

Englischer Suchbegriff	**Deutscher Suchbegriff**
machine learn	maschinell lernen
deep learn	
computer vision	computervision
image recognition	bilderkennung
artificial intelligence (ai)	künstlich intelligenz (ki)
neural network	neuronal netz
face tracking	gesichtserkennung
face detection	
face recognition	
gesture recognition	gestenerkennung
natural language processing	
algorithm	algorithmus

Quelle: Eigene Darstellung.

Somit kann mit einem Stichwort (z.B. künstlich intelligenz) nach sämtlichen flektierten Formen dieses Stichwortes (z.B. künstlich intelligenter, künstlicher intelligenz, künstlich intelligente) gesucht werden. Mittels der Stichwortsuche konnte die Anzahl der zu betrachtenden Unternehmenswebseiten deutlich reduziert werden. So umfasste der Datensatz nach der Stichwortsuche 63.187 Unternehmenswebseiten von 26.915 Unternehmen.

Da die reine Existenz einer der Begriffe auf der Webseite eines Unternehmens noch keine Rückschlüsse auf den Einsatz von KI-Technologien im Unternehmen zulässt, wurde der Datensatz nochmals feiner klassifiziert. Um die Unternehmen zu ermitteln, die selbst KI-Technologien nutzen, bzw. Produkte oder Dienstleistungen mit KI-Bezug anbieten, wurde ein Modell zur überwachten Textklassifikation trainiert. Dazu wurden die im ersten Schritt gefundenen Texte in Absätze aufgeteilt. Anschließend wurden manuell 1.069 dieser Paragraphen annotiert.

Die Paragraphen wurden jeweils einer von zwei Kategorien zugeordnet. Kategorie eins umfasst Absätze, in denen Unternehmen berichten, dass sie KI-Produkte oder Dienstleistungen anbieten bzw. diese nutzen. Kategorie zwei sind Absätze, in denen zwar einer der Begriffe steht, jedoch kein klarer Bezug zwischen dem Unternehmen und dem Begriff besteht. Dies ist beispielsweise dann der Fall, wenn allgemein über KI-Technologie gesprochen wird oder Partner des Unternehmens KI einsetzen. Tabelle 12 zeigt jeweils zwei deutsche und zwei englische Absätze auf, die mittels der Stichwortsuche identifiziert und anschließend als Trainingsdaten verwendet wurden.

Tabelle 12: Auszüge aus den Trainingsdaten des Klassifikationsmodells.

Absatz	Label
Wir nutzen künstliche Intelligenz (KI) und maschinelles Lernen (ML) zur Optimierung digitaler Erlebnisse. Unter KI verstehen wir die Fähigkeit von Maschinen und Systemen, intelligentes menschliches Verhalten nachzuahmen. Unterscheiden muss man hier zwischen General KI, also vollständig maschinengestützten Geschäftsmodellen, und Functional KI, also einzelnen Entscheidungszweigen, bei denen die künstliche Intelligenz in Teilbereichen mitwirkt.	Direkter KI-Bezug
Your health data is yours. Own it, train it, grow it. Our deep learning computing technology is designed for vertical medical domains and to generate personalized health insights you can act on.	Direkter KI-Bezug
Künstliche Intelligenz und Machine Learning revolutionieren die Wirtschaftsprüfung. Die Entwicklung schreitet rasant voran.	Kein direkter KI-Bezug
The EU Commission plans to overturn all existing IT security measures. The plan: automatic scans of all content on the Internet and the use of “artificial intelligence” are to fight crime. The regulation was announced by EU Commissioner Ylva Johansson for the beginning of December. A group of academic experts called the goals of the EU Commission’s monitoring plans simply “illusory”.	Kein direkter KI-Bezug

Quelle: Eigene Darstellung.

Die ersten beiden Absätze beschreiben die direkte Nutzung von KI durch den Webseitenbetreiber, da in der ersten Person von der KI-Nutzung gesprochen wird.. Aus den anderen beiden Absätzen lassen sich keine direkten Verbindungen zwischen der Technologie und dem Unternehmen ablesen, da eher allgemein über die Technologie gesprochen wird.

Der manuell annotierte Datensatz wurde anschließend verwendet, um ein vortrainiertes Transformermodell mittels des sogenannten Transfer-Learnings abzustimmen. Wie bereits in Kapitel 4.8 erläutert, verfügen diese vortrainierten Transformermodelle bereits a priori über ein umfassendes Textverständnis, sodass rund 1.000 annotierte Datenpunkte ausreichten, um ein performantes Klassifikationsmodell zu trainieren. Konkret wurde für diese Fallstudie das XLM-RoBERTa Modell verwendet (HUGGING FACE 2022b). Dieses wurde auf Basis von über zwei Terrabyte Textdaten in über 100 Sprachen aus dem CC-Projekt vortrainiert (RUDER et al. 2019b).

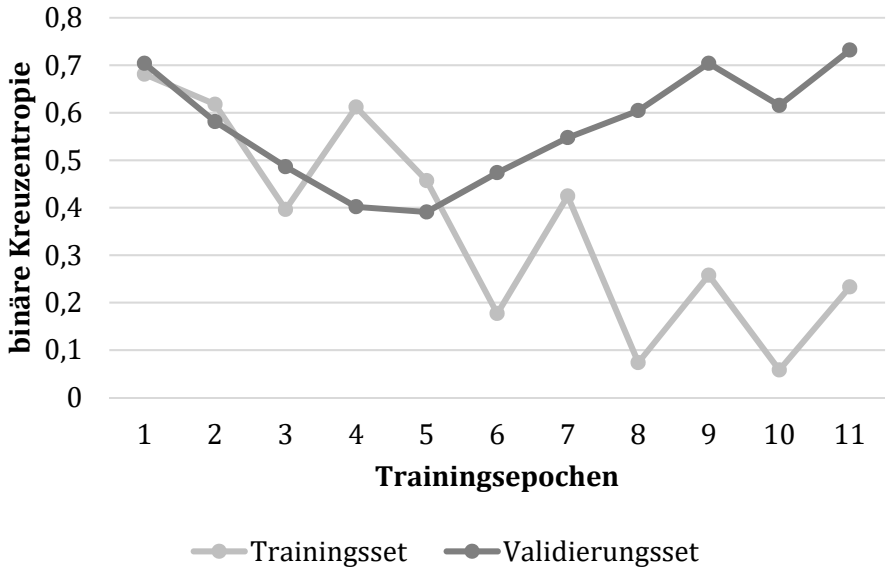
Für Training, Validierung und Evaluation wurde der Datensatz in drei Subsets aufgeteilt (65 % Training; 10 % Validierung; 25 % Evaluation). Diese Aufteilung ist ein gängiges Verfahren des ML (GÉRON 2022), wodurch nach jeder Epoche eines Trainingslaufs mittels des Validierungssets geprüft wird, ob das Modell allgemein übertragbare Fähigkeiten erlernt hat. Durch diese Validierung wird verhindert, dass das Modell sich zu stark an die Trainingsdaten anpasst und weiterhin allgemein übertragbare Fähigkeiten erlernt. Andererseits wird sichergestellt, dass das Training nicht vorzeitig – also während der Lernphase - abgebrochen wird. Dies geschieht durch Überwachung des Ergebnisses der Verlustfunktion (engl. loss).

Im Fall eines binären Klassifikationsproblems wird in der Regel die binäre Kreuzentropie als Verlustfunktion verwendet. Diese ist zwischen 0 und 1 normiert und nimmt ab, wenn sich die vorhergesagten Werte den tatsächlichen Werten annähern. Vereinfacht lässt sich mittels der Veränderung der binären Kreuzentropie also nach jeder Epoche der Lernerfolg der letzten Epoche quantifizieren (GOODFELLOW et al. 2016). Für das Training dieses Modells wurde die binäre Kreuzentropie des Validierungssets betrachtet, um die optimale Anzahl von Trainingsepochen zu bestimmen. Nimmt die binäre Kreuzentropie über fünf Epochen hinweg nicht weiter ab, wird das Training abgebrochen und die Modellparameter des besten Ergebnisses werden gespeichert.

Abbildung 34 zeigt den Verlauf der binären Kreuzentropie für das Trainings- und Validierungsset. Wie aus der Abbildung hervorgeht, nimmt die binäre Kreuzentropie des Validierungssets bis Epoche fünf kontinuierlich ab. Ab Epoche fünf nimmt das Maß wieder stetig zu. Somit ist anzunehmen, dass nach fünf Trainingsepochen die Lernerfolge auf dem Trainingsset nicht mehr auf das Validierungsset übertragbar sind. Das Phänomen, dass sich Modelle zu sehr an die Trainingsdaten anpassen ist unter dem Begriff des Overfittings gefasst (JURAFSKY und MARTIN 2019). Dabei passt sich das Model nach und nach perfekt an die Trainingsdaten an, allerdings sind die erlernten Fähigkeiten nicht mehr auf ungesehene Daten übertragbar.

Die finale Vorhersagegüte des Klassifikationsmodells wird abschließend anhand des Testsets evaluiert. Dieses wurde während des Trainings nicht verwendet, sodass mittels des Testsets geprüft werden kann, ob das Modell während des Trainings allgemeingültig übertragbare Fähigkeiten erlernt hat.

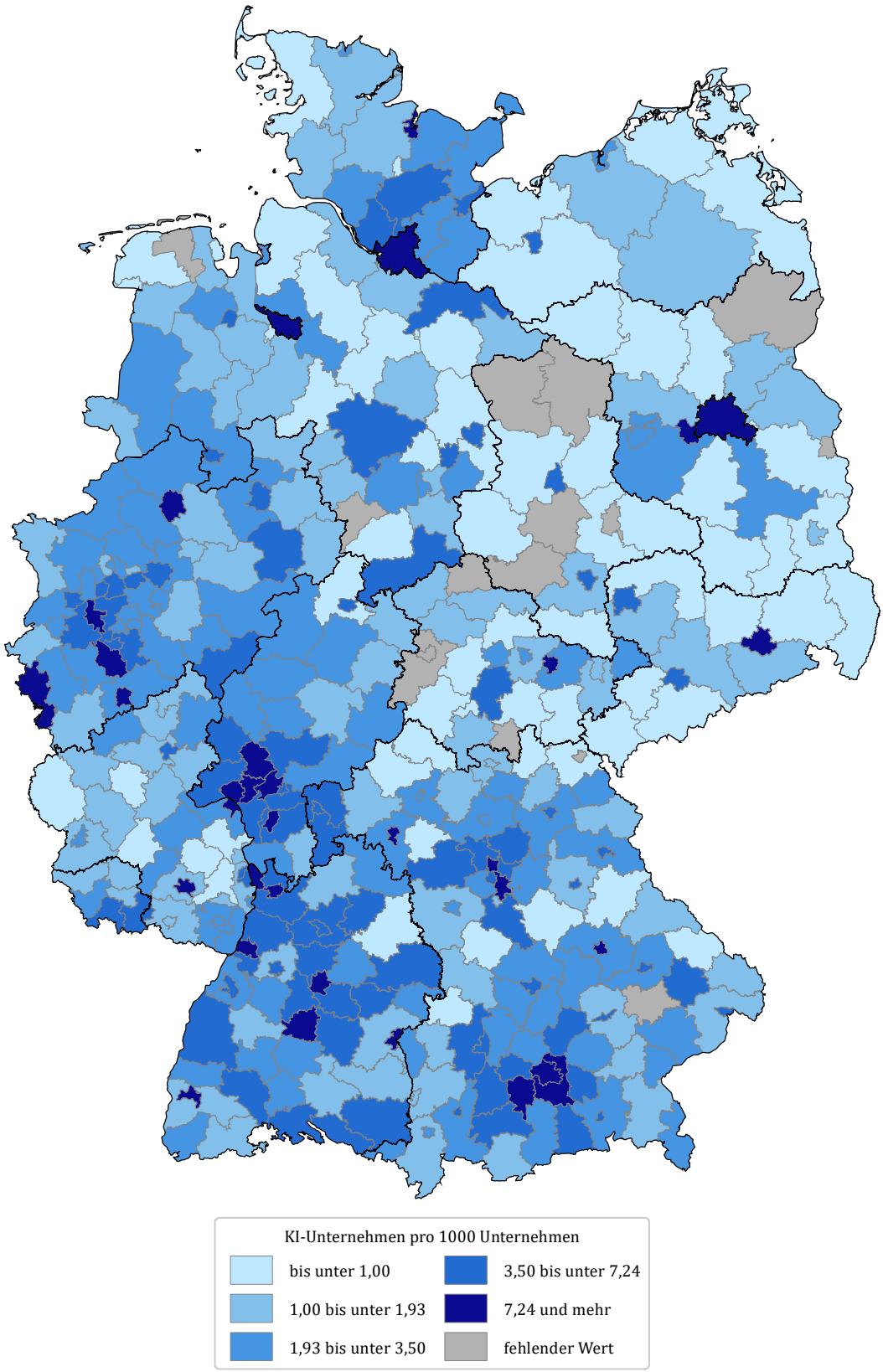
Abbildung 34: Lernraten des Textklassifikationsmodells.



Quelle: Eigene Darstellung.

Hierzu werden die vorhergesagten Kategorien des Modells den manuell annotierten Kategorien gegenübergestellt. Insgesamt erzielt das Modell auf dem Testset eine Vorhersagegenauigkeit von 85,04 %. Die Sensitivität, also die Wahrscheinlichkeit, mit der ein KI-Unternehmen auch als ein solches erkannt wird, beträgt 81,33 %. Die Spezifität, also die Wahrscheinlichkeit, dass ein Unternehmen, welches keine KI verwendet, auch als ein solches erkannt wird, liegt etwas höher bei 85,92 %. Anschließend wurden die einzelnen Absätze der Unternehmenswebseiten mit dem Klassifikationsmodell kategorisiert. Nach abgeschlossener Klassifikation blieben Webseiten von 10.453 Unternehmen im Datensatz erhalten. Im Vergleich mit den Ergebnissen der reinen Stichwortsuche zeigt sich, dass durch die zusätzliche Filterung mit dem Klassifikationsmodell 63,9 % der initial detektierten Unternehmen entfernt wurden. Eine reine Stichwortsuche hätte somit aufgrund der Missachtung des Kontexts enorm viele falsch-positive Fälle erzeugt. Abbildung 35 zeigt die geographische Verteilung der identifizierten KI-Unternehmen aggregiert auf Kreisebene.

Abbildung 35: Verteilung der identifizierten KI-Unternehmen.



Quelle: Eigene Darstellung.

Bei Betrachtung von Abbildung 35 fällt ein markanter Unterschied zwischen Kreisen der neuen Bundesländer und ehemaligen BRD-Kreisen auf. Während der Anteil von KI-Unternehmen pro 1.000 Unternehmen im Osten Deutschlands größtenteils unter dem Mittelwert der Verteilung von 3,19 liegt, nehmen die Kreise im Westen der Republik häufig Werte zwischen 1,93 und 7,24 an.

Besonders hohe Werte sind vor allem in kreisfreien Städten zu beobachten. So sind unter den zehn Kreisen mit der höchsten KI-Dichte neun kreisfreie Städte zu finden. Die höchste KI-Dichte weist die kreisfreie Stadt Heidelberg mit 19,81 KI-Unternehmen pro 1.000 Unternehmen auf. Es folgen die kreisfreie Stadt München (18,86), der Landkreis München (16,93), die kreisfreie Stadt Karlsruhe (16,11), die kreisfreie Stadt Darmstadt (16,04) sowie die kreisfreie Stadt Potsdam (13,48).

7.3 Analyse der Standortfaktoren

Zur präziseren Erklärung der Standortmuster wurde ein multiples, lineares Regressionsmodell aufgestellt. Dazu wurden die identifizierten KI-Unternehmen auf Kreisebene aggregiert und mit der Gesamtzahl der im Kreis niedergelassenen Betriebe normalisiert. Zur Erklärung der so erzeugten Variable *Anteil KI-Unternehmen an allen Unternehmen pro Kreis (KI-Dichte)* wurden Faktoren betrachtet, welche der im ersten Teil dieser Fallstudie vorgestellten Theorien entnommen wurden.

Das Hauptaugenmerk der Untersuchung liegt auf der Prüfung der zu Beginn des Kapitels erläuterten Agglomerationsvorteile. Im Kontext von KI-Unternehmen sind dies neben der Unternehmens- und Einwohner:innendichte vor allem die Verfügbarkeit von technisch-affinem Personal mit Hochschulabschluss. Zur Anwendung und Entwicklung von KI benötigt dieses Personal verschiedene Kompetenzen, wie Auswertungen von Stellenbeschreibungen zeigen (ZSCHECH et al. 2018; BENSBERG und BUSCHER 2016; KIM und CHOONG 2016). So wird ein profundes Verständnis für Daten und deren Verarbeitung sowie Programmierkenntnisse, Kenntnisse im Umgang mit Software und Konzeptwissen in Mathematik und Statistik verlangt. Darüber hinaus spielen personenbezogene Kompetenzen wie eine ausgeprägte Kommunikations- und Kollaborationsfähigkeit sowie Neugier und Kreativität eine zentrale Rolle (ZSCHECH et al. 2018).

Da diese Kompetenzen insbesondere in Studiengängen der Mathematik, Informatik, Naturwissenschaften und Technik (MINT) vermittelt werden, wurde die Anzahl dieser Studiengänge pro 1.000 Einwohner:innen als erklärende Variable in das Regressionsmodell aufgenommen. Diese stellen zwar latent die Forschungsschwerpunkte der jeweiligen Hochschulen dar. Dennoch wurde die Anzahl außeruniversitärer Forschungseinrichtungen der Wissenschaftszweige Naturwissenschaften, Ingenieurwissenschaften und Technologie pro 1.000.000 Einwohner:innen ergänzend in das Modell integriert. Die Studiengangsdaten stammen von der Bundesagentur für Arbeit

(BUNDESAGENTUR FÜR ARBEIT 2022). Diese können effizient über eine Programmierschnittstelle abgerufen werden. Die Daten zu außeruniversitären Forschungseinrichtungen stammen von dem BMBF (BMBF 2020).

Zur Vervollständigung des regionalen Humankapitalbestands wurde darüber hinaus der Akademiker:innenanteil inkludiert. Ferner spielen Faktoren der Lebensqualität eine wichtige Rolle, um hochqualifizierte Arbeitskräfte einerseits am Standort halten zu können und andererseits weitere Talente anzuziehen. Um diese Faktoren ebenfalls im vorliegenden Modell zu berücksichtigen, wurde die Anzahl der Straftaten pro 1.000 Einwohner:innen, die Erholungsfläche pro Einwohner:innen sowie die Baulandpreise pro m^2 beleuchtet. Zur Kontrolle für Effekte bisher unberücksichtigter sozioökonomischer Faktoren wurde das Bruttoinlandsprodukt (BIP) pro Kopf in das Modell aufgenommen.

Neben einem ausgeprägten und spezifischen Humankapitalbestand sind nach PORTER (1990) Agglomerationsfaktoren zentrale Aspekte zur Erklärung von Technologieclustern. Zur Kontrolle dieser Dimension enthält das Regressionsmodell neben der Einwohner:innen- und Unternehmensdichte, den lokalen Gewerbesteuerhebesatz und den Anteil der Großunternehmen am lokalen Unternehmensbestand. Letztere dienen gleichermaßen, dazu die in der Literatur diskutierte Relevanz der Ankerunternehmen zu prüfen (FELDMAN 2003; AGRAWAL und COCKBURN 2003).

Während für industrielle Cluster eine leistungsstarke Infrastruktur vor allem zur Beförderung von Waren und Gütern einen erfolgskritischen Faktor darstellt, nimmt für IT-Unternehmen eher die Personenmobilität einen relevanten Stellenwert ein. Darüber hinaus ist anzunehmen, dass die Qualität der digitalen Infrastruktur eine wichtige Komponente für die Entstehung von KI-Clustern darstellt. Somit bildet die Qualität der physischen und digitalen Infrastruktur einen weiteren Untersuchungsschwerpunkt im vorliegenden Regressionsmodell.

Wie Abbildung 33 zu entnehmen ist, weisen besonders die Kreise in den neuen Bundesländern niedrige KI-Dichten auf. Um diese visuelle Annahme prüfen zu können, wurde eine Dummy-Variable in das Modell integriert. Die Dummy-Variable bestimmt, ob ein Kreis in der ehemaligen DDR oder BRD liegt. Tabelle 13 gibt einen Überblick über die Ergebnisse des Regressionsmodells. Sämtliche Variablen des Modells wurden im Vorfeld der Parameterschätzung logarithmiert, sodass ein Log-Log-Modell vorliegt. Nach Bereinigung fehlender Werte betrachtet das vorliegende Regressionsmodel insgesamt 384 Kreise. In Klammern hinter den Beta-Koeffizienten sind jeweils die Standardfehler der Variable aufgetragen.

Tabelle 13: Ergebnisse des OLS-Modells.

Variablenname	Beschreibung	Beta-Koeffizient
	Agglomerationsfaktoren	
Unternehmensdichte	Unternehmen pro 1.000 Einwohner:innen	0,3758*** (0,140)
Einwohner:innendichte	Einwohner:innen pro km^2	0,0933** (0,037)
Anteil Großunternehmen	Anteil der Niederlassungen mit mehr als 250 SV-Beschäftigten an den Niederlassungen insgesamt in %	0,5585** (0,238)
Gewerbesteuer	Gewerbesteuer in € je Einwohner:in	0,0477 (0,062)
	Sozioökonomische Faktoren	
BIP pro Kopf	Bruttoinlandsprodukt je Einwohner:in	-0,14443 (0,108)
Anteil Akademiker:innen	Anteil der SV Beschäftigten am Arbeitsort mit akademischem Berufsabschluss an den SV Beschäftigten in %	0,4971*** (0,094)
MINT-Studiengänge	MINT-Studiengänge pro 1.000 Einwohner:innen	0,3931** (0,182)
Baulandpreise	Durchschn. Kaufwerte für Bauland in € je m^2	0,1007*** (0,034)
FuE-Einrichtungen	Außeruniversitäre FuE-Einrichtungen der Wissenschaftszweige Naturwissenschaften und Ingenieurwissenschaften und Technologie pro 1.000.000 Einwohner:innen	0,1271*** (0,020)
	Infrastruktur	
Autobahnanbindung	Durchschn. Pkw-Fahrzeit zur nächsten BAB-Anschlussstelle in Minuten	-0,0437 (0,038)
Breitbandanbindung	Anteil der Haushalte mit Breitbandversorgung mit 100 Mbit/s in %	0,1958 (0,135)
Fernverkehranbindung	Durchschn. Pkw-Fahrzeit zum nächsten Fernverkehrsbahnhof in Minuten	0,0233 (0,020)
Erreichbarkeit Flughafen	Durchschn. Pkw-Fahrzeit zum nächsten internationalen Flughafen in Minuten	0,0082 (0,042)
	Lebensqualität	
Lebenserwartung	Durchschnittliche Lebenserwartung eines Neugeborenen	6,7147*** (1,967)
Straftaten	Straftaten pro 1.000 Einwohner:innen	-0,0601 (0,065)
Erholungsflächen	Erholungsflächen pro Einwohner:innen	-0,0451 (0,044)
Dummy-Variable West/Ost		0,1712*** (0,061)

***p < 0,01; **p<0,05; *p<0,1; adj. R^2: 0,768

Quelle: Eigene Berechnung.

Insgesamt ist das Modell in der Lage mehr als drei Viertel der Varianz der abhängigen Variable *KI-Dichte* zu erklären. Dieser Erklärungsgehalt entspricht damit nach COHEN (1992) einem starken Effekt.

Tabelle 14 zeigt, dass besonders sozioökonomische- und Agglomerationsfaktoren einen signifikanten Effekt auf die *KI-Dichte* haben, während die Infrastrukturvariablen keine signifikanten Werte annehmen. Dem Modell zufolge siedeln sich KI-Unternehmen verstärkt in Agglomerationsräumen mit höherer Einwohner:innen- und Unternehmensdichte an. Des Weiteren spielt die räumliche Nähe zu Großunternehmen eine bedeutende Rolle. Auf Seiten der sozioökonomischen Faktoren sind vor allem die Bildungsindikatoren von Relevanz. Ein hoher Akademiker:innenanteil sowie viele MINT-Studiengänge und fachspezifische außeruniversitäre Forschungseinrichtungen haben positive Auswirkungen auf die KI-Dichte einer Region. Darüber hinaus spielt die Attraktivität der Region eine wichtige Rolle für die Erklärung der regionalen KI-Dichte. So sind hohe Baulandpreise und hohe Lebenserwartungen ebenfalls mit einer hohen KI-Dichte assoziiert. Letztlich bestätigt das Regressionsmodell die visuelle Annahme, dass Kreise in den neuen Bundesländern niedrigere KI-Dichten aufweisen, als Kreise in den alten Bundesländern. Das Modell schätzt die KI-Dichte in westdeutschen Kreisen um 18,6 % höher ein als in ostdeutschen Kreisen.

7.3.1 Interpretation und Diskussion der Analyseergebnisse

Zusammenfassend lässt sich festhalten, dass starke regionale Innovationssysteme auch für die Entwicklung und Anwendungen von KI-Technologien eine große Bedeutung haben. So spielen akademische Forschung und Lehre als Wissensproduzenten eine zentrale Rolle bei der Standortwahl von KI-Unternehmen. Die signifikant positiven Effekte von Studiengängen und Forschungseinrichtungen der MINT-Disziplinen sowie der Akademiker:innenanteil sind gut mit etablierten regionalen Innovationstheorien erklärbar. Einerseits sorgt ein hoher Anteil von MINT-Studiengängen für fachlich passende und hochausgebildete Arbeitskräfte, welche direkt im jeweiligen Kreis eine Arbeitsstelle finden (BRESCHI 2001; BRAMWELL und WOLFE 2008). Andererseits erreicht Forschung an Universitäten und außeruniversitären Einrichtungen, dass globales, fachspezifisches Wissen in die Region gelangt (FROMHOLD-EISEBITH und WERKER 2013; BENNEWORTH et al. 2012; BENNEWORTH und HOSPERS 2007; ETZKOWITZ und LEYDESDORFF 1997). Insbesondere für die Entwicklung und Anwendung von KI nimmt aktuelle internationale Forschung einen hohen Stellenwert ein. Die Entwicklungsmotoren von KI sind zumeist große amerikanische Technologieunternehmen, die leistungsstarke Frameworks und Programme zur Entwicklung von KI bereitstellen (RASCHKA et al. 2020). Das Forschungsgeschehen in diesem Feld ist sehr dynamisch, sodass Forschungs- und Entwicklungseinrichtungen für die regionale Inwertsetzung dieser Technologien eine wichtige Rolle einnehmen.

Neben der Relevanz von Universitäten und Forschungseinrichtungen als Ausbildungsstätten und Wissensgeneratoren ist der lokale Firmenbesatz für die Erklärung der regionalen KI-Dichte von Relevanz. So siedeln sich KI-Unternehmen verstärkt in urbanen Gebieten mit hoher Firmendichte

und hohem Anteil an Großunternehmen an. Letztere können für dienstleistungsorientierte KI-Unternehmen wichtige Kunden darstellen. Insbesondere das Zusammenspiel von KI-Entwickler:innen und KI-Anwender:innen können somit eine wichtige Grundlage für die Entstehung von KI-Clustern bilden.

Die Qualität der Infrastruktur steht insgesamt in keinem signifikanten Zusammenhang mit der KI-Dichte. Für reine Softwareunternehmen lässt sich der nicht vorhandene Zusammenhang logisch erklären, da diese kaum auf den Transport von Waren und Gütern angewiesen sind. Allerdings werden mit der Variable KI-Dichte theoretisch auch beispielsweise Industrie- oder Agrarunternehmen betrachtet, welche KI aktiv nutzen. Die sektoralen Charakteristika der betrachteten Unternehmen können aufgrund mangelnder Granularität sekundärstatistischer Daten im Rahmen dieser Arbeit nicht genauer beleuchtet werden. Darüber hinaus hat auch die Breitbandverfügbarkeit keinen signifikanten Einfluss auf die KI-Dichte. Erklären lässt sich dies lediglich mit der räumlichen Maßstabsebene. Die verwendete Kreisebene ist für die Betrachtung von Breitbandverfügbarkeiten vermutlich zu grob, da die Qualität der digitalen Infrastruktur eines gesamten Landkreises nicht zwangsläufig mit der eines Unternehmens korreliert.

Bei Betrachtung der Faktoren zur Lebensqualität weist lediglich die Variable Lebenserwartung eines Neugeborenen einen signifikanten Effekt auf die KI-Dichte eines Kreises auf. Die Lebenserwartung ist von vielen Faktoren abhängig. Einerseits spielt der persönliche Lebensstil sowie der Zugang zu Ärzt:innen eine bedeutende Rolle. Andererseits zeigen Studien deutliche Zusammenhänge zwischen sozialem Status der Eltern und der Lebenserwartung eines Neugeborenen (LAMPERT et al. 2007; MACKENBACH 2006). Die Variablen Straftaten pro 1.000 Einwohner:innen und Erholungsflächen pro Einwohner:innen weisen keine signifikanten Zusammenhänge mit der Zielvariable auf. Dies ist primär mit dem Zusammenhang der KI-Dichte und der Variable Einwohner:innendichte zu erklären. Dichter besiedelte Kreise bieten aufgrund des Siedlungsdrucks weniger Erholungsflächen und leiden gleichzeitig unter erhöhten Straftaten im Vergleich zu dünner besiedelten Regionen.

Darüber hinaus wurde das vorliegende auf autoregressive räumliche Prozesse geprüft. Die Morans I Statistik ($p = 0.0871$) zeigt, dass die räumliche Autokorrelation zwischen den betrachteten Kreisen nur auf einem Signifikanzniveau von 10 % vorliegt. Folglich ist nicht von signifikanten autoregressiven Prozessen zwischen den betrachteten Kreisen auszugehen. Die KI-Dichte der einzelnen Kreise ist also nicht signifikant von benachbarten Kreisen beeinflusst, sondern lässt sich durch endogene Prozesse erklären. Dieses Ergebnis steht in Einklang mit den theoretischen Überlegungen der Clustertheorie. Die dort beschriebenen Spillovereffekte zielen insbesondere auf persönliche Kontakte ab, die leichter und intensiver innerhalb einer Region entstehen können, als über Kreisgrenzen hinweg. Da sich die KI-Technologie insbesondere in Deutschland noch in einem

Entwicklungsstadium befindet, siedeln sich entsprechende Unternehmen besonders häufig in Regionen mit entsprechender Humankapitalausstattung an. Diese wird einerseits durch Hochschulen und Forschungseinrichtungen sichergestellt. Andererseits sorgt die Nähe zu großen Unternehmen für die Möglichkeit, KI-Anwendungen zu entwickeln und einzusetzen. Zur empirischen Validierung dieser Annahme werden im folgenden Kapitel die Ansiedlungsmuster der identifizierten KI-Unternehmen auf Mikroebene betrachtet.

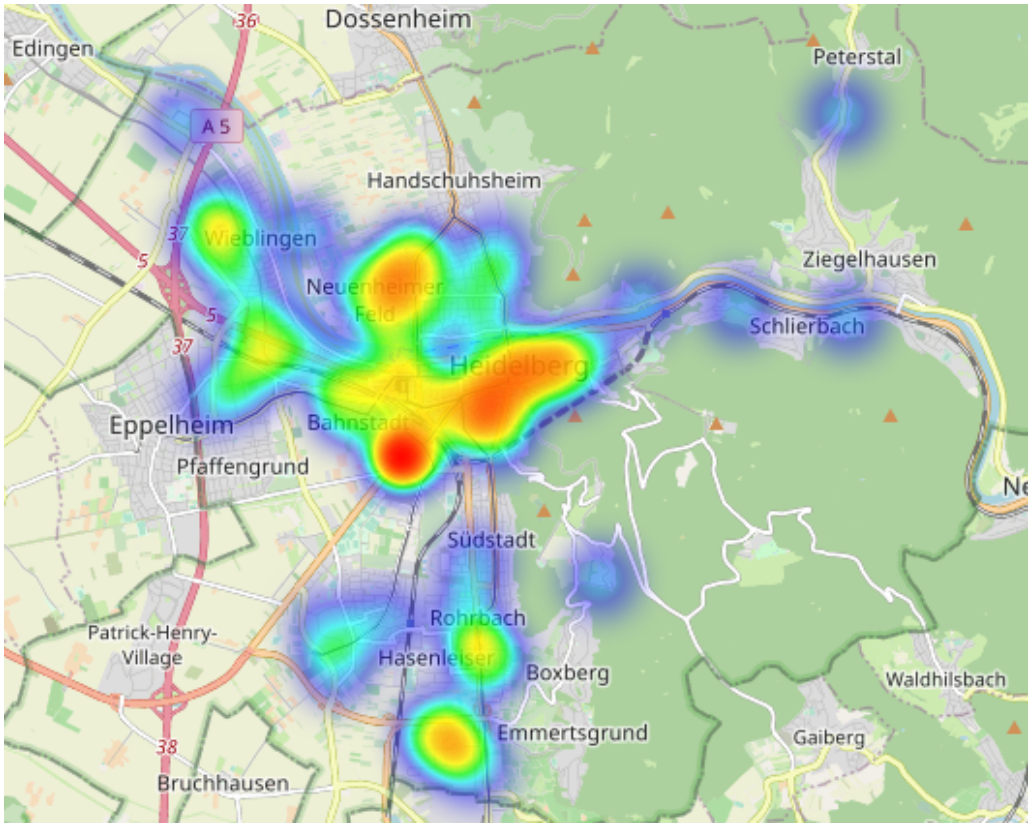
7.3.2 Mikrobetrachtung ausgewählter Kreisstädte

Nachdem in den vorherigen Kapiteln die geographische Verteilung visualisiert und die Standortbedingungen deutscher KI-Unternehmen analysiert wurden, können in einem weiteren Schritt dank der fein aufgelösten Datengrundlage kleinsträumige Betrachtungen durchgeführt werden. Diese können speziell für wirtschaftsgeographische Fragestellungen einen erheblichen Mehrwert bieten, da sie tiefgreifende Einblicke in regionale Innovationssysteme ermöglichen, die über die Körnung gängiger Aggregationsniveaus hinausgeht. Entsprechend soll im folgenden Kapitel exemplarisch dargestellt werden, inwiefern eine mikrogeographische Betrachtung klassische ökonometrische Modelle ergänzen und zu einem zusätzlichen Evidenzgewinn beitragen können.

Hierzu werden die mikrogeographischen Verteilungsmuster der beiden Städte mit den höchsten KI-Dichten betrachtet: Heidelberg und München. In Heidelberg wurden insgesamt 92 KI-Unternehmen identifiziert (19,81 pro 1.000 Unternehmen), während in München 893 Unternehmen (18,86 pro 1.000 Unternehmen) mit unmittelbarem KI-Bezug erkannt wurden. Abbildung 36 zeigt eine Heatmap der KI Unternehmen in Heidelberg.

Auffallend ist eine Ballung von KI-Unternehmen im Stadtzentrum. Im Osten der Stadt nimmt die Intensität im Vergleich zum Zentrum deutlich ab. Im Norden des Untersuchungsgebiets zeigt sich eine weitere Häufung von KI-Unternehmen im Neuenheimer Feld des Stadtteils Neuenheim. Das Neuenheimer Feld gilt als Wissenschaftscampus sowie Sondergebiet für Wissenschaft, Medizin und Krankenversorgung und wird zukünftig im Rahmen eines Masterplanverfahrens weiter entwickelt (STADT HEIDELBERG 2022b). Allerdings ist das Neuenheimer Feld auch heute schon ein bedeutender Wissenschafts- und Forschungstandort. 2021 flossen rund 245 Millionen Euro Drittmittel in die universitäre Forschung, die am Standort über 4.000 Doktorand:innen beschäftigt. Derzeit sind 105 Unternehmen im Neuenheimer Feld sowie ein 40.000 m^2 großer Technologiepark angesiedelt (STADT HEIDELBERG 2022c). Während im Technologiepark hauptsächlich Start-Ups vertreten sind, beherbergt das Neuenheimer Feld auch multinationale Unternehmen wie die Springer-Verlag GmbH oder die Octapharma Biopharmaceuticals GmbH (STADT HEIDELBERG 2022a).

Abbildung 36: Heatmap der KI-Dichte in Heidelberg.



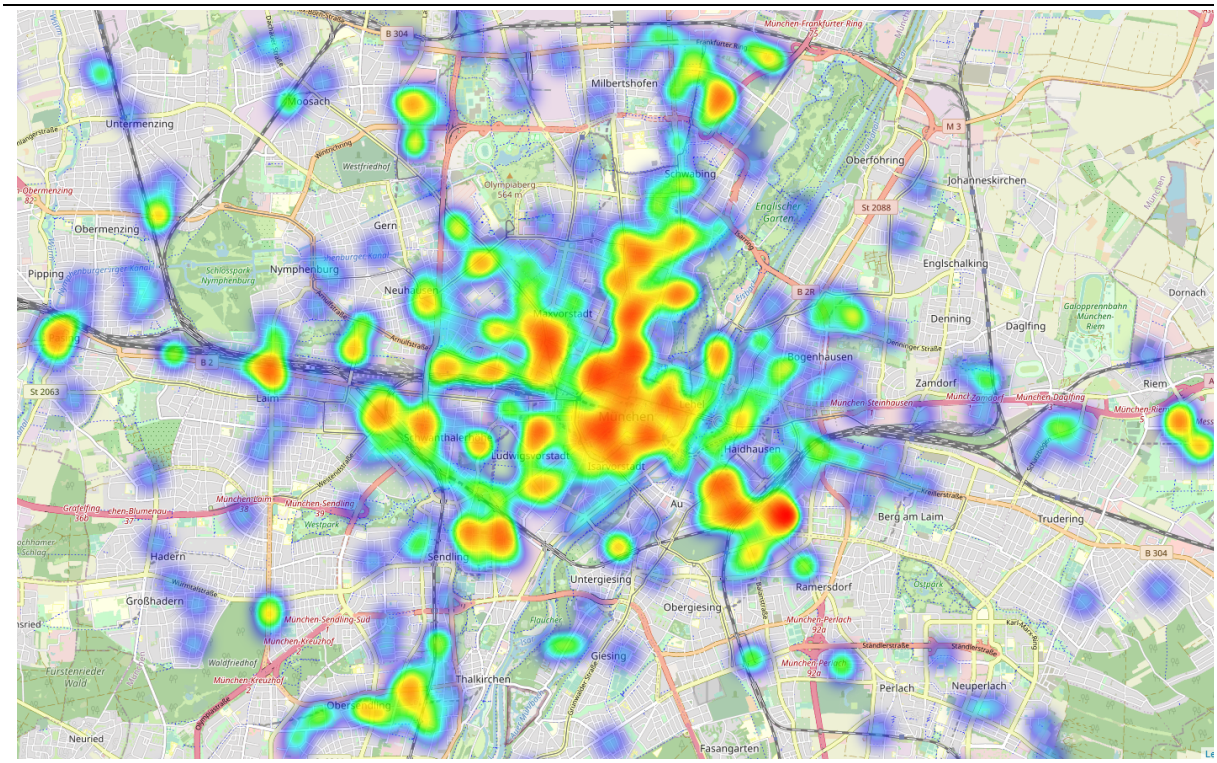
Quelle: Eigene Darstellung.

Ein weiterer KI-Hotspot in Heidelberg ist der Stadtteil Bahnstadt. Der Stadtteil ist der jüngste Heidelbergs. Seit 2012 vereinen sich auf dem ehemaligen Güterbahnhofareal die Nutzungen Wohnen, Forschen und Arbeiten. Auf dem 22 Hektar großen Gebiet sind ebenfalls zwei Dependancen des Heidelberger Technologieparks angesiedelt. Allein am Standort *Innovation Lab* an der Speyerer Straße sind über 190 Forschende beschäftigt. Mit dem *Forum organische Elektronik* hat ebenso ein internationales Spitzencluster seinen Sitz in der Bahnstadt wie andere forschungsnahe Unternehmen der Biotechnologie und IKT (STADT HEIDELBERG 2018).

Abbildung 37 zeigt die Standortmuster der ansässigen KI-Unternehmen für die kreisfreie Stadt München, welche mit 18,86 KI-Unternehmen pro 1.000 Unternehmen die zweithöchste KI-Dichte bundesweit aufweist. Insgesamt wurden in München 893 KI-Unternehmen identifiziert. Hier ballen sich die ansässigen KI-Unternehmen ebenfalls stark im Stadtzentrum. Insgesamt ist dennoch ein diffuseres Bild hinsichtlich der Verteilungsmuster als in Heidelberg zu sehen. Ausgehend vom Altstadtring sind KI-Unternehmen primär in den benachbarten Bezirken Maxvorstadt, Schwabing, Ludwigsvorstadt, Schwanthalerhöhe und Isarvorstadt zu finden. Südöstlich des Stadtzentrums tritt auf dem neu entwickelten Werksviertel am Rand des Ostbahnhofs eine weitere Ballung auf

(HELLER & PARTNER 2022). Am östlichen Rand haben weitere KI-Unternehmen ihren Standort direkt am Messegelände. Im südlich gelegenen Bezirk Obersendling ist der *Büropark Omega* sowie der *Sirius Business Park München* auf dem ehemaligen Siemensgelände Heimat weiterer KI-Unternehmen.

Abbildung 37: Heatmap der KI-Dichte in München.



Quelle: Eigene Darstellung.

Auf Basis dieser Analyse lassen sich politische Handlungsempfehlungen formulieren. Die Ergebnisse des Regressionsmodells untermauern die Relevanz von Bildung und Forschung für die Integration von KI-Technologien in den Geschäftsbetrieb von Unternehmen. Neben der Förderung eines hohen Anteils an Akademiker:innen gilt es insbesondere MINT-Fächer an Universitäten und Hochschulen zu fördern, um spezialisiertes Fachpersonal für die Entwicklung und Integration von KI-Technologien in den lokal ansässigen Unternehmen auszubilden. Darüber hinaus sollte ein allgemeines Verständnis von KI-Technologien und deren Implikationen für das jeweilige Fach in sämtlichen Studiengängen vermittelt werden. Des Weiteren gilt es KI-Forschung in Universitäten und Forschungseinrichtungen ebenfalls voranzutreiben. Insbesondere in den Bereichen Bildung und Forschung bestehen seitens der Politik beispielsweise durch gezielte Fördermaßnahmen vielseitige Eingriffsmöglichkeiten, um die Entwicklung von KI-Kompetenzen regional steuern zu können. Am Beispiel der Stadt Heidelberg konnte darüber hinaus exemplarisch gezeigt werden, dass Technologieparks sowie Netzwerke zwischen Wissenschaft und Wirtschaft ebenfalls Nährboden

für eine stärkere regionale KI-Kompetenz sein können. Hier können auch Maßnahmen der Stadtentwicklung und –planung Grundlage für einen einfacheren Austausch zwischen Akteuren des regionalen Innovationssystems sein.

Abschließend lässt sich festhalten, dass mittels Web Mining und NLP Indikatoren erzeugt werden können, die sich passend in quantitative Forschungsdesigns integrieren lassen. Aufgrund der breiten Datengrundlage lassen sich somit auch sehr spezifische Indikatoren aus den Textdaten ableiten. Dies stellt zweifelsohne einen Mehrwert für quantitativ-orientierte Forschung in der Wirtschaftsgeographie dar, da solch spezifische Indikatoren kaum aus etablierten Datenquellen (z.B. Sekundärstatistik, Patentdaten, Publikationen) extrahierbar sind. Ein weiterer Vorteil stellt die räumliche Granularität dar. Die Erzeugung von Punktdaten ermöglicht tiefgreifendere mikrogeographische Analysen, die insbesondere für wirtschaftsgeographische Untersuchungen von zentraler Bedeutung sind.

8 Fallstudie 3: Dynamisches Topic Modeling wirtschaftsgeographischer Literatur

Ein weiterer Anwendungsbereich von NLP ist die unüberwachte Analyse von großen Textmengen. Fragestellungsabhängig stehen unterschiedliche Verfahren, wie z.B. das Training von Word-Embeddings und Sprachmodellen (vgl. Kapitel 4), Clusterverfahren oder Topic Modeling, zur Verfügung. Topic Modelle werden genutzt, um latente Themenzusammenhänge mittels unüberwachtem ML aufzudecken. Die Anwendungsgebiete von Topic Modellen sind vielschichtig und reichen von der Themenmodellierung historischer Zeitungsartikel (YANG et al. 2011) über die textbasierte Klassifikation von Patenten (YUN und GEUM 2020), die Analyse von Nachrichtenmeldungen während der Coronapandemie (KIM 2020), der Dimensionsreduktion biologischer Datensätze (LIU et al. 2016b), der Diffusion von Technologieinnovationen (LENZ und WINKER 2020) bis hin zur Analyse von Programmcode (PANICHELLA et al. 2013). Weitere Anwendungsgebiete setzen observierte Themenentwicklungen mit gesamtwirtschaftlichen Entwicklungen in Zusammenhang (LARSEN und THORSRUD 2019; MIZUNO et al. 2017)

8.1 Problemstellung und Hintergrund

Induziert durch den rapiden Anstieg digitaler Texte steigt in der jüngeren Vergangenheit die Nachfrage nach automatisierten Verfahren der Textverarbeitung, um die Informationsdichte unstrukturierter Texte quantitativ fassen zu können. Die automatisierte Strukturierung großer Textmengen ist insbesondere für die Bibliothekswissenschaften von Relevanz. Wissenschaftliche Literatur wird primär digital publiziert und stellt eine umfassende Wissensressource dar (GRIFFITHS und STEYVERS 2004; GLENISSON et al. 2005). Topic Modelle bieten der Bibliometrie und Szientometrie erstmals die Möglichkeit die reichhaltigen textuellen Informationen von wissenschaftlichen Artikeln quantifizieren zu können. Beispielsweise untersuchen BLEI und LAFFERTY (2007) über 16.000 Artikel der Zeitschrift *Science* und leiten aus den Textdaten ein Themennetzwerk ab. Ein ähnliches Vorgehen wählen GLENISSON et al. (2005) in ihrer Studie, welche die Themen eines Special Issues der Zeitschrift *Scientometrics* modelliert. Neben der Themenmodellierung einzelner Zeitschriften haben ein Großteil der bestehenden Arbeiten, die wissenschaftliche Publikationen mit Topic Modellen untersuchen, einen klaren inhaltlichen Fokus. Beispielsweise existieren Themenmodellierungen zu Literatur in der Biomedizin (KAVVADIAS et al. 2020), der Wasserkraftforschung (JIANG et al. 2016) oder der Fertigungsforschung (XIONG et al. 2019). Untersuchungen zu wirtschaftsgeographischen Themenzusammenhängen und –veränderungen bestehen bis dato nicht, sodass dieses Kapitel die Tauglichkeit von Topic Modellen zur Exploration einschlägiger wirtschaftsgeographischer Literatur analysiert werden soll.

Die Fähigkeit von Topic Modellen, abstrakte Themen in Dokumentensammlungen zu modellieren und deren semantische Verwandtschaft zu quantifizieren, wird über die klassischen Anwendungsbiete hinausgehend zudem zur Dokumentenzuordnung verwendet. Bedeutende Verwendungszwecke sind beispielsweise die Zuordnung von Patentdokumenten und wissenschaftlicher Literatur (HAIN et al. 2022; RANAEI et al. 2017).

Auf methodischer Seite leisteten insbesondere die Modellarten LDA (BLEI et al. 2003) und Non-Negative Matrix Factorization (FÉVOTTE und IDIER 2011) Pionierarbeit. Entsprechend nutzen viele der empirischen Studien LDA-Modelle, um Themen aus Textkorpora zu extrahieren. Das LDA-Modell greift zur numerischen Einbettung von Wörtern auf den BOW-Algorithmus zurück (vgl. Kapitel 4.1). Eine zentrale Limitation dieser Modelle ist, dass sie nicht in der Lage sind kontextabhängige semantische Beziehungen zwischen den einzelnen Wörtern zu modellieren.

Daher nutzen moderne Verfahren zur Themenmodellierung die kontextuelle und semantische Leistungsstärke von Transformermodellen, um Text als hochdimensionale Wortvektoren einzubetten. Ein hochmodernes und leistungsstarkes Modell zur Themenmodellierung mit Transformermodellen ist das BERTopic-Modell (GROOTENDORST 2022). Der Algorithmus nutzt dabei ein sogenanntes Sentence-Transformermodell, welches in der Lage ist, ganze Sätze bzw. Absätze in Vektorrepräsentationen zu überführen (REIMERS und GUREVYCH 2019). Diese Fähigkeit ist für die Themenmodellierung von Dokumenten besonders interessant, da der gesamte semantische Inhalt einzelner Sätze und Absätze relevant ist, um die Themen eines Dokuments vollständig abbilden zu können.

8.2 Methodische Vorgehensweise

Mittels dynamischem Topic Modeling lassen sich Themen erkennen, deren Entwicklung im Laufe der Zeit betrachten und verschiedene Akzentuierungen der jeweiligen Themen im Zeitverlauf darstellen. Der Untersuchungsgegenstand der Fallstudie ist dabei die wirtschaftsgeographische, wissenschaftliche Literatur seit dem Jahr 1990. Um die relevantesten Zeitschriften für die Wirtschaftsgeographie zu ermitteln, wurde Google Scholar genutzt. Google Scholar weist registrierten Autor:innen verschiedene Labels zu, welche die jeweiligen Forschungsthemen der Autor:innen beschreiben. Zunächst wurden die 100 meistzitierten Autor:innen mit dem Label *Economic Geography* selektiert.

Anschließend wurden die Literaturangaben zu sämtlichen Arbeiten dieser Autor:innen heruntergeladen und im nächsten Schritt nach Zeitschrift aggregiert, sodass eine sortierte Liste wirtschaftsgeographischer Zeitschriften mit Publikationshäufigkeit abgeleitet werden konnte. Um die bedeutendsten Zeitschriften zu selektieren, wurde ein Grenzwert berechnet. Dieser setzt sich aus dem Mittelwert aller Publikationshäufigkeiten, addiert mit der Standardabweichung aller Publikationshäufigkeiten, zusammen. Übersteigt die Publikationshäufigkeit einer Zeitschrift diesen

Grenzwert, wurden alle bibliometrischen Angaben sowie Abstracts der veröffentlichten Artikel dieser Zeitschrift aus der Literaturdatenbank Web of Science heruntergeladen. Insgesamt wurden somit 49.514 Abstracts und bibliometrische Angaben aus 43 Zeitschriften bezogen. Tabelle 14 zeigt die zehn größten enthaltenen Zeitschriften nach Artikelanzahl.

Tabelle 14: Die zehn größten betrachteten Zeitschriften nach Artikelanzahl.

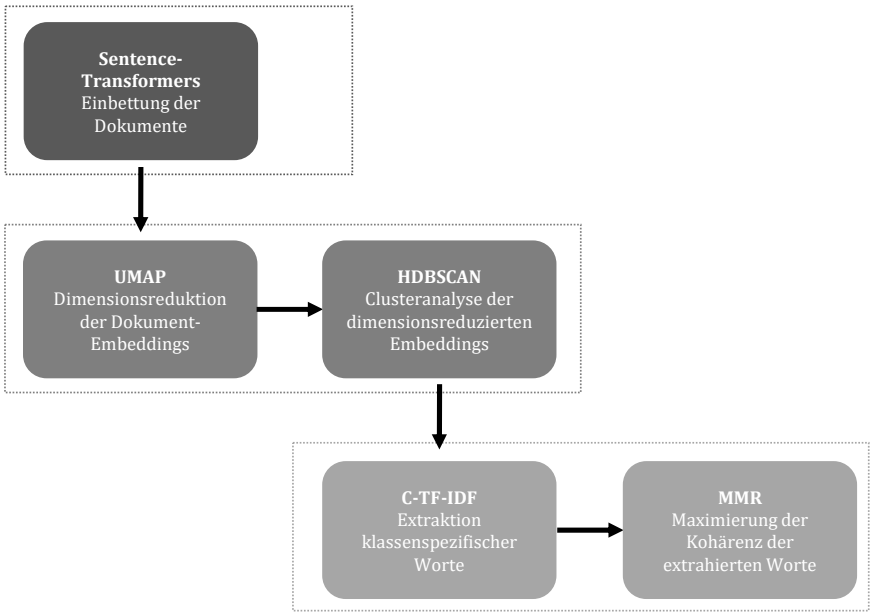
Zeitschrift	Anzahl Artikel
URBAN STUDIES	3.586
RESEARCH POLICY	2.819
GEOFORUM	2.817
REGIONAL STUDIES	2.305
AMERICAN ECONOMIC REVIEW	2.238
WORLD DEVELOPMENT	2.018
SMALL BUSINESS ECONOMICS	1.972
EUROPEAN PLANNING STUDIES	1.907
JOURNAL OF INTERNATIONAL ECONOMICS	1.865
INTERNATIONAL JOURNAL OF URBAN AND REGIONAL RESEARCH	1.496

Quelle: Eigene Berechnung basierend auf Web of Science.

Die Abstracts dieser Artikel wurden anschließend in ein Topic Model überführt. Die Abstracts eignen sich aus zweierlei Gründen besonders gut für die Themenmodellierung. Einerseits sind in der Kurzzusammenfassung sämtliche relevante Inhalte, wie übergeordnete Problemstellung, Fragestellung, Methodik und Ergebnisse des Papiers enthalten. Andererseits eignet sich die verwendete Embedding-Technik besonders gut für kürzere Dokumente (vgl. Kapitel 4). Die durchschnittliche Länge der betrachteten Abstracts beträgt 152 Wörter und ist somit geeignet, um diese mithilfe von Sentence-Transformermodellen numerisch einzubetten.

Abbildung 38 veranschaulicht den Prozessablauf des Algorithmus des zugrundeliegenden Topic Models. Im ersten Schritt dieser Fallstudie werden entsprechend die Abstracts der ausgewählten Artikel in dichte numerische Wortrepräsentationen übersetzt. Diese hochdimensionalen Vektoren werden genutzt, um semantische Ähnlichkeiten zwischen den Dokumenten aufzudecken. Zur Einbettung der Dokumente können prinzipiell alle verfügbaren vortrainierten Sentence-Transformersprachmodelle verwendet werden. Für diese Fallstudie wurde das Modell „all-mpnet-base-v2“ verwendet (HUGGING FACE o.J.). Dieses Modell wurde auf Basis von über einer Milliarde Sätzen aus unterschiedlichen Quellen trainiert. Ein großer Teil der Trainingsdaten sind wissenschaftliche Artikel, sodass sich das gewählte Sprachmodell gut zur Prozessierung wissenschaftlicher Literatur eignet.

Abbildung 38: BERTopic Algorithmus.



Quelle: verändert nach GROOTENDORST (2022).

Anschließend folgte der Einbettung der Dokumente ein zweiter Schritt zur Dimensionsreduktion und Clusteranalyse der Sentence-Embeddings. Standardmäßig wird mittels des gewählten Transformermodells ein Dokument in einem 768-dimensionalen Vektorraum verortet. Typischerweise haben Clusteralgorithmen Probleme, solche hochdimensionalen Daten zu verarbeiten. Entsprechend nutzt das BERTopic-Modell die „Uniform Manifold Approximation and Projection for Dimension Reduction“ (UMAP), um die komplexen Vektoren in eine niedrigdimensionalere Version umzuwandeln (MCINNES et al. 2020). Die dimensionsreduzierten Vektoren wurden anschließend in einen hierarchischen und dichtebasierten Clusteralgorithmus überführt, um die einzelnen Dokumente in Cluster einzuteilen.

Für die berechneten Cluster wurden in einem dritten Schritt klassenspezifische Wörter extrahiert, um den einzelnen Clustern einen Titel zuweisen zu können. Dies geschieht mittels einer adaptierten Version der TF-IDF-Technik (vgl. Kapitel 3.1.2). Diese bewertet typischerweise die Bedeutung einzelner Wörter für ein Dokument innerhalb eines Textkorpus. Der BERTopic-Algorithmus hingegen betrachtet alle Dokumente eines Clusters als ein Dokument. Somit konnte die Bedeutung einzelner Wörter für die jeweiligen Cluster berechnet werden. Die entstehenden Titelkandidaten wurden in einem letzten Schritt nochmals gefiltert, um besonders kohärente Clusterbeschreibungen zu erhalten (GROOTENDORST 2022).

Neben der Auswahl des Sprachmodells zur Einbettung der Dokumente lassen sich noch weitere Einstellungen bei der Themenmodellierung vornehmen. Entsprechend wurde festgelegt, dass der Algorithmus Bi- und Trigramme bei der Berechnung berücksichtigt und englische Stopwörter aus

den Texten nach der Dokumenteneimbettung entfernt. Darüber hinaus wurde als minimale Topicgröße der Wert 50 gewählt. Einem Topic müssen also mindestens 50 Dokumente (circa 0,001 % aller Dokumente) zugewiesen werden, um im Topic Model enthalten zu sein. Das verwendete Clusterverfahren bestimmt die initiale Clusteranzahl automatisch, wobei diese im Nachgang noch weiter reduziert werden kann.

8.3 Analyse des Topic Modelings

In dieser Fallstudie beträgt die initiale Clusteranzahl 109. Dabei annotiert der Clusteralgorithmus vorerst nicht jedes Dokument mit einem Topic, sondern weist Ausreißern, die nicht eindeutig zugeordnet werden können, kein Topic zu. Im ersten Durchlauf war dies bei 21.624 Dokumenten der Fall. Allerdings wird für jedes Dokument eine Wahrscheinlichkeitsmatrix ausgegeben, sodass im Nachgang nahezu jedes Dokument mit einem dominanten Topic versehen werden kann. Hierzu wurde den Dokumenten das Topic mit der höchsten Wahrscheinlichkeit zugewiesen. Falls kein Wert der Wahrscheinlichkeitsmatrix für ein Topic den Wert 0,01 übersteigt, wurde dem zugehörigen Dokument kein Topic zugewiesen. Dies war final für 2.164 Dokumente der Fall.

8.3.1 Deskriptive Analyse und Fusion der Topics

Standardmäßig wird jedes Topic mit zehn Begriffen beschrieben. Tabelle 15 gibt einen Überblick über die zehn größten Topics, deren Beschreibungen und den Anteil an allen klassifizierten Dokumenten. Aufgrund der hohen Anzahl unterschiedlicher Topics beschreiben die initial berechneten Topics jeweils nur eine geringe Zahl an Dokumenten.

Tabelle 15: Überblick über die zehn größten Topics des initialen Topic Models.

#	Topic Beschreibung	Anteil
0	innovation, knowledge, regional, cluster, clusters, regions, regional innovation, firms, innovative, networks	4,65 %
6	entrepreneurship, entrepreneurial, entrepreneurs, self, employment, business, growth, self employment, firms, firm	3,85 %
1	planning, governance, local, policy, regional, spatial planning, spatial, development, european, urban	3,61 %
3	geography, research, geographers, work, geographical, human, geopolitics, critical, political, military	2,94 %
2	trade, export, countries, model, country, rights, tariff, tariffs, firms, exports	2,82 %
4	exchange, rate, monetary, model, debt, shocks, currency, countries, financial, inflation	2,44 %
7	migration, immigrants, migrants, immigration, labor, employment, job, workers, data, market	2,01 %
8	commuting, travel, congestion, transport, transit, parking, time, commute, car, road	1,99 %
5	china, urban, chinese, land, state, development, housing, rural, economic, city	1,98 %
9	innovation, firms, technological, knowledge, product, design, firm, technology, performance, external	1,96 %

Quelle: Eigene Darstellung.

Bei Betrachtung der Tabelle 16 fällt auf, dass die jeweiligen Topic-Beschreibungen relativ kohärent sind. So subsumieren sich unter Topic 0 Artikel zur Rolle von Nähe auf die Innovativität von Unternehmen und Regionen, Clustern, Unternehmensnetzwerken und Wissensaustausch. Topic 6 stellt das zweitgrößte Topic dar und fällt ebenfalls in die Sparte der unternehmensbezogenen Arbeiten. Inhaltlich sammeln sich unter Topic 6 Arbeiten, die sich mit den Themen Selbständigkeit, Unternehmensführung und –wachstum auseinandersetzen. Topic 1 hingegen hat einen deutlich anderen inhaltlichen Fokus und adressiert Planungs- und Governancethemen, politische Eingriffsmöglichkeiten und Regionalentwicklung.

Unter Topic 3 subsumieren sich Arbeiten, die sich mit Geopolitik und Militärgeographie beschäftigen. Außenwirtschaftsbeziehungen auf Länderebene, Zölle und Steuern behandeln unter Topic 2 gefasste Artikel. Topic 4 weist ebenfalls einen Finanzfokus auf und beinhaltet Artikel, die sich mit Wechselkursen, Schuldenkrisen und Inflation befassen. Topic 7 fokussiert Themen der arbeitsbezogenen Migration. Die Themen öffentlicher Verkehr, Stau und Pendeln sind unter Topic 8 gesammelt. Topic 5 hat einen regionalen Fokus auf China und beschreibt Arbeiten zur Stadt- und Wirtschaftsentwicklung. Topic 9 weist augenscheinlich eine gewisse Schnittmenge mit Topic 0 auf. Es geht ebenfalls um unternehmerische Innovation. Allerdings mit einem stärkeren Produktfokus, da dieses Topic unter anderem mit den Wörtern Design oder Produkt beschrieben wird. Zusammenfassend lässt sich also festhalten, dass sowohl die zehn größten Themen des initialen Topic Modelings in sich semantisch kohärente Beschreibungen aufweisen als auch ein deutlicher Differentierungsgrad zwischen den Themen besteht.

Insgesamt bestehen zwischen den initial berechneten 109 Topics teilweise größere Ähnlichkeiten. Um diese Überschneidungen aufzulösen und die Topicanzahl in eine übersichtlichere Größe zu überführen, können benachbarte Topics fusioniert werden, wobei auch die Topic-Beschreibungen entsprechend der neuen Dokumentenzusammensetzungen der Cluster aktualisiert werden. Zur Reduktion der Topicanzahl kann entweder eine gewünschte Topicanzahl vorgegeben oder eine automatische Themenreduktion vorgenommen werden. Erstere Option geht mit dem Nachteil einher, dass Themen unabhängig von ihrer Ähnlichkeit fusioniert bzw. aufgeteilt werden, bis die gewünschte Topicanzahl erreicht ist. Bei einer automatischen Reduktion wird die Anzahl der Themen, beginnend mit dem kleinsten Thema, solange reduziert, bis eine Cosinusähnlichkeit zwischen den Themen von 0,915 unterschritten wird.

Nach der automatischen Themenreduktion umfasst das Topic Model noch 79 unterschiedliche Themen. Durch die Fusionierung einiger Themen ändern sich sowohl deren Beschreibung und Nummerierung als auch die Größe der einzelnen Topics. Tabelle 16 zeigt die zehn häufigsten Themen nach der automatischen Topicreduktion.

Tabelle 16: Überblick über die zehn größten Topics nach der Themenreduktion.

#	Topic Beschreibung	Anteil
0	innovation, firms, knowledge, firm, business, regional, research, new, performance, technology	17,62 %
1	urban, city, social, housing, state, cities, china, development, land, political	8,16 %
2	environmental, climate, conservation, forest, change, climate change, political, land, management, governance	5,77 %
3	planning, policy, local, governance, regional, development, urban, european, political, spatial	3,61 %
4	geography, research, geographers, work, political, human, geographical, critical, geopolitics, geographies	2,94 %
5	trade, countries, export, model, country, rights, firms, tariff, exports, tariffs	2,82 %
6	migration, employment, workers, labor, job, wage, immigrants, data, areas, market	2,81 %
7	exchange, rate, exchange rate, model, monetary, shocks, debt, financial, countries, trade	2,43 %
8	housing, mortgage, property, market, rent, homeownership, income, urban, home, households	2,43 %
9	research, university, universities, scientific, science, academic, knowledge, scientists, technology, researchers	2,34 %

Quelle: Eigene Darstellung.

Die Themenfusion hat eine Verschiebung der Themengrößen und –beschreibungen zur Folge. Die einzelnen Topics sind inhaltlich nun weiter gefasst und repräsentieren folglich eine größere Zahl an Dokumenten. Beispielsweise wurden einige Themen, die sich inhaltlich mit Unternehmen, Innovationen, Technologien und wissensbasierter Regionalentwicklung beschäftigen, fusioniert, sodass Topic 0 nach der Themenreduktion das deutlich größte Topic darstellt. Ähnlich verhält es sich mit Topic 1, welches inhaltlich etwas breiter gefasst ist und daher nun mehrere semantisch ähnliche Topics zu Urbanisierung und Stadtentwicklung vereint. Topic 2 erscheint erst durch die Themenfusion in der Liste der zehn größten Topics. Inhaltlich sind unter Topic 2 sämtliche Artikel kategorisiert, die sich mit Umwelt- und Naturschutz sowie Klimawandel auseinandersetzen. Die Topics 3 bis 7 weisen nach der Themenreduktion nur marginale Veränderungen in Größe und Semantik im Vergleich zum initialen Topic Model auf. Topic 8 hingegen erscheint nach der Themenfusion neu in den zehn größten Topics und setzt sich mit Kauf- und Mietpreisen, dem Immobilienmarkt in Städten sowie Haushaltseinkommen auseinander. Topic 9 kategorisiert Arbeiten, die Universitäten, Wissenschaftler:innen, Wissen und Technologie in den Fokus der Betrachtung nehmen.

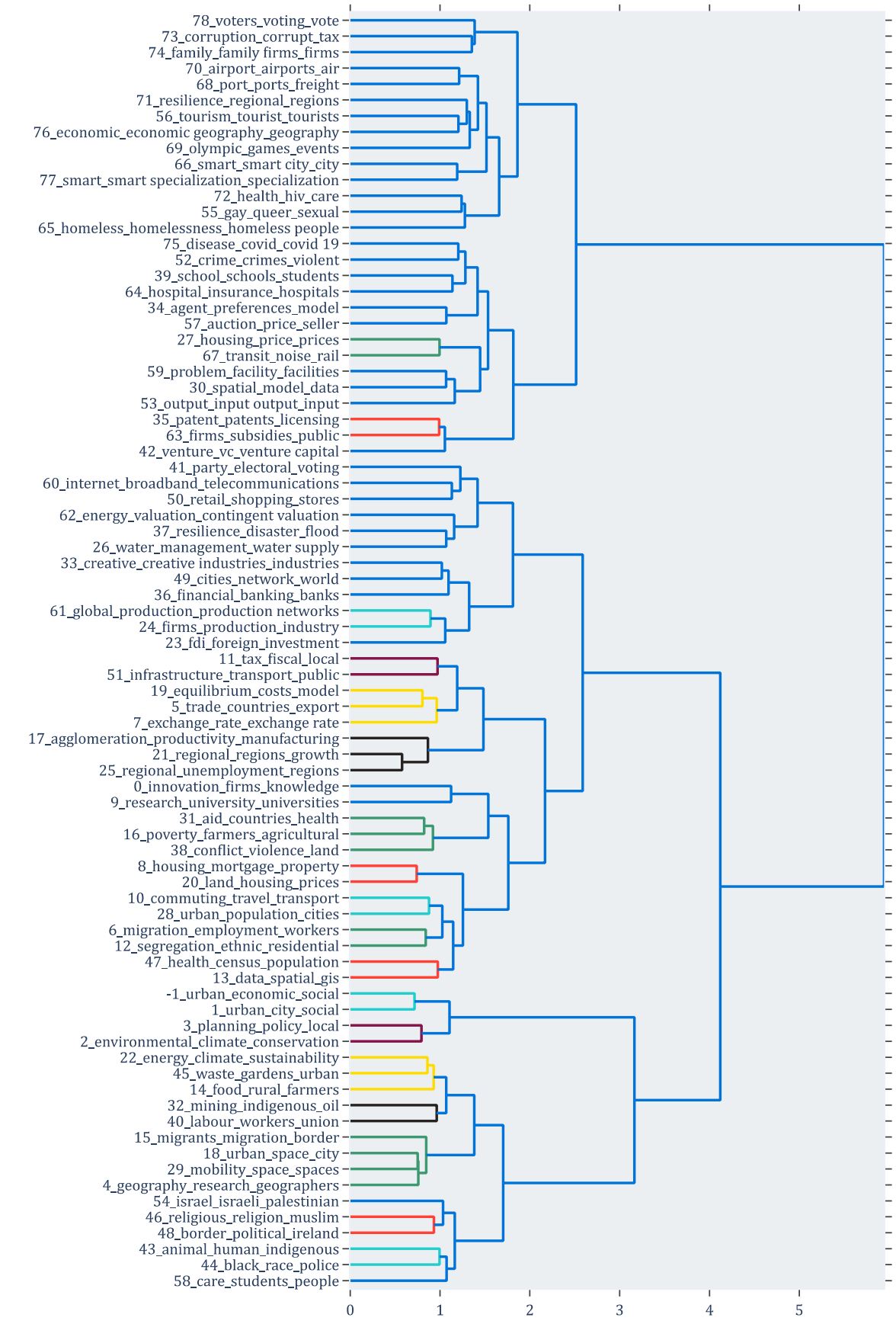
Des Weiteren lässt sich nachvollziehen, welche zuvor disjunkten Themen durch die automatische Themenreduktion fusioniert wurden. Das Topic 0 vereint nach der Themenfusion beispielsweise insgesamt 12 Themen des initialen Modells. Diese 12 Themen unterscheiden unter anderem Arbeiten zu KMU, ausländischen Direktinvestionen, Agglomerationsfaktoren, Internationalisie-

rungsstrategien oder Firmenkultur. Das neue Topic 1 wurde aus sieben Topics des initialen Modells gebildet. Die initialen Topics unterscheiden thematisch beispielsweise nach Arbeiten zu Bevölkerungs- und Städtewachstum sowie -schrumpfung, Gentrification und nehmen räumlich insbesondere Schwellen- und Entwicklungsländer wie Indien, China oder Südafrika in den Fokus. Das neue Topic 2 besteht ebenfalls aus insgesamt sieben initialen Topics. Thematisch unterscheiden die nun fusionierten Themen nach Vegetation und Niederschlag, Umwelt- und Luftverschmutzung, dem Klimawandel, CO^2-Emissionen, Kohlenstoffmärkten, Rodung von Wäldern und Umwelt- und Naturschutz.

8.3.2 Semantische Verwandtschaften der Topics

Da das BERTopic-Modell sowohl numerische Repräsentationen von Dokumenten als auch von ganzen Topics berechnet, kann auf Basis der Topic-Embeddings die Verwandtschaft der einzelnen Themen quantifiziert werden. Diese thematische Ähnlichkeit kann einerseits genutzt werden, um Themen zu fusionieren. Andererseits hilft sie bei der Strukturierung von Themenzusammenhängen. Methodisch wird hierfür auf Basis der Cosinus-Ähnlichkeit der Topic-Embeddings eine hierarchische Clusteranalyse durchgeführt. Die Linkages zwischen den Themen wurden mittels des Ward-Algorithmus bestimmt. Abbildung 39 zeigt ein Dendrogramm, welches sämtliche Topics sowie die Stufen der hierarchischen Klassifikation darstellt. Besonders ähnliche Topics mit Abstandswerten unter eins wurden nochmals farblich hervorgehoben. Der geringste Abstand besteht beispielsweise zwischen Topic 21 und Topic 25, welche sich mit regionalem Wachstum und Konvergenz respektive regionalen Beschäftigungsquoten auseinandersetzen. Weitere Topics mit großen Ähnlichkeiten sind Topic 35 und Topic 63, die das Patentverhalten von Unternehmen, Firmensubventionen und Fördergelder betrachten. Topic 27 beleuchtet das Thema der Preise für Häuser, Wohnungen sowie Bauland und weist eine starke Ähnlichkeit zu Topic 67 auf, welches Papiere zur Erreichbarkeit von Bahnstationen und Verkehrslärm subsumiert. Weitere thematische Gruppierungen sind bezüglich der Themen Steuern und staatliche Infrastruktur (Topic 11 & Topic 51), Produktivität von globalen Unternehmensnetzwerken (Topic 61 & Topic 24), Preisbildung, Wechselkurse und Wettbewerb (Topics 19, 5 und 7), Agglomerationsfaktoren, regionaler Wachstum und regionale Beschäftigung (Topics 17, 21, 25), Entwicklungshilfe, Armut von Landwirten und Landnutzungskonflikte (Topics 31, 16, 38), Immobilienpreisen, Hypotheken und Wohneigentum (Topics 8 und 20), öffentlicher Transportnetzwerke und Urbanisierung/Städtewachstum (Topics 10 und 28), Migration und Segregation (Topics 6 und 12), Gesundheits- bzw. Zensusdaten und GIS-Methoden (Topics 47 und 13), Raumplanung und Klimaschutz bzw. –adaption (Topics 3 und 2), erneuerbare Energien, Müll und Landwirtschaft (Topics 22, 45 und 14), Gewerkschaften und der Abbau fossiler Rohstoffe (Topics 32 und 40), Migration, urbane Räumen, öffentlicher Vekehr und Geopolitik (Topics 15, 18, 29 und 4), religiöse bzw. politische Konflikte (Topics 46 und 48) sowie ethnische Minderheiten (Topics 43 und 44) zu beobachten.

Abbildung 39: Dendrogramm des fusionierten Topic Models.

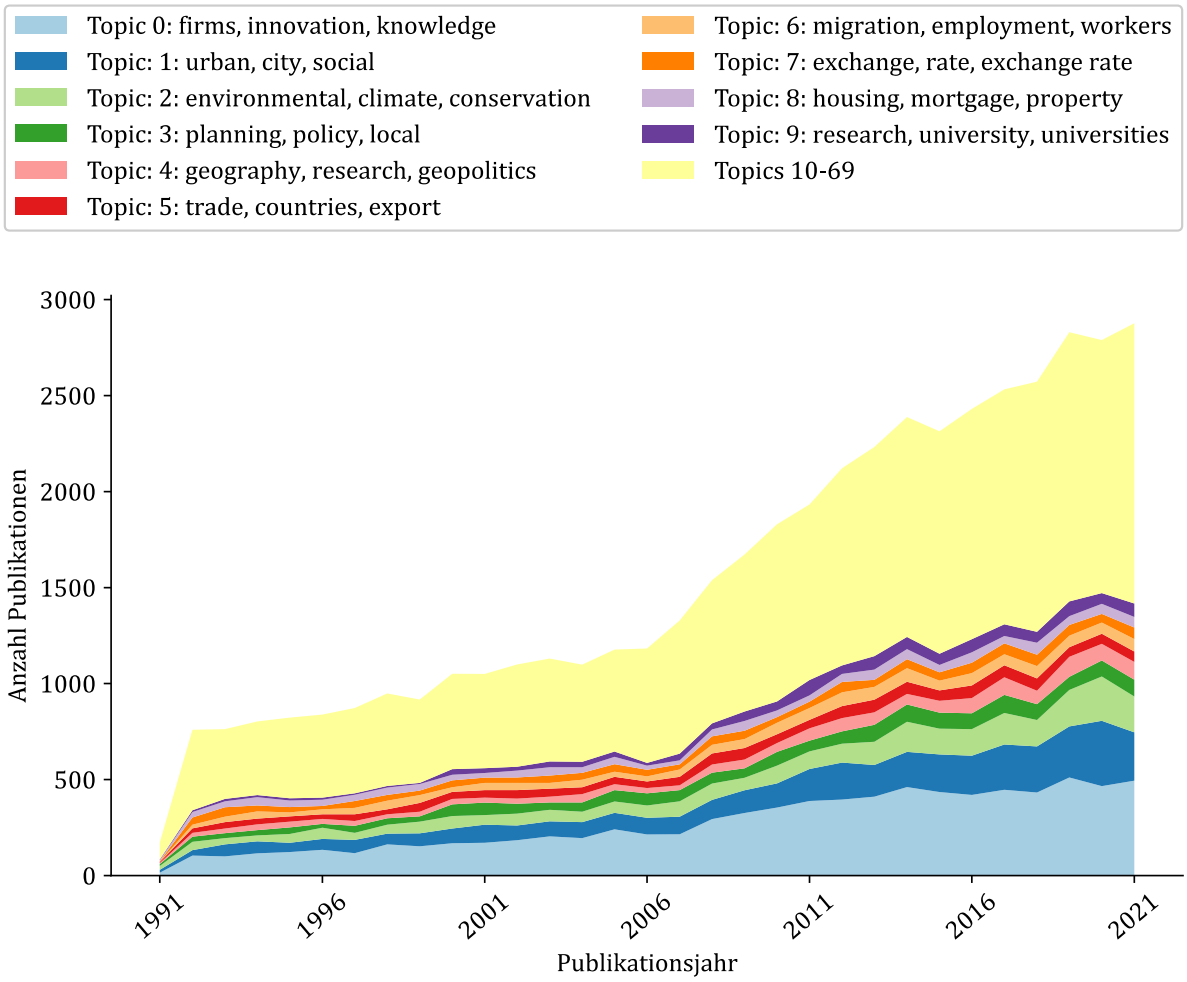


Quelle: Eigene Darstellung.

8.3.3 Dynamisches Topic Modeling

Neben der Querschnittsbetrachtung der Themen und deren Beziehung untereinander lassen sich die Themenentwicklungen ferner über Zeit betrachten. Abbildung 40 gibt einen allgemeinen Überblick über die Themenverteilung sowie Artikelanzahl im Zeitverlauf. Die Themenbenennung wurde für Abbildung 40 zur Wahrung der Übersichtlichkeit nochmals verkürzt.

Abbildung 40: Anzahl der Publikationen nach Thema im Zeitverlauf.



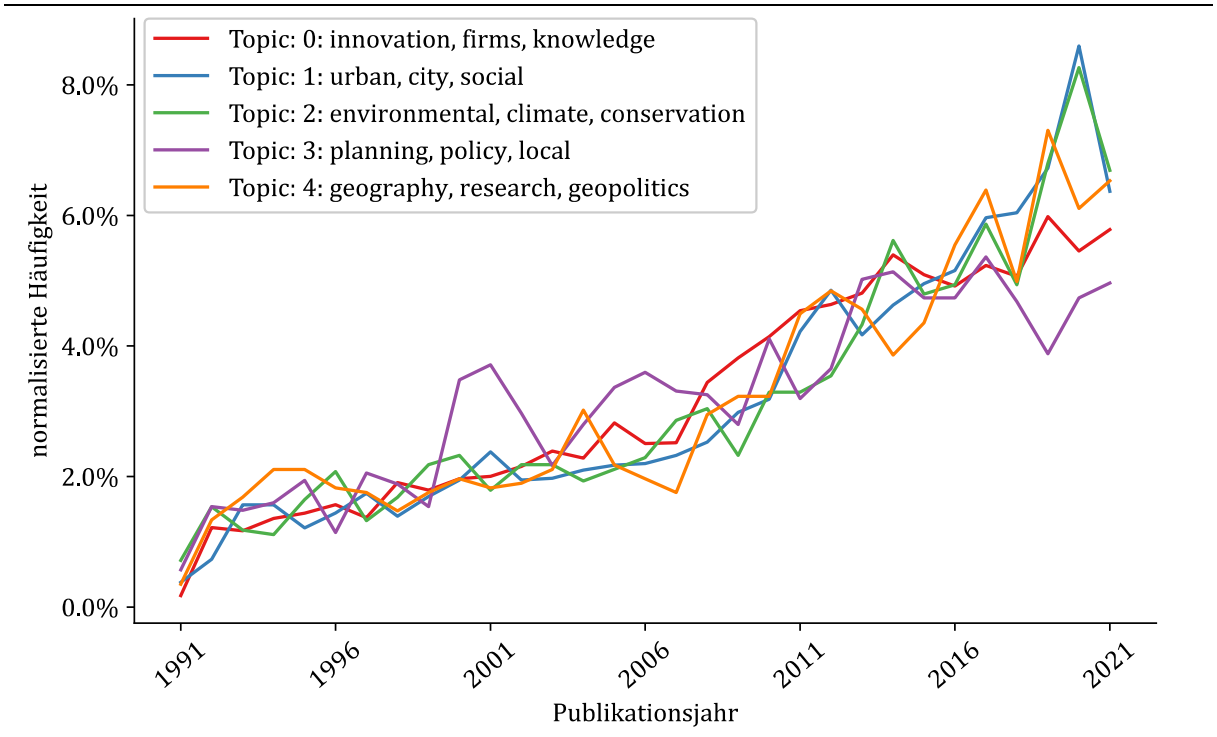
Quelle: Eigene Darstellung.

Insgesamt decken die zehn größten Themen fast die Hälfte aller betrachteten Publikationen ab. Auffällig ist ebenfalls, dass die Gesamtzahl der Publikationen im Zeitverlauf insgesamt stark zunimmt. Während zu Beginn des Jahrtausends jährlich rund 1.000 Artikel in den ausgewählten Journals publiziert wurden, sind es 2021 über 2.500 Artikel pro Jahr. Gleichzeitig findet über die Zeit eine stärkere Ausdifferenzierung der Themen statt. So nimmt der Anteil der kleineren Topics im Zeitverlauf stetig zu. Dies ist neben der vermehrten Publikation von Randthemen ebenso mit der Entstehung neuer Topics zu begründen. Beispiele für Topics, die sich erst in der jüngeren Vergangenheit als eigene Forschungsrichtung herausgebildet haben, sind Topic 77 (Smart Specialization), Topic 74 (Familienunternehmen), Topic 66 (Smart City) oder Topic 69 (Sport-Events,

olympische Spiele). Themen, die im Laufe der Zeit zunehmend an Bedeutung gewonnen haben, sind unter anderem Topic 75 (Krankheiten, Epidemien), Topic 71 (regionale Resilienz), Topic 55 (sexuelle Minderheiten), Topic 37 (Naturkatastrophen) und Topic 22 (erneuerbare Energien). Auch innerhalb der zehn größten Themen sind in Abbildung 39 leichte Verschiebungen zu beobachten. Beispielsweise zeigt sich eine sukzessive Zunahme von Topic 2 insbesondere seit Beginn der 00er Jahre. Auch die Topics 0, 1, 3, 5 und 9 weisen moderate Zunahmen in der Artikelanzahl über Zeit auf.

Um diese Verschiebungen genauer analysieren zu können, lassen sich die normalisierten Häufigkeiten pro vergebenem Topic nach Publikationsjahr analysieren. Abbildung 41 zeigt die Entwicklung der fünf größten Topics auf. Hierzu wurden die thematisch zugeordneten Artikel nach Publikationsjahr aggregiert und auf Basis der Gesamthäufigkeit eine normalisierte Häufigkeit pro Jahr errechnet. Trotz der allgemeinen Zunahme der Publikationstätigkeit weisen die Trendlinien der einzelnen Topics dennoch individuelle Entwicklungen auf. Beispielsweise weist Topic 3 im Vergleich zu den anderen vier Themen eine flachere Entwicklung auf.

Abbildung 41: Dynamisches Topic Modeling der fünf größten Topics.



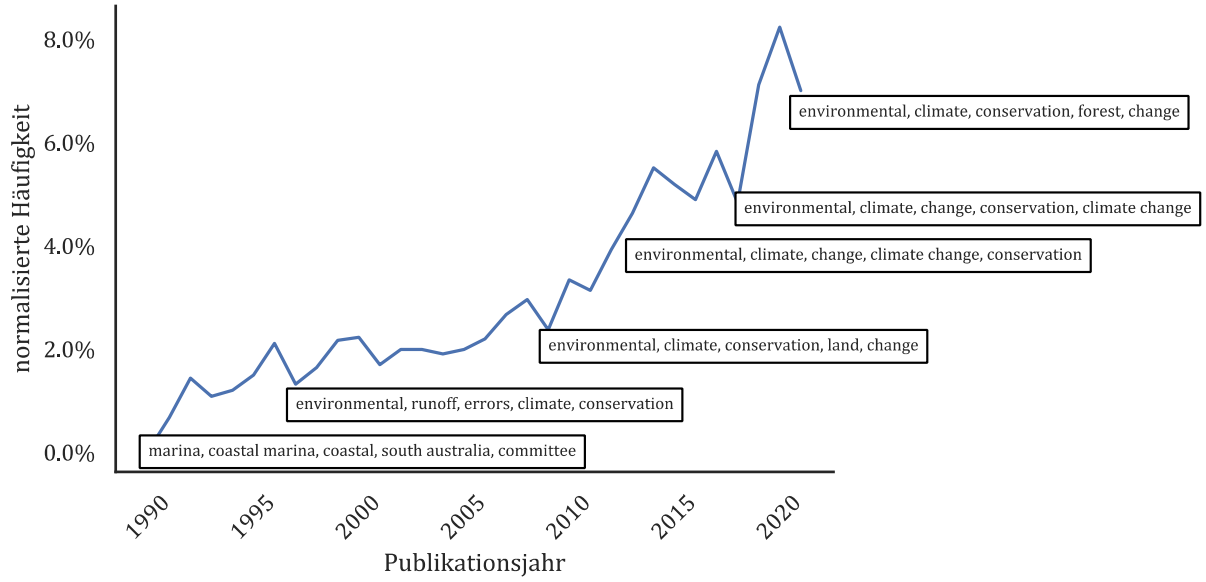
Quelle: Eigene Darstellung.

Deutliche Anstiege sind seit Beginn der 2010er Jahre bei Topic 1 und Topic 2 zu beobachten. Bezüglich Topic 2 fällt besonders die Entwicklung ab Ende der 2010er Jahre auf, die im Jahr 2020 mit über 8 % normalisierter Häufigkeit ihr derzeitiges Maximum erreicht. Um die Gründe für diese Entwicklungen besser verstehen zu können, kann neben einer globalen Betrachtung der Themenentwicklungen über Zeit auch die lokale Themenentwicklung analysiert werden. Methodisch wird

dazu für jedes Publikationsjahr und jedes Topic eine gesonderte Themenrepräsentation berechnet, um die jeweiligen Themenschwerpunkte einzelner Themen zu bestimmten Zeitpunkten analysieren zu können. Hierzu können die a priori berechneten Dokumentencluster genutzt werden. Diese werden nach Zeitpunkt und Cluster aggregiert, sodass je Cluster und Zeitpunkt eine gesonderte Themenbeschreibung aus den bestehenden Clustereinbettungen erzeugt werden kann. Dieses Vorgehen geht mit dem Vorteil einher, dass nicht für jedes Themencluster und jeden Zeitpunkt ein gesondertes Modell berechnet werden muss, sondern lediglich eine Anpassung des Querschnittmodells erfolgt.

Wie bereits aus Abbildung 41 hervorgeht, ist das Publikationsgeschehen von Topic 2 (enviromental, climate, conversion) besonders dynamisch, sodass eine dedizierte Betrachtung der lokalen Themenentwicklung dieses Topics sinnvoll erscheint. Abbildung 42 zeigt dementsprechend diese Entwicklung auf. Die Verlaufslinie ist an ausgewählten Zeitpunkten mit den lokalen Themenbeschreibungen des Topic Models annotiert. Bei Betrachtung dieser lokalen Themen innerhalb von Topic 2 fällt auf, dass mit dem rapiden Anstieg der relativen Publikationshäufigkeiten ein veränderter thematischer Schwerpunkt einhergeht.

Abbildung 42: Lokale Themen innerhalb von Topic 2 im Zeitverlauf.



Quelle: Eigene Darstellung.

Während bis Anfang der 2010er Jahre unterschiedliche Themen wie Küsten, Wälder und Erosion im Fokus standen, ist ab 2011 der Klimawandel das dominierende Thema innerhalb des Topics 2. Somit unterstreicht die Untersuchung, dass das Phänomen des Klimawandels ab Ende der 00er Jahre den auschlaggebenden Grund für eine intensivere wissenschaftliche Auseinandersetzung im (wirtschafts-)geographischen Forschungskanon darstellt. Diese Entwicklung hält bis heute an

und kann anhand der lokalen Themenbeschreibungen nachvollzogen werden. So treten die Wörter *climate* und *change* in der lokalen Themenbeschreibung erstmals im Jahr 2000 gemeinsam auf. Das Bigramm *climate change* erscheint 2010 erstmals in der lokalen Themenbeschreibung des Topic 2. Auffällig ist ebenfalls, dass das Thema des Klimawandels eine immer zentralere Rolle in den lokalen Topicbeschreibungen einnimmt, während andere Themenschwerpunkte (z.B. Wälder) sukzessive aus den lokalen Themenbeschreibungen verschwinden. Die Ergebnisse weisen damit Überschneidungen mit vergleichbaren Studien auf, die mit klassischen bibliometrischen Methoden (soziale Netzwerkanalysen, Zitationsanalysen, Kookurrenzanalysen) wissenschaftliche Klimaliteratur beleuchten. So zeigen HOU und WANG (2021) eine deutliche Zunahme von Klimastudien ab Mitte der 2000er Jahre, sodass zwischen 2006 und 2016 mehr als die Hälfte aller von den Autor:innen untersuchten Artikel publiziert wurden. Auch die Ergebnisse von TAN et al. (2021) sowie von HUANG et al. (2020) bescheinigen eine moderate Zunahme der klimabezogenen Publikationen ab Mitte der 2000er Jahre, an welche sich ein starkes Wachstum ab Mitte der 2010er Jahre anschließt. HAUNSCHILD et al. (2016) analysieren die Titel von über 220.000 klimabezogenen Publikationen und zeigen eine deutliche Zunahme des Begriffs „Klimawandel" im Zeitverlauf. Ferner belegen die Ergebnisse der Autor:innen ein gestiegenes Interesse an Klimaforschung jenseits der Naturwissenschaften ab Beginn der 2010er Jahre.

Diese Entwicklungen der klimabezogenen Forschung in der Wirtschaftsgeographie weisen Überlagerungen mit klimapolitischen Zielen und Vereinbarungen auf. Zwar finden seit den 1970er Jahren internationale Treffen und Verhandlungen zur Klimapolitik statt. Allerdings bestehen bindende Verpflichtungen zum Schutz des Klimas erst seit 2005. In diesem Jahr sorgte das Inkrafttreten des Kyoto-Protokolls erstmals für verbindliche Zielwerte für den Treibhausgasausstoß. Im weiteren Zeitverlauf löste der 5. Sachstandsbericht des Intergovernmental Panel on Climate Change (IPCC) im Jahr 2014 intensivere Diskussionen aus. Dieser betonte nochmals eindringlich den Einfluss des Menschen auf Klimaveränderungen sowie die Tatsache, dass Mensch und Natur bereits jetzt die Auswirkungen des Klimawandels spüren (BULKELEY und NEWELL 2015). Im Jahr 2015 wurden gleich zwei zentrale multilaterale Vereinbarungen getreffen. Das Übereinkommen von Paris, welches als Nachfolger des Kyoto-Protokolls verabschiedet wurde, fordert die Staaten dazu auf, Maßnahmen zu ergreifen, um die globale Erwärmung auf deutlich unter 2 Grad Celsius im Vergleich zum vorindustriellen Niveau zu begrenzen (VEREINTE NATIONEN 2015a). Die globalen Nachhaltigkeitsziele wurden ebenfalls in 2015 von den vereinten Nationen verabschiedet und verfolgen das übergeordnete Ziel, bis 2030 eine nachhaltige Entwicklung zu fördern (VEREINTE NATIONEN 2015b).

Diese politischen und gesamtgesellschaftlichen Entwicklungen werden, wie den Ergebnissen des Topic Models zu entnehmen ist, intensiv in wirtschaftsgeographischer Forschung diskutiert. Für

Volkswirtschaften und Unternehmen führt beispielsweise der Emissionshandel zu einer stärkeren Integration von Klimaschutzaspekten in das wirtschaftliche Handeln.

Zusammenfassend lässt sich festhalten, dass das verwendete Topic Model viele unterschiedliche Möglichkeiten zur unüberwachten Analyse großer Textmengen bereitstellt. Der Modellierungsprozess ist modular aufgebaut, sodass daten- und fragestellungsabhängig unterschiedliche Algorithmen beispielsweise zur Dokumenteneinbettung, Dimensionsreduktion oder Clusterbildung verwendet und kombiniert werden können. Ein zentraler Vorteil von BERTopic ist die Integration moderner vortrainierter Sprachmodelle. Diese ermöglichen eine individuellere Kontextualisierung der Dokumente und somit eine adäquatere numerische Repräsentation der Semantik, als BOW-basierte LDA-Modelle (GROOTENDORST 2022).

Neben einer überblickshaften Darstellung von Themen innerhalb einer Dokumentensammlung können deren semantische Verwandtschaft betrachtet und Hierarchien innerhalb der Themen modelliert werden. Insbesondere die dynamische Betrachtung des Topic Models erlaubt tiefgreifendere Analysen, die sowohl eine quantitative als auch eine qualitative Betrachtung von Themenveränderungen ermöglichen. Einen besonderen Mehrwert stellt die Analyse lokaler Themenveränderungen im Zeitverlauf dar. Diese ermöglicht neben einer rein quantitativen Betrachtung von Zu- oder Abnahmen von Themen über Zeit ebenfalls eine qualitativere Analyse durch die Beleuchtung von Änderungen in den Themenbeschreibungen. So können Auslöser und Triebkräfte quantitativer Veränderungen qualitativ begründet und gestützt werden.

Kritisch anzumerken ist einerseits, dass BERTopic standardmäßig jedem Dokument lediglich ein Thema zuordnet. In der Praxis kann ein Dokument jedoch mehrere Themen behandeln, sodass diese Differenzierung zunächst ignoriert wird. Allerdings können längere Dokumente nochmals in Absätze geteilt werden, um jeden Absatz als einzelnes Dokument zu behandeln und so mehrere Themen aus einem Dokument extrahiert werden können. Darüber hinaus können aus dem Topic Model keine Begründungen für die Ähnlichkeiten zwischen Themen entnommen werden.

9 Abschließende Synthese der Empirie

Im folgenden Kapitel werden die Ergebnisse der drei Fallstudien zusammengeführt und diskutiert. Das Analysedesign der Arbeit orientierte sich an bestehenden Systematiken des Web Minings und wurde durch die Integration unterschiedlicher Verfahren des NLP ergänzt. Der vierstufige Prozess bestehend aus Datenerhebung, Datenselektion und –aufberereitung, Datenanalyse sowie Interpretation konnte auf alle drei Fallstudien der Arbeit übertragen werden. Trotz der gemeinsamen Systematik weisen die unterschiedlichen Fallstudien verschiedene Foki auf. Im Rahmen der ersten Fallstudie wurden schwerpunktmäßig Ressourcen und Verfahren zur Datenbeschaffung und –erhebung vorgestellt. Die zweite Fallstudie setzte den methodischen Fokus auf die Analyse von Webmassendaten sowie die Integration dieser in klassische ökonometrische Modelle. Die dritte Fallstudie hat methodisch bedingt einen stärker interpretativen Charakter. Die Synthese der Fallstudien orientiert sich am Analysedesign der Arbeit und diskutiert jeweilige Besonderheiten.

9.1 Datenzugang und –erhebung

Datenzugang und –erhebung können im Rahmen von umfangreichen Web Mining-Vorhaben als größte Herausforderung gesehen werden. Der ausschlaggebende Grund ist die mangelnde Systematisierung von Webdomains. Der Umstand, dass bis dato keine strukturierte Webseitendatenbank besteht, führt zu erheblichen Anlaufkosten bei umfangreichen Web Mining-Vorhaben. Unter Umständen haben fehlende systematische Datenzugänge in der Vergangenheit eine Auseinandersetzung mit Webdaten behindert. In der Regel zielt Web Mining darauf ab eine bestimmte Gruppe von Webseiten zu untersuchen (beispielsweise Unternehmenswebseiten, Universitätswebseiten, Pressemeldungen, soziale Medien). In den wenigsten Fällen existiert ein zentrales Register, welches den Beobachtungseinheiten (z.B. Unternehmen) eine Webseite zuordnet. Kommerzielle Unternehmensdatenbanken bieten zwar durchaus solche Informationen an. Allerdings geht die Nutzung dieser neben finanziellen Aufwänden mit weiteren Limitationen einher (beispielsweise. eingeschränkte gemeinschaftliche Forschung). Somit ist ein zentraler erster Schritt eines jeden Web Mining-Projekts eine Liste mit zu untersuchenden Webdomains zu erstellen. Für Studien mit übersichtlichen Untersuchungsumfängen können diese Listen beispielsweise mithilfe von Suchmaschinen manuell erstellt und gepflegt werden.

Bei umfangreicheren Vorhaben ist ein automatisiertes Suchverfahren notwendig, um den Untersuchungseinheiten eine Webseite zuordnen zu können. Allerdings benötigen Forschende zunächst Zugang zu einer möglichst vollständigen Zahl an Webdomains, um darin nach den gewünschten Domains suchen zu können. Bisher existieren nur wenige umfangreiche Webrepositorien, die hierfür als Datenbasis dienen können. Die erste Fallstudie nutzte das CC-Projekt als Grundlage, um aus diesem deutsche Unternehmenswebseiten zu identifizieren. Die deskriptive

Analyse des Datensatzes zeigte, dass das CC als durchaus umfangreiche Datenquelle für Web Mining-Vorhaben dienen kann. Das CC offeriert die Daten in verschiedenen Formaten, sodass fragestellungsabhängig das passende gewählt werden kann. Im Rahmen von Fallstudie 1 hat sich der URL-Index, welcher neben einer Sammlung von Domains auch deren Sprache enthält, als wertvolle Grundlage erwiesen. Dieser ermöglicht einerseits eine vergleichsweise schnelle Prozessierung der Inhalte. So ist der URL-Index eines Crawls durchschnittlich um dem Faktor 300 kleiner als der vollständige HTML-Crawl (ca. 0,26 Tebibyte vs. ca. 82 Tebibyte). Der Datensatz, der neben den URLs und Metadaten noch die jeweiligen Webseitentexte enthält, ist im Durchschnitt um dem Faktor 35 größer als der verwendete URL-Index. Andererseits ist der Index mit Metadaten annotiert, sodass beispielsweise Webseiten effizient nach Sprache gefiltert werden können. Weiterhin zeigte die deskriptive Analyse, dass bereits wenige Datensätze des CC ausreichen, um einen Großteil der im CC enthaltenen deutschsprachigen Webseiten zu extrahieren. Beispielsweise enthält der erste betrachtete Crawl bereits über 42 % aller betrachteten Domains dieser Arbeit. Fallstudie 1 zeigte jedoch ebenfalls, dass ein eigenständiger Abruf der enthaltenen Webdomains notwendig ist, bevor diese für tiefgreifendere Analysen genutzt werden können. Webdomains weisen eine enorme Dynamik auf, sodass es bereits nach kurzer Zeit zu Änderungen oder Weiterleitungen kommen kann. Um eine aktuelle Liste an Domains zu erhalten, bedarf es dementsprechend nochmals eigenständiger Abrufe. Außerdem ist die Auswahl an Webseiten einer Domain, die im CC enthalten ist, zufällig. Falls Forschende an bestimmten Webseiteninhalten (z.B. dem Impressum) einer Domain interessiert sind, ist ebenfalls ein eigenständiger Abruf der Domain und eine entsprechende Suche nach den gewünschten Inhalten notwendig.

Neben einer umfangreichen Datengrundlage sind performante Werkzeuge zum Abruf und zur Analyse der Webseiten relevant. Für das Webscraping stehen in unterschiedlichen Programmiersprachen verschiedenste Pakete zur Verfügung. Im Rahmen dieser Arbeit wurde das Scrapy Framework (SCRAPY COMMUNITY 2022b) genutzt. Dieses kann für umfangreiche Web Scraping-Vorhaben abschließend als gut geeignet bewertet werden. Das wohl ausschlaggebenste Argument für die Nutzung von Scrapy sind die enorm geringen Laufzeiten. Beispielsweise ermöglicht Scrapy den Abruf von knapp 5.000 Webseiten pro Minute auf einem Laptop (16GB Arbeitsspeicher, 8 logische Prozessoren). Neben der Leistungsfähigkeit von Scrapy bietet das Framework nahezu beliebige Individualisierungsmöglichkeiten an. Es kann sowohl für systematisches Crawling als auch gezieltes Scraping genutzt werden. Scrapy bietet viele nützliche Komponenten an, die für das eigene Vorhaben lediglich aktiviert bzw. deaktiviert werden müssen. Das Framework ist ferner kompatibel mit gängigen Datenbanksystemen, sodass Verwaltung und Speicherung der Daten ebenfalls sehr komfortabel erfolgen kann.

Darüber hinaus besteht die Möglichkeit, eigene Logiken und Regeln anzuwenden. Beispielsweise wurde in Fallstudie 1 ein Suchverfahren implementiert, welches auf der Startseite der Domains

gezielt nach Links zu Impressumswebseiten suchte. Somit konnten die gewünschten Inhalte gezielt abgerufen und der Suchaufwand im Rahmen gehalten werden. Für Fallstudie 2 wurde ebenfalls Scrapy genutzt, um weitere Webseiten der identifzierten Unternehmensdomains zu beziehen. Das Programm wurde entsprechend angepasst, um alle internen Links der jeweiligen Domains ausgehend von der Startseite zu sammeln und deren Inhalte abzurufen. Fallstudie 3 zeigte hingegen, dass je nach Anwendungsfall auch strukturiertere Zugänge zu digitalen Textdaten bestehen. Für diese Fallstudie konnten die betrachteten Abstracts über den Browser abgerufen werden.

Für einige Sonderfälle ist dennoch das Selenium-Framework Scrapy vorzuziehen. Wenn gezielt mit Webseiten interagiert werden soll, beispielsweise um Formulare auszufüllen, Inhalte zu aktivieren bzw. anzuklicken oder über die Webseite zu scrollen, ist das Selenium-Framework besser geeignet. Dieses simuliert einen tatsächlichen Browseraufruf, sodass eben auch Interaktionen mit der Webseite möglich sind. Diese Simulation des realen Browserverhaltens geht allerdings zu Lasten der Abrufgeschwindigkeit. So können lediglich 2-3 Webseiten pro Minute mit Selenium abgerufen werden (eigene Berechnung).

9.2 Datenselektion und -aufbereitung

Nach dem Webseitenabruf gilt es den HTML-Code der Webseite auszulesen. Im Rahmen dieser Arbeit wurde für diesen Zweck das Python-Paket Beautiful Soup verwendet (RICHARDSON 2022). Dieses bietet eine Vielzahl von Möglichkeiten, HTML- und XML-Code auszulesen und zu verarbeiten. Darüber hinaus ist das Programm einfach parallelisierbar, sodass auch mehrere Millionen Webseiten in überschaubarer Zeit ausgelesen werden können. Für Fallstudie 1 wurden mittels Beautiful Soup sämtliche Textelemente aus dem HTML-Code extrahiert und an das NER-Modell überführt.

Webseiten bestehen neben dem zentralen Webseitentext aus weiteren Textelementen, die aufsetzende Analysen behindern können. Beispielhaft zu nennen sind Textelemente in Kopf- und Fußzeilen, Verzeichnisse oder Hyperlinks. Zur Extraktion des zentralen Webseitentexts nutzte Fallstudie 2 das Python-Paket Trafilatura (BARBARESI 2021). Dieses ermöglichte es, automatisiert und zuverlässig die Volltextpassagen der Webseiten zu extrahieren und somit für die weitere Analyse vorzubereiten. Für Fallstudie 3 konnte aufgrund der standardisierten und qualitativ hochwertigen Datengrundlage gänzlich auf eine Vorverarbeitung verzichtet werden.

Neben der Textextraktion existieren im Bereich des NLP viele weitere Methoden zur Vorverarbeitung von Text. Beispielhaft zu nennen sind die Entfernung von Stopwörtern, Lemmatisierung, Stemming oder Tokenizing. Diese Verfahren zur Textstandardisierung waren in der Vergangenheit notwendig, um die standardisierten Wörter mittels starrer Word-Embeddings in numerische

Repräsentationen überführen zu können. Die Entwicklung von vortrainierten Transformermodellen machen diese Vorverarbeitungsschritte weitestgehend obsolet. Während starre Word-Embeddings (z.B. Word2Vec, BOW), wie sie vor der Einführung von Transformermodellen verwendet wurden, auf standardisiertem Text trainiert wurden, wird für das Training von Transformermodellen nicht standardisierter Text genutzt (vgl. Kapitel 4.6). Damit ist Textstandardisierung in Zeiten der Transformermodelle lediglich für spezielle Aufgaben relevant. Soll beispielsweise ein Text nach bestimmten Stichwörtern durchsucht werden, ergibt es Sinn diesen im Vorfeld zu standardisieren, um sämtliche grammatikalische Formen zu berücksichtigen. Diese Entwicklungen gehen für Anwender:innen mit enormen Vorteilen einher. Einerseits sind Transformermodelle per se performanter als neuronale Netze mit fixen Worteinbettungen, andererseits fällt mit der Textvorverarbeitung ein fehleranfälliger und arbeitsintensiver Prozessierungsschritt weitestgehend weg.

9.3 Datenanalyse

Hinsichtlich der NLP-Methoden lag der Fokus von Fallstudie 1 auf der überwachten Textklassifikation auf Wortebene. Das verwendete NER-Modell benötigte lediglich einige hundert Trainingsbeispiele, um generalisierbare Fähigkeiten zur Annotation von Impressumstexten zu erlernen. Somit konnte das Modell eingesetzt werden, um Unternehmensnamen und Adressen in den Impressumstexten zu erkennen und zu annotieren. Dieser Veredelungsschritt ist für viele geographische Web Scraping-Projekte höchst relevant. Für Geograph:innen ist das Internet eine bisher vernachlässigte Datenquelle für aufsetzende fachliche Analysen. Um Webinhalte in geographische Untersuchungen einbetten zu können, ist die räumliche Verortung von Webseiteninhalten eine notwendige Bedingung. Diese ist jedoch nicht automatisch gegeben, sondern muss erst durch den Forschenden vorgenommen werden, sodass dieser Arbeitsschritt geographische Webforschung bisher stark erschwerte. Insbesondere durch das in Deutschland geltende Telemediengesetz und der darin verankerten Impressumspflicht können Impressen von Webdomains als generischer Ansatz dienen, um Webseiten systematisch verorten zu können. Das verwendete NER-Modell konnte im Rahmen dieser Arbeit knapp 4 Millionen Domains mit Adressdaten versehen. Damit ist davon auszugehen, dass das vorliegende Modell eine zentrale Grundlage sämtlicher raumbezogener Web Scraping Vorhaben darstellen kann. NER-Modelle können für geographische Forschung auch losgelöst von Impressen eine enorme Relevanz aufweisen. So können auch in Fließtexten a priori definierte Entitäten erkannt und benannt werden, sodass beispielsweise im Rahmen von Dokumentenanalysen Akteure und Orte systematisch extrahiert werden können.

Die in Fallstudie 1 aus den Impressen extrahierten benannten Entitäten konnten ohne weitere Vorverarbeitung mithilfe eines Adressgeokodierungsverfahrens mit geographischen Koordinaten versehen werden. Hierbei ist anzumerken, dass derzeit nur wenige Möglichkeiten bestehen, Adressen in großem Umfang zu geokodieren. Die im Rahmen dieser Arbeit verwendete Variante

nutzt einen lokal installierten Nominatim-Server, um die Geokodierung vorzunehmen. Webdienste können für solche Vorhaben nicht genutzt werden, da diese in der Regel lediglich die Prozessierung einiger tausend Anfragen zulassen.

Fallstudie 2 demonstrierte ebenfalls ein Verfahren zur Textklassifikation. Während das NER-Modell aus Fallstudie 1 Wörter mit einem Label annotierte, kategorisierte das Klassifikationsmodell in Fallstudie 2 ganze Textpassagen. Die Klassifikation erfolgte in Fallstudie 2 in zwei Schritten. Zunächst wurden mittels des KeyBERT-Pakets (GROOTENDORST 2021) aus den jeweiligen Absätzen der Webseitentexte Stichwörter extrahiert. Diese sind standardisiert, berücksichtigen Bi- und Trigramme sowie die Semantik der Inputsequenz und können als Kurzzusammenfassung des Inhalts eines jeden Absatzes gesehen werden. Somit konnten die Absätze effizient auf Basis der Stichwörter gefiltert werden. Eine Stichwortsuche empfiehlt sich insbesondere dann, wenn erwartbar ist, dass lediglich ein Bruchteil der betrachteten Textpassagen für die Fragestellung relevante Inhalte enthält. Am Beispiel der Identifizierung von deutschen KI-Unternehmen konnte somit die Anzahl genauer zu betrachtender Textpassagen massiv reduziert werden. Gleichzeitig stellt die Auswahl geeigneter Stichwörter je nach Anwendungsfall eine Herausforderung dar. Die Hinzunahme respektive das Streichen einzelner Stichwörter kann sensitive Auswirkungen auf das Filterergebnis haben. Im Rahmen dieser Arbeit wurden in Fallstudie 2 vortrainierte Word-Embeddings verwendet, um die semantische Nachbarschaft einzelner Ausgangsbegriffe systematisch absuchen zu können. Dieses Vorgehen macht die Stichwortauswahl intersubjektiv nachvollziehbar und erleichtert es Forschenden möglichst neutral die finale Stichwortauswahl durchzuführen.

Nach abgeschlossener Stichwortsuche wurden die identifizierten Absätze nochmals feiner klassifiziert. Hierzu wurde ein vortrainiertes Transformermodell mittels Transfer Learning an die vorliegende Problemstellung angepasst. Da ein vortrainiertes Modell genutzt wurde, reichten für Training, Validierung und Evaluation insgesamt rund 1.000 Datenpunkte aus, um das Modell an die vorliegende Klassifikationsaufgabe anzupassen. Das verwendete Modell wurde mit Textdaten in über 100 Sprachen vortrainiert, sodass auch multilinguale NLP-Probleme gelöst werden können. Das Modell war somit in der Lage zu klassifizieren, ob in den gefilterten Absätzen ein direkter Bezug zur KI-Technologie vorliegt. Die empirischen Ergebnisse zeigen, dass die alleinige Existenz eines der gewählten Stichwörter noch kein ausreichender Indikator für die tatsächliche KI-Nutzung sind: Lediglich rund 38 % der mittels Stichwortsuche identifizierten Unternehmen wurden durch das Klassifikationsmodell als Unternehmen kategorisiert, welches in unmittelbarem KI-Bezug steht. Die Fähigkeit moderner Transformermodelle, Textpassagen anhand des Kontexts in vordefinierte Kategorien einzuteilen, kann im Rahmen geographischer Forschung zu unterschiedlichen Zwecken eingesetzt werden. Neben binären Klassifikationsproblemen können mittels des vorgestellten Verfahrens prinzipiell auch Mulit-Label-Klassifikationsprobleme gelöst werden.

Fallstudie 3 hingegen stellte im Gegensatz zu den beiden vorherigen Fallstudien ein unüberwachtes Verfahren zur Themenmodellierung in unstrukturierten Textdaten vor. Unüberwachte Verfahren gehen mit dem Vorteil einher, dass keine manuell annotierten Daten benötigt werden, um Texte Themen zuzuordnen. Unüberwachte Verfahren, wie das Topic Modeling, haben daher eher einen explorativen Charakter. Sie legen offen, welche Themen in einem Dokumentenkorpus enthalten sind, welche Dokumente welche Themen adressieren und wie ähnlich sich die jeweiligen Themen sind. Somit können Topic Models dabei helfen große Textmengen inhaltlich zu strukturieren.

Im Rahmen von Fallstudie 3 wurden Abstracts aus ausgewählten wirtschaftsgeographischen Journals analysiert. Durchgeführt wurde die Themenmodellierung mit dem Python-Paket BERTopic (GROOTENDORST 2022). Dieses nutzt ebenfalls vortrainierte Sprachmodelle, um Textpassagen kontextuell einzubetten. Das Paket ist darüber hinaus weitgehend individualisierbar. Beispielsweise können unterschiedliche Algorithmen zur Dimensionsreduktion oder zur Clusterbildung der eingebetteten Textsequenzen genutzt und kombiniert werden. Neben der Verwendung unterschiedlicher Algorithmik bietet BERTopic diverse Möglichkeiten, das Modell frage- bzw. zielstellungsabhängig zu parametrisieren. Beispielsweise kann a priori eine gewünschte Zahl an Themen definiert werden. Ferner können zu extrahierende Themen in Form von Stichwörtern vorgegeben werden und somit mittels überwachtem Topic Modeling in Textkorpora gesucht werden. Hierarchisches Topic Modeling ermöglicht es, Themen in Subgruppen zu ordnen und zu organisieren. Die empirischen Ergebnisse zeigen, dass das gewählte Modell durchaus präzise abstrakte Themen in Dokumenten benennen kann. Ferner konnten Themen im Zeitverlauf modelliert werden. Auf Basis dieses dynamischen Topic Modelings konnten Trends im wissenschaftlichen Diskurs identifiziert werden. Die Möglichkeit Themenbeschreibungen differenziert nach Zeitpunkt betrachten zu können, sorgt für tiefgreifendere Interpretationsmöglichkeiten und stützt die quantitative Betrachtung mit qualitativen Elementen.

Eine Herausforderung bei unüberwachten NLP-Verfahren ist es die Ergebnisse zu evaluieren. Dies geschieht in der Regel qualitativ durch die fortlaufende Bewertung durch den Forschenden im Analyseprozess. Quantitative Evaluationsansätze bestehen zwar. Allerdings hängt die finale Bewertung der Nützlichkeit eines Topic Models von der subjektiven Bewertung der Anwender:innen ab (DIENG et al. 2020; BIANCHI et al. 2021). Quantifizierende Ansätze untersuchen die Kohärenz bzw. Diversität der Themenbeschreibungen, um diese bewerten zu können.

Zusammenfassend lässt sich festhalten, dass moderne Datenanalysemethoden größerer Textmengen stark auf den Fähigkeiten vortrainierter Sprachmodelle beruhen. Diese erleichtern die Analyse massiv, da sie nahezu ohne Vorverarbeitungsschritte einsetzbar sind. Ferner ist zu konstatie-

ren, dass moderne NLP-Verfahren eine enorme Performanz aufweisen und sehr vielseitig einsetzbar sind. So wurden im Rahmen der Fallstudien Verfahren zur Klassifikation von Wörtern (NER-Modell), Absätzen (Text Klassifikation) und Dokumenten (Topic Modeling) verwendet. Es wurden Texte in unterschiedlichen Sprachen und aus unterschiedlichen Quellen verarbeitet sowie Quer- als auch Längsschnittsbetrachtungen durchgeführt. Die vorgestellten Analysemethoden umfassen, jedoch nur einen Teil der grundsätzlich verfügbaren NLP-Verfahren, sodass diese auch jenseits der skizzierten Anwendungsmöglichkeiten nutzbar sind. Darüber hinaus ist das Konzept des Transfer Learnings von bedeutender Relevanz für die Anwendung. Transfer Learning ermöglicht es mit geringem Initialisierungsaufwand komplexe NLP-Aufgaben zu lösen und das Sprachverständnis aufwändig vortrainierter Modelle zu nutzen.

9.4 Dateninterpretation

Bezüglich der Interpretation der analysierten Textdaten bestehen je nach Analysemethode unterschiedliche Möglichkeiten. In Fallstudie 1 wurden die gewonnenen Daten zunächst kartographisch visualisiert und anschließend regionale Disparitäten hinsichtlich der Webseitenabdeckung dargestellt. Darüber hinaus wurden die generierten Daten in Bezug auf geographische Repräsentativität mit offiziellen Unternehmensstatistiken verglichen und somit validert. Die kartographische Betrachtung regionaler Disparitäten hinsichtlich der Webseitenabdeckung von Unternehmen stellt somit einen gesonderten Indikator dar, welcher in weiteren Studien als Proxy zur Messung der regionalen Digitalisierungsintensität von Unternehmen genutzt werden kann.

Fallstudie 2 zeigte auf, wie web-generierten Daten genutzt werden können, um diese als Komplement innerhalb eines ökonometrischen Modells zu verwenden. Hierzu wurden die extrahierten Webdaten als abhängige Variable in ein lineares Regressionsmodel integriert. Die erzeugte Variable zur Messung der KI-Dichte in deutschen Kreisen und kreisfreien Städten wurde anschließend mit sekundärstatistischen Daten erklärt. Die Ergebnisse des Modells zeigen, dass insgesamt ein hoher Anteil der Varianz der abhängigen Variable KI-Dichte erklärt werden konnte. Eine besondere Relevanz für die Erklärung der Zielvariable hatten sozioökonomische Faktoren sowie Agglomerationsfaktoren. Somit zeigte sich, dass sich webgenerierte Daten gut als Komplement zu tradierten Datenquellen der empirischen Sozialforschung eignen. Aufgrund der Granularität der erzeugten Daten konnte an die Betrachtung auf Kreisebene eine feinkörnigere Analyse der Standortmuster von KI-Unternehmen durchgeführt werden. Auf Basis dieser Betrachtung können weitere detaillierte Standortbedingungen von KI-Unternehmen untersucht werden.

Die Interpretation der Ergebnisse von Fallstudie 3 erfolgte zweistufig. Zunächst wurden deskriptive Statistiken zur Themenzuordnung der Dokumente betrachtet. Darüber hinaus konnten die semantischen Verwandtschaften von Themen und Dokumenten quantifiziert werden. Auf Basis der deskriptiven Analyse konnten weitere Adaptionen des Modells vorgenommen und Themen

fusioniert werden. Im nächsten Schritt erfolgte die Interpretation der Ergebnisse auf Basis von Visualisierungen. Diese spielen eine zentrale Rolle bei der Bewertung und Interpretation von Themenmodellierungen. Sie schlagen eine Brücke zwischen quantitativen Beziehungen und dem qualitativen Verständnis des Forschenden. Zur Interpretation wurden beispielsweise Dendrogramme zur Bewertung von weiteren Clustermöglichkeiten betrachtet, die Entwicklung ausgewählter Themenkomplexe im Zeitverlauf analysiert und die lokalen Themenveränderungen eines ausgewählten Topics über Zeit beleuchtet. Darüber hinaus stellen interaktive Visualisierungen ein wichtiges Werkzeug bei der Exploration und Interpretation von Topic Modellen dar. Sie bieten die Möglichkeit, Themenzuordnungen bis auf Dokumentenebene nachzuvollziehen und zu evaluieren.

9.5 Limitationen

Insbesondere empirische Arbeiten, welche die Exploration neuartiger Methoden in den Fokus der Betrachtung stellen, weisen inhaltliche und methodische Limitationen auf. Diese sollen im folgenden Kapitel näher aufgezeigt und diskutiert werden. Die erste Limitation, welche Fallstudie 1 und 2 betrifft, stellt die verwendete Datengrundlage dar. Die im CC enthaltenen Domains wurden durch zufällige Crawls identifiziert. Ob und in welchem Umfang Unternehmensdomains nicht in dem betrachteten Datensatz enthalten sind, kann mangels passender Vergleichsdaten nicht abschließend geklärt werden. Ausgehend von dem Crawlingverhalten des CC-Crawlers kann davon ausgegangen werden, dass relevante Domains, die eine hohe Zentralität im Webgraphen aufweisen, inkludiert sind. Falls eine Unterdeckung von Unternehmensdomains besteht, ist diese daher eher bei kleineren bzw. neu gegründeten Unternehmen zu vermuten. Darüber hinaus betrachtete diese Arbeit lediglich deutschsprachige Domains. Es ist allerdings davon auszugehen, dass es ein Teil der deutschen Unternehmen fremdsprachige Domains betreibt. Diese sind ebenfalls nicht im untersuchten Datensatz enthalten und stellen eine weitere Unterdeckung dar. Um diesen Fehler feiner quantifizieren zu können, fehlen erneut passende Vergleichsdaten.

Darüber hinaus wurden im Rahmen dieser Arbeit Unternehmendomains anhand ihres Impressums identifiziert. Zwar besteht für deutsche Unternehmen laut § 5 Telemediengesetz eine Impressumspflicht. Allerdings ist davon auszugehen, dass einige Unternehmen entweder kein Impressum führen oder dieses nicht mittels der beschriebenen Heuristik identifizierbar ist. Beispielsweise können Impressumsinhalte, die als Bilddateien oder PDF-Dokumente gespeichert sind, nicht von dem eingesetzten Scraping Framework verarbeitet werden. Ferner wurde ein NER-Modell zur Annotierung der Adressdaten in den Impressumstexten verwendet. Dieses erreicht keine perfekte Vorhersagegenauigkeit, sondern lediglich Werte von knapp 90 %. Insbesondere die Entität *Unternehmensname* wird nicht in allen Fällen vollumfänglich korrekt erkannt. Allerdings bildet diese Entität eine Grundlage für die Identifizierung, da auf Basis der im Namen

enthaltenen Rechtsform Unternehmensdomains extrahiert wurden. Entsprechend ist hinsichtlich der Identifizierung mit einer gewissen Anzahl von falsch-negativen Fällen zu rechnen. Um diesen Fehler so gering wie möglich zu halten, wurde das Modell mehrfach auf den Datensatz angewandt. Neben der Identifizierung von Unternehmensdomains bildet die Geokodierung weiteres Fehlerpotential. Einerseits ist die Qualität der Geokodierung durch die Qualität der OSM-Daten limitiert, andererseits können Fehler in der Adressextraktion zu Folgefehlern bei der Geokodierung führen.

Neben potentiellen Fehlern bei der Adressextraktion und Geokodierung konnte im Rahmen dieser Arbeit keine tiefgreifenderen Validierungen der Unternehmensdatenbank durchgeführt werden. Somit bleibt ungeklärt, inwiefern die Abdeckung der Datenbank beispielsweise nach Unternehmensalter, Unternehmensgröße oder Branche variiert. Auf Basis von Sekundärstatistiken sowie vergleichbaren Forschungsarbeiten lassen sich erwartbare Verteilungsunterschiede zwar approximieren. Jedoch konnte mangels unternehmensscharfer Vergleichsdaten kein Vergleich mit der vorliegenden Datenbank durchgeführt werden.

In Fallstudie 2 wurden aufbauend auf den Ergebnissen von Fallstudie 1 aus der Gesamtheit der identifizierten Unternehmen, die Unternehmen extrahiert, welche KI in ihrem Geschäftsbetrieb verwenden. Die Klassifikation von KI-Unternehmen geht ebenfalls mit einigen methodischen Einschränkungen einher. Erstens wurden lediglich Webseiten betrachtet, zu denen ein Link ausgehend von der Startseite der Unternehmensdomain führt. Somit werden ggf. nicht alle Inhalte der Domain betrachtet, sondern lediglich Webseiten, die weiter oben in der Linkstruktur der Domain zu finden sind. Somit kann es bei der Klassifikation zu falsch-negativen Fällen kommen, falls die relevanten Webseiten nicht auf der Startseite der Domain verlinkt sind. Gleiches gilt für Texte, die im PDF-Format veröffentlicht wurden. Diese werden ebenfalls nicht betrachtet und können daher auch zu falsch-negativen Klassifikationen führen.

Zweitens stellt die vorgeschaltete Stichwortsuche zur Identifizierung von KI-Unternehmen eine weitere potentielle Limitation dar. Unterschiedliche Stichwortlisten können hier zu unterschiedlichen Ergebnissen führen. Zur Prüfung der Sensitivität wurden unterschiedliche Listen geprüft. Es zeigte sich, dass sich ein Großteil der Webseiten anhand der Überbegriffe „KI“, „AI“, „künstliche Intelligenz“ und „artificial intelligence“ identifizieren lassen. Somit ist insgesamt von einer gewissen Robustheit der Ergebnisse auszugehen, falls die genannten Überbegriffe in der Stichwortliste enthalten sind. KI-Unternehmen die keines der Stichworte auf ihren Webseiten verwenden, werden ebenfalls von der Analyse ignoriert und stellen eine weitere potentielle Fehlerquelle dar.

Drittens erreicht das verwendete Textklassifizierungsmodell ebenfalls keine perfekte Vorhersagegenauigkeit. Diese lag bei unterschiedlichen Testläufen relativ konstant bei rund 85 %. Da sich Sensitivität und Spezifität des Tests kaum unterscheiden, ist gleichermaßen von einigen wenigen falsch-negativen sowie falsch-positiven Fällen auszugehen. Darüber hinaus ist die Reliabilität der

Methodik von den Texten der Unternehmensdomains abhängig. Diese entsprechen nicht notwendigerweise wissenschaftlichen Standards. Vielmehr sind die Texte durch einen Selbstauskunftscharakter geprägt, sodass Abweichungen zwischen der Selbstdarstellung eines Unternehmens auf der Webseite und der Realität bestehen können. Ferner erfolgte die Annotation der Trainingsdaten durch lediglich einen Forscher, sodass die Intercoderreliabilität eingeschränkt ist.

Auch Fallstudie 3 weist Limitationen hinsichtlich der Datengrundlage auf. Zur Selektion relevanter Journals respektive Artikel einer Wissenschaftsdisziplin können unterschiedliche Methoden und Heuristiken zum Einsatz kommen. Ferner bietet das Verfahren des Topic Modelings fragestellungsabhängig unterschiedliche Möglichkeiten, Themen feiner aufzuschlüsseln oder zu fusionieren. Die Evaluation der Eignung der Themengruppen und –beschreibungen ist somit stark von dem Ziel der Untersuchung abhängig. Entsprechend wären auf Basis des vorliegenden Topic Models weitere Adaptionen denkbar, die gegebenenfalls abweichende Ergebnisse produzieren könnten. Darüber hinaus kann es bei der Replikation der vorliegenden Themenmodellierung zu leichten Abweichungen kommen. Dies ist mit den stochastischen Elementen des BERTopic-Algorithmus zu begründen, sodass eine exakte Replikation nur durch Unterdrückung stochastischer Elemente zu Lasten der Leistungsfähigkeit erreichbar ist (GROOTENDORST 2023).

10 Integration der untersuchten Methoden in die empirische Sozialforschung

Ausgangspunkt dieser Arbeit war es zu prüfen, inwiefern die Methoden des Web Minings und NLP einen Mehrwert für Forschungsvorhaben in der Wirtschaftsgeographie darstellen können. Aus methodischer Perspektive bietet Web Mining und NLP eine Vielzahl an Anwendungsmöglichkeiten, sodass die Methoden keiner etablierten Methodik der empirischen Sozialforschung zugeordnet werden können. Zwar nutzen die vorgestellten Verfahren ähnlich der qualitativen Forschung Text als Datenquelle. Allerdings in deutlich größeren Fallzahlen als es in tradierten qualitativen Forschungsdesigns gängig ist. Quantitative Methodik ist zwar auf die Analyse großer Fallzahlen ausgelegt. Jedoch fokussieren sich die Auswertungsmethoden auf die statistische Analyse numerischer Daten. Während Stärken und Schwächen qualitativer und quantitativer Methodik in Lehrbüchern und Fachbeiträgen hinlänglich diskutiert sind, ist bis dato noch ungeklärt, inwiefern sich Text Mining insbesondere im Kontext der Wirtschaftsgeographie von den beiden etablierten Ansätzen abgrenzen lässt.

10.1 Erkenntnistheoretische Konzeption

Grundlegend unterscheiden sich qualitative und quantitative Forschungsdesigns bereits hinsichtlich ihrer wissenschaftstheoretischen Fundamente. Während quantitative Forschung der aus den Naturwissenschaften stammenden Denkrichtung des Positivismus folgt, ist qualitative Forschung eher dem philosophischen Konstruktivismus zugeordnet. Quantitative Forschung folgt demnach einer deduktiven Denklogik. Auf Basis theoretisch-konzeptioneller Überlegungen werden Erklärungsansätze respektive Hypothesen formuliert, welche anschließend auf Basis des empirischen Materials geprüft werden. Qualitativen Forschungsdesigns sind hingegen induktive Logiken inhärent. Hypothesen und Erklärungsansätze entstammen nicht theoretischen Annahmen oder Modellen, sondern entstehen aus Einzelfallbeobachtungen heraus (BAUR und BLASIUS 2022; HÄDER 2010).

Die methodischen Vorgehensweisen der Fallstudien dieser Arbeit sind sowohl deduktiv als auch induktiv geleitet. Fallstudie 1 nutzte als Kerninstrument ein NER-Modell zur Annotation von Adressdaten. Überwachte Verfahren der NER suchen in Texten nach a priori fest definierten Entitäten. Es findet zwar keine Hypothesenprüfung statt. Dennoch kann die grundlegende Strategie als deduktiv orientiert eingeordnet werden. In Fallstudie 2 wurde ebenfalls ein überwachtes Lernverfahren zur Klassifizierung von Textsequenzen angewandt. Diese Textsequenzen wurden hinsichtlich fest definierter Kategorien eingeordnet und somit ebenfalls deduktiv verarbeitet. Fallstudie 3 hingegen verwendete mit dem Verfahren des Topic Modelings ein unüberwachtes Ver-

fahren, welches einen induktiv-explorierenden Charakter aufweist. Hypothesen bzw. Erklärungsmuster entstehen in Fallstudie 3 induktiv aus dem empirischen Material heraus. Vorannahmen spielen bei der methodischen Umsetzung lediglich eine untergeordnete Rolle. Vielmehr helfen unüberwachte Verfahren dabei große Textmengen inhaltlich zu strukturieren und einen Überblick über im Textkorpus enthaltene Themen und deren Zusammenhänge zu geben. Insgesamt zielen Text Mining-Verfahren also auf die Entdeckung latenter semantischer Strukturen in Texten ab. Unter dem Überbegriff sammeln sich sowohl deduktiv orientierte als auch induktiv geprägte Analyseverfahren. Entsprechend kann Text Mining allgemein keiner fixen Denkrichtung zugeordnet werden. In wirtschaftsgeographischer Forschung kommen induktiv und deduktiv orientierte Verfahren gleichermaßen zum Einsatz. Klassischerweise dienen induktive Verfahren, die auf qualitativen Untersuchungen weniger Fälle basieren zur Theoriebildung. Zur Prüfung bzw. zur Makrofundierung bestehender Theorien nutzen Wirschaftsgeograph:innen deduktiv-quantitative Analysemethoden. Die Möglichkeiten, die sich aus der Integration von Text Mining in die wirtschaftsgeographische Forschungspraxis ergeben, werden in Kapitel 10.5 nochmals expliziter beleuchtet.

10.2 Datengrundlagen, Untersuchungsumfänge und Forschungsprozess

Darüber hinaus unterscheiden sich qualitative und quantitative Verfahren hinsichtlich der Untersuchungsumfänge. Quantitative Verfahren zielen auf große Fallzahlen ab, um generalisierbare Aussagen treffen zu können. Qualitative Forschungsdesigns hingegen untersuchen Einzelfälle, die mittels theoretischem Sampling ausgewählt werden. Verfahren des Text Minings lassen sich bezüglich dieser Dimension den quantitativen Ansätzen zuordnen. Das Forschungsfeld hat sich explizit zur Analyse großer Datenmengen entwickelt.

Die Datengrundlagen qualitativer und quantitativer Methoden unterscheiden sich ebenfalls. Quantitative Forschung nutzt standardisiertes, numerisches Datenmaterial, welches mittels statistischer Analyseverfahren ausgewertet wird. Das Datenmaterial qualitativer Methodik ist im Vergleich deutlich unstrukturierter. Es handelt sich hierbei häufig um unstrukturiertes Text-, Ton- oder Bildmaterial, welches interpretativ ausgewertet wird. Methoden des Text Minings sind bezüglich der Datenerhebungs- und Auswertungsmethoden keinem Paradigma eindeutig zuordenbar. Einerseits nutzen Verfahren des Text Minings – ähnlich der qualitativen Forschung - unstrukturierte Textdaten als Datenmaterial. Andererseits handelt es sich nicht, um manuelle interpretative Auswertungsmethoden, sondern um computerbasierte und daher systematische Verfahren. Damit weist Text Mining Überschneidungen mit quantitativen Inhaltsanalysen auf. Jedoch ermöglicht moderne Algorithmik deutlich umfassendere semantische Analysen als quantitative Textanalysen, welche Texte nur sehr grob z.B. anhand von Wortzählungen beschreiben.

Aus Datenmaterial und Untersuchungsumfang ergibt sich ein weiterer zentraler Unterschied zwischen qualitativer und quantitativer Forschung. Qualitative Verfahren können die Einzelfälle sehr

detailliert und umfassend untersuchen. Daraus ergibt sich die Möglichkeit auch latente, schwer formalisierbare Sachverhalte wie Stimmungen, Einstellungen, Werte oder Denkweisen zu untersuchen. Im Vergleich zu quantitativen Ansätzen sind qualitative Forschungsergebnisse in der Regel hinsichtlich ihrer Reichweite und Generalisierbarkeit limitiert, weisen jedoch eine größere Tiefenschärfe auf als quantitative Forschungsarbeiten.

Quantitatives Datenmaterial liegt meist geographisch bzw. inhaltlich aggregiert vor, sodass kein detailliertes Verständnis zu Kontexten oder Hintergründen geschaffen werden kann. Text Mining ist auch hinsichtlich dieser Dimension keinem Wissenschaftsparadigma präzise zuordenbar. Vor dem Hintergrund einer steigenden Verfügbarkeit von Textdaten – auch zu Einzelfällen – können mittels Text Mining unscharfe bis dato schwer quantifizierbare Sachverhalte auf Mikroebene analysiert werden. Text Mining-Verfahren erreichen bezüglich dieser qualitativen Dimension nicht die Tiefenschärfe von klassischer qualitativer Forschung, da computerbasierte Verfahren bis dato nicht in der Lage sind Textelemente in größere politische, gesellschaftliche oder soziale Sinnzusammenhänge zu bringen bzw. diese zu interpretieren. Gleichzeitig ermöglichen die computergestützten Textanalysen die Betrachtung deutlich höherer Fallzahlen im Vergleich zu qualitativen Forschungsdesigns.

Auch hinsichtlich des Forschungsablaufs bestehen markante Unterschiede zwischen den beiden klassischen Analyseverfahren der empirischen Sozialforschung. Quantitative Ansätze verfolgen einen starren, linearen und geschlossenen Ablauf, während qualitative Verfahren als offen, zirkulär und flexibel beschrieben werden können. Zentrales methodisches Element von Text Mining-Verfahren bildet das Zusammenspiel aus großen Datenmengen und KI-Modellen, welche versuchen Gesetzmäßigkeiten, Muster und Zusammenhänge aus den hochdimensionialen Massendaten abzuleiten. Der Trainings-, Validierungs- und Evaluationsprozess besteht dabei aus einer permanenten Adaptierung verschiedener Modellgattungen auf die vorliegenden Daten. Der Prozess weißt somit einen eher zirkulären Charakter auf, da die Modellspezifikation permanent unter Rückbezug auf die Daten erfolgt.

10.3 Gütekriterien und Qualitätssicherung

Quantitative und qualitative Methoden der empirischen Sozialforschung unterscheiden sich darüber hinaus hinsichtlich ihrer Gütekriterien. Die Gütekriterien quantitativer Forschung sind unter den Begrifflichkeiten Reliabilität, Validität und Objekivität fest definiert (HÄDER 2010). Gütekriterien qualitativer Forschung werden seit vielen Jahren kontrovers diskutiert, sodass unterschiedliche Denkrichtungen verschiedene Kriteriensets hervorgebracht haben. STEINKE (2007) formuliert vier allgemeine Kriterien zur Qualitätssicherung in der qualitativen Sozialforschung: die Indikation der methodischen Vorgehensweise, die empirische Verankerung der gewonnenen Theorie in den Daten, das Aufzeigen der Verallgemeinerbarkeit der Ergebnisse und die Herstellung

intersubjektiver Nachvollziehbarkeit. Text Mining-Verfahren hingegen basieren hauptsächlich auf ML- sowie KI-Verfahren. In der Informatik bestehen verschiedene Verfahren zur Modellselektion, -abstimmung und –evaluation, die eine umfassende Qualitätssicherung der trainierten Modelle ermöglichen. Beispielhaft zu nennen sind die Evaluationsmetriken Precision, Recall, Accuracy und Fehlerrate sowie Verfahren zur Evaluierung der Robustheit von Modellen wie zum Beispiel Kreuzvalidierung (RASCHKA 2020).

Wie die empirischen Ergebnisse dieser Arbeit zeigen, ist die Datenprozessierung und –verarbeitung großer Textmengen unmittelbar mit der Nutzung von KI-Modellen verbunden. Wie in Kapitel 4 beschrieben sind moderne KI-Modelle in der Lage, hochkomplexe und nicht-lineare Muster und Beziehungen in großen Datenmengen zu erlernen. Die Komplexität dieser Modelle übersteigt dabei bei weitem die kognitive Kapazität von Menschen. Entsprechend können Menschen gegebenenfalls einzelne Rechenschritte eines komplexen Algorithmus nachvollziehen. Allerdings ist es quasi unmöglich zu verstehen auf Basis welcher Merkmale der Inputdaten eine Prognose entsteht. Somit werden moderne KI-Modelle häufig als Black-Box beschrieben (CARABANTES 2020; CASTELVECCHI 2016).

Mittels Test- und Validierungsdaten kann die Prognoseleistung von KI-Modellen zwar umfassend evaluiert, jedoch keine Aussagen über das Zustandekommen des Ergebnisses getroffen werden. Diese mangelnde Interpretierbarkeit und intersubjektive Nachvollziehbarkeit stellt aus erkenntnistheoretischer Perspektive eine Unzulänglichkeit dar. Abstrakte Muster und latente Zusammenhänge können im Ergebnis zwar zu guten Vorhersageleistungen führen. Allerdings stellen sie keine kausal begründete Entscheidungsfindung dar. Folglich kann die Intransparenz von KI-Modellen als größte Herausforderung bei der Integration moderner Algorithmik in (sozial)wissenschaftliche Forschungsmethoden betrachtet werden. Bestehende Gütekriterien zur Qualitätssicherung von ML-Modellen sind stark technisch orientiert. Entsprechend wird derzeit im ML-Umfeld intensiv an Verfahren geforscht, die exakt die beschriebene Black-Box Problematik im Zusammenhang mit ML-Systemen adressieren. Diese werden im folgenden Kapitel nochmals detaillierter beschrieben.

10.4 Zusammenfassende Einordnung

Zusammenfassend lässt sich festhalten, dass die untersuchte Methode des Text Minings hinsichtlich der betrachteten Dimensionen keinem Paradigma der empirischen Sozialforschung eindeutig zuzuordnen ist. Tabelle 17 fasst die zentralen Gemeinsamkeiten sowie Unterschiede der drei verglichenen Methodiken abschließend zusammen.

Tabelle 17: Unterschiede zwischen tradierten Forschungsdesigns sowe Text Mining.

Dimension	**Qualitative Methodik**	**Quantitative Methodik**	**Text Mining**
Erkenntnistheorie	Induktion	Deduktion	Beides möglich
Datengrundlage	Unstrukturiert	Strukturiert	Unstrukturiert
Untersuchungsumfang	Einzelfälle	Hohe Fallzahlen (zumeist aggregiert)	Hohe Fallzahlen (Einzelfälle)
Forschungsprozess	Zirkulär, offen, flexibel	Linear, geschlossen, strikt	Zirkulär
Gütekriterien	nachvollziehbar, theoretisch verankert, Generalisierbarkeit	Reliabilität, Validität, Objektivität	Metriken zur Modellevaluation

Quelle: Eigene Darstellung.

Somit stellt die Methodik des Text Minings eine dritte Säule dar. Diese zeichnet sich dadurch aus, dass sie detaillierte und der Forschungsfrage angepasste Daten generieren kann. Dies wird durch das Kategorisieren von Einzelbeobachtungen erreicht. Somit sind die Daten auch geographisch hoch aufgelöst. Diese Eigenschaften werden in der empirischen Sozialforschung mit qualitativen Forschungsdesigns verbunden. Während qualitative Forschung jedoch auf die umfassende Analyse ausgewählter Einzelbeobachtungen abzielt, ermöglichen es Verfahren des NLP deutlich höhere Fallzahlen zu verarbeiten, die je nach Untersuchungsgegenstand einer Vollerhebung nahekommen oder gar entsprechen. Hinsichtlich der Stichprobengrößen bestehen also größere Schnittmengen zwischen Text Mining und der quantitativen Forschung. Allerdings untersuchen quantitative Studien Einzelbeobachtungen nur im Rahmen standardisierter Primärerhebungen. Diese weisen jedoch hinsichtlich der Stichprobengröße Limitierungen auf. Einerseits erreichen Rücklaufquoten selten Werte über 50 % (SHIH und FAN 2009; STEDMAN et al. 2019). Außerdem beobachten STEDMAN et al. (2019) eine Abnahme der Rücklaufquoten von 0,76 % pro Jahr. Ferner leiden insbesondere Stichproben von Onlineumfragen unter Verzerrungen durch Selbstselektion und -verständnis (BETHLEHEM 2010; GREENACRE 2016). Andere Datenquellen der quantitativen Forschung sind häufig sektoral, technologisch oder geographisch hoch aggregiert, sodass präzise Aussagen über individuelle Effekte auf Mikroebene nicht möglich sind.

Obgleich der Potentiale von Text Mining als methodisches Komplement in der Wirtschaftsgeographie entstehen aus den empirischen Ergebnissen dieser Arbeit Fragen nach methodischen Standards für diese Forschungspraxis. Während qualitative und quantitative Methodiken seit Jahrzehnten etablierte und differenziert ausgestaltete Forschungzugänge darstellen, ist Text Mining noch in einem explorativen Stadium. Entsprechend bedarf es noch weiterer Untersuchungen, welche die Methodik kritisch evaluieren und bestehende Herausforderungen adressieren.

Im Kontext groß angelegter Web Mining-Vorhaben sollten diese bereits bei der Datengrundlage ansetzen. Das in dieser Arbeit betrachtete CC-Projekt kann durchaus als umfassende Datengrundlage zur Identifizierung von Unternehmensdomains herangezogen werden. Inwieweit das CC für

die Identifizierung anderer akteursbezogener Webseiten geeignet ist, gilt es in weiteren Forschungsprojekten zu klären. Um Web Mining als standardisierten Datenzugang in der empirischen Sozialforschung zu etablieren bedarf es folglich weiterer Forschung zu Aufbau und Struktur des Internets. Aus diesen Arbeiten lassen sich Annahmen ableiten, die zu einem besseren Verständnis der Grundgesamtheit der im Internet enthaltenen Domains beitragen können. Entsprechend werden freie, standardisierte und barrierearme Datenzugänge benötigt, welche gemeinschaftliche Forschung und darüber hinaus einen wissenschaftlichen Diskurs ermöglichen.

Aus einer intensiveren wissenschaftlichen Auseinandersetzung heraus lassen sich weitere Best Practices hinsichtlich des Datenbezugs ableiten. An dieser Stelle gilt es weitere Evidenz zu Struktur und Umfang von Webseiten zu sammeln und zu konsolidieren. Insbesondere zur Frage, in welchem Umfang und in welcher Tiefe Texte von Webseiten abgefragt werden müssen, um ein repräsentatives Bild der zugrundeliegenden Domain zu erlangen, bedarf es weiterer Studien. Auf Basis der empirischen Ergebnisse dieser Arbeit sowie verwandter Studien (KINNE und AXENBECK 2020) lässt sich die Annahme formulieren, dass bereits wenige Webseiten einer Domain ausreichen, um deren Profil und Inhalte anhand der Texte schätzen zu können. Inwiefern diese Erkenntnisse, die sich auf Unternehmensdomains beziehen, übertragbar sind, stellt ein weiteres Forschungsdesiderat dar.

Ebenso gilt es Validität und Objektivität von Webseitentexten besser zu verstehen. Wie bereits in Kapitel 9.5 erläutert sind Webseitentexte nicht nach wissenschaftlichen Standards erhoben oder formuliert. Sie können entsprechend durch Selbstdarstellung der Akteure in ihrer Objektivität und Validität eingeschränkt sein. Um Sentiment und Inhalt von Webseitentexten besser einordnen zu können, sind neben statistischen Validierungen auf Basis von Vergleichsdaten insbesondere qualitative Studien nötig. Im Rahmen von qualitativen Untersuchungen gilt es zu prüfen, inwiefern Themen, Charakteristika und Wahrnehmungen, die Webseitentexten entnommen wurden, mit Ergebnissen qualitativer Studien übereinstimmen. Ferner lassen sich aus qualitativen Validierungsstudien heraus Limitationen und Implikationen der Datenquelle Webseite einordnen.

Auch hinsichtlich der Datenverarbeitung und –analyse gilt es Standards und Praktiken zu etablieren, die eine Integration von Text Mining in das sozialwissenschaftliche Methodeinstrumentarium ermöglichen. Entsprechend ist es zukünftig notwendig, den Prozess der Entscheidungsfindung komplexer KI-Modelle für den Menschen nachvollziehbar zu machen. Unter dem Stichwort *erklärbare KI* (engl. explainable AI) werden bereits erste Ansätze diskutiert (ARRIETA et al. 2020). Nach ARRIETA et al. (2020) kann erklärbare KI erstens dazu beitragen sicherzustellen, dass im Prozess der Entscheidungsfindung lediglich im jeweiligen Kontext sinnvolle Variablen betrachtet werden. Zweitens helfen sie dabei Verzerrungen in den Trainingsdaten aufzudecken und auszugleichen. Drittens erhöht sie die Robustheit der Modelle, da sensitive Variablen aufgedeckt werden können.

Im Bereich der maschinellen Textverarbeitung kann beispielsweise mittels Salienzverfahren visualisiert werden, welche Token der Inputsequenz eine besondere Relevanz zur Vorhersage des Outputs haben (TENNEY et al. 2020; WALLACE et al. 2019)

Neben technischen Aspekten hinsichtlich der Interpretierbarkeit und Robustheit von KI-Modellen, gilt es weitere Standards bezüglich der methodischen Vorgehensweise zu definieren. Zwar benötigen vortrainierte Transformermodelle immer weniger manuell annotierte Trainingsdaten. Jedoch sollte der Annotierungsprozess als kritisches Element der Modellabstimmung sorgfältig dokumentiert und reflektiert werden. An dieser Stelle bedarf es weiterer methodischer Richtlinien, welche den Kodierprozess rahmen. Denkbar wäre es beispielsweise Forschungspraktiken aus der qualitativen Inhaltsanalyse aufzugreifen, die mittels Kodierleitfäden Kategorien, Definitionen, Ankerbeispiele und Kodierregeln dokumentieren (MAYRING 2010). Gleiches gilt ferner für die Parametrisierung und Auswahl von Algorithmen und Modellen. Hier bedarf es ebenfalls weiterer methodischer Standards zur Dokumentation des Analyseprozesses, um eine Reproduzierbarkeit der Ergebnisse zu gewährleisten.

10.5 Beispiele integrativer Forschungsdesigns

Aufgrund der verschiedenen Charakteristika der drei vorgestellten methodischen Zugänge ermöglicht deren Kombination neue Forschungsdesigns. Die Verbindung qualitativer und quantitativer Analysedimensionen in einem Forschungsdesign ist bereits hinlänglich unter dem Überbegriff Mixed-Methods-Ansatz bekannt (vgl. Kapitel 2.2). Durch die Einführung eines dritten methodischen Zugangs ergeben sich demnach neue Methodenkombinationen in der Mixed-Methods-Forschung.

Die Potentiale einer Verbindung von Text Mining-Verfahren mit qualitativer Forschung wurden zwar in Kapitel 2.3 dieser Arbeit bereits kurz angeschnitten, sollen im Folgenden jedoch nochmals unter Reflektion der empirischen Analysen beleuchtet werden. Text Mining kann insbesondere bei der Fallauswahl für tiefgreifendere qualitative Auseinandersetzungen eingesetzt werden. Sowohl induktive, unüberwachte Verfahren (z.B. Topic Modeling) als auch deduktive, überwachte Ansätze (z.B. Textklassifikation) bieten die Möglichkeit umfangreiches Textmaterial zu strukturieren. Speziell unüberwachtes Topic Modeling ermöglicht es zu Beginn eines Forschungsprojekts einen Untersuchungsgegenstand (in Form von Textmaterial) besser zu verstehen. Angesichts stetig wachsender digitaler Textrepositorien kann diesen Verfahren eine besondere Relevanz zugeschrieben werden. Auf Basis der computergestützten Strukturierung der Themen(zusammenhänge) kann anschließend eine gezielte Fallauswahl für die qualitative Analyse folgen. Anknüpfend an Fallstudie 3 könnten Autor:innen von Artikeln beispielsweise bestimmter Themengrup-

pen nochmals gezielter befragt werden. Außerdem könnten im Rahmen von Fokusgruppendiskussionen und Expert:innengesprächen die quantitativ explorierten Strukturen und Trends qualitativ evaluiert werden.

Außerdem können mit dynamischen Topic Modellen Veränderungsprozesse in umfangreichem Textmaterial automatisiert quantifiziert werden. Somit können enorm große Textbestände auch über lange Zeiträume hinweg analysiert und der Bedeutungszuwachs respektive die Bedeutungsabnahme bestimmter Themen im Zeitverlauf quantitativ gefasst werden. Diese Daten können sowohl singulär analysiert als auch als zusätzliche Variable in quantitative und qualitative Forschungsdesigns eingebunden werden.

Gilt es einen umfangreichen Textkorpus im Vorfeld einer qualitativen Untersuchung gezielt nach bestimmten Themen zu filtern, können überwachte Klassifikationsverfahren eingesetzt werden. Somit lassen sich aus einer unstrukturierten Textsammlung gezielt relevante Dokumente extrahieren. Aus dieser Teilmenge können dann erneut gezielt Fälle für tiefere Analysen identifiziert werden. Am Beispiel von Fallstudie 2 könnten z.B. identifizierte KI-Akteure und Multiplikatoren regionaler Innovationssysteme befragt werden, um ein bessereres Verständnis für zugrundeliegenden Mechanismen einer fortschrittlichen regionalen KI-Entwicklung zu erlangen. Die analysierten Mikrodaten sind an dieser Stelle von besonderer Relevanz, um betreffende Akteure effizient identifizieren und die quantitativen Ergebnisse adäquat mit qualitativen Studien synthetisieren zu können.

Insgesamt ermöglichen Text Mining-Bausteine in Kombination mit qualitativen Elementen Forschenden eine Art „Vogelperspektive auf große Textkorpora einzunehmen“ (DUMM und NIEKLER 2016: 91). MORETTI (2000) fasst diese Verknüpfung unter den Begriffen *close* und *distant reading* zusammen, wobei *close reading* einer qualitativen Inhaltsanalyse gleichkommt und *distant reading* die computerbasierte großskalige Analyse darstellt (vgl. Kapitel 2.3). Die Möglichkeit, während des Forschungsprozesses permanent zwischen diesen Lesemodi wechseln zu können, sehen DUMM und NIEKLER (2016) als den zentralen Vorteil dieses kombinierenden Forschungsdesigns an.

Gleichwohl besteht die Möglichkeit Text Mining in quantitativ orientierte Forschungsdesigns zu integrieren. Für quantitative Forschungsdesigns können aus großen Textkorpora neue Variablen erzeugt werden, die in bestehenden Datensätzen nicht oder nur in aggregierter Form enthalten sind. Hierfür können sowohl überwachte als auch unüberwachte Verfahren eingesetzt werden. Allerdings sind überwachte Verfahren aufgrund des stark linearen und deduktiven Charakters quantitativer Forschung geeigneter, um eine stringente Operationalisierung von Hypothesen zu gewährleisten. Beispielhaft wurde dies in dieser Arbeit anhand der ersten beiden Fallstudien illustriert. Diese zeigen auf, wie mittels Text Mining umfangreiche und sowohl inhaltlich als auch

räumlich granulare Datenzugänge geschaffen werden können. Speziell in der Wirtschaftsgeographie spielt die räumliche Maßstabsebene sowie die Übersetzbarkeit von Hypothesen in messbare Indikatoren eine zentrale Rolle. Die mangelnde Verfügbarkeit entsprechender Daten stellt häufig eine Limitierung der empirischen Operationalisierbarkeit von theoretisch abgeleiteten Annahmen und Modellen dar. Fallstudie 1 illustrierte exemplarisch wie entsprechende Datenzugänge geschaffen werden können, die exakt die genannten Limitationen bestehender Ansätze adressieren. Fallstudie 2 skizzierte anschließend wie Variablen aus Webtexten abgeleitet und in ein klassisches ökonometrisches Modell integriert werden können.

Somit kann die Integration von Text Mining Elementen in eine quantitativ ausgerichtete Forschungsarchitektur eine wichtige Ergänzung des Indikatorensets darstellen. Für die Spezifikation ökonometrischer Modelle können so bestehende Datensätze, z.B. Sekundärstatistiken, Unternehmensbefragungen oder Patentdaten, mit Text Mining ergänzt werden. Solche neuen Indikatoren bieten einen besonderen Nutzen für bis dato schwer quantifizierbare Hypothesen. Speziell in der Wirtschaftsgeographie bilden unscharfe Interaktionen zwischen Akteuren im Raum das Fundament vieler etablierter Theoriekonstrukte. Mittels Text Mining kann versucht werden diese weichen und unscharfen Konzepte quantitativ zu fassen und zu erklären. Damit eröffnen sich neue Möglichkeiten einerseits zur Mikrofundierung und Erweiterung bestehender Theorien sowie andererseits zur Formulierung neuer Erklärungsansätze.

11 Handlungsempfehlungen

Aus der Synthese der Empirie und unter Einbezug der theoretisch-konzeptionellen Überlegungen sollen folgend Handlungsempfehlungen für die Praxis abgeleitet werden. Aufgrund des methodenexplorierenden Charakters der Arbeit, richten sich diese insbesondere an (wirtschafts)geographische Forschung und Lehre sowie übergeordnet an die Hochschulpolitik.

11.1 Handlungsempfehlungen für die wirtschaftsgeographische Forschung

Aus den empirischen Ergebnissen dieser Arbeit lässt sich ableiten, dass Webdaten im Allgemeinen und Webtextdaten im Speziellen bedeutende Potentiale für wirtschaftsgeographische Forschung bereithalten. Entsprechend sollten sich Wirtschaftsgeograph:innen zukünftig verstärkt mit den Potentialen dieser neuartigen Methoden für die eigene Forschung auseinandersetzen. Geograph:innen sind auf räumlich fein aufgelöste Daten angewiesen, um theoretische Modellannahmen prüfen und erweitern bzw. aus gesammelten Datenmaterial heraus neue Erklärungsansätze ableiten zu können. Bestehenden quantitativen Daten fehlt es häufig an kontextualisierenden Elementen, die differenzierte Interpretationen statistischer Muster und Regelmäßigkeiten erleichtern. Fehlendes bzw. unzureichendes Datenmaterial speziell auf Mikroebene sorgt in der Wirtschaftsgeographie für eine tiefe Kluft zwischen quantitativer und qualitativer Forschung (Bathelt und Li 2020).

Aufgrund der weit fortgeschrittenen und weiter zunehmenden Digitalität von Textpublikationen stellt digitaler Text bereits heute eine zentrale Wissensressource dar, die neue Untersuchungsmöglichkeiten schafft. Auf methodischer Seite fehlt Wirtschaftsgeograph:innen bisweilen das notwendige Werkzeug, um digitalen Text sowohl in großen Mengen als auch semantisch umfänglich „abzubauen“. Neben der Prüfung der Potentiale des Text Minings für die eigene Forschung sind folglich methodische Weiterbildungen notwendig. Über die zentralen Werkzeuge des NLP und Web Minings hinaus sind Kenntnisse im Bereich Data-Engineering sowie allgemein solide Programmierkenntnisse notwendig, um Text Mining in der eigenen Forschung anwenden zu können. Gleichwohl bestehen zu diesen Themengebieten vielfältige, teilweise kostenfreie Lernangebote im Internet. Diese ermöglichen einen schnellen Einstieg in die wichtigsten Verfahren und Konzepte. Auf diesem Grundverständnis aufbauend kann dann die Entwicklung speziell angepasster Algorithmik für die eigene Forschung erfolgen. Darüber hinaus kann interdisziplinäre Forschung ein Vehikel sein, um Textdaten aus dem Internet in wirtschaftsgeographische Forschungsprogramme einfließen zu lassen. Beispielsweise könnte gemeinschaftliche Forschung von Wirtschaftsgeograph:innen mit Informatiker:innen respektive Vertreter:innen der Computerlinguistik die Möglichkeit bieten, technischen und fachliche Expertise zu vereinen. Insbesondere in der (Wirtschafts)geographie fehlen hier bisweilen nützliche Tools, die speziell für raumbezogene For-

schung notwendig sind, sodass es weiterer technisch orientierter Forschung in der Wirtschaftsgeographie bedarf, um die Potentiale von Textdaten auch weniger technisch-affinen Wirtschaftsgeograph:innen zugänglich zu machen. Ferner sollten weitere integrative Forschungsdesigns exploriert werden, um ein besseres Verständnis zu erhalten, inwiefern sich Text Mining in die empirische wirtschaftsgeographische Forschung einfügen kann.

11.2 Handlungsempfehlungen für die geographische Methodenausbildung

Da die bisherige Methodenausbildung der geographischen Curricula kaum Grundlagen schafft, die einen barrierearmen Einstieg in das Themengebiet ermöglichen, können ebenso Handlungsempfehlungen für die Methodenausbildung gegeben werden. Das folgende Kapitel beschäftigt sich dediziert mit den notwendigen Kompetenzen, die eine zentrale Grundlage für die eigenständige Anwendung der Methodik darstellen.

Angesichts der enormen Menge digital verfügbarer sozialwissenschaftlich relevanter Informationen bedarf es einer Erweiterung der geographischen Methodenausbildung. Derzeit besteht eine sich weitende Kluft zwischen Potentialen von Big Data und den Kompetenzen von Geograph:innen, diese wissenschaftlich in Wert setzen zu können. Um diese Lücke zu schließen, existieren bereits seit einigen Jahren Vorschläge zur Anpassung sozialwissenschaftlicher Lehre (MUNZERT 2014, 2018). Grundsätzlich schließt sich dieser Beitrag den bestehenden Forderungen an, welche insbesondere eine fundierte Softwareausbildung in den Sozialwissenschaften verlangen. Einerseits sind Programmierkenntnisse die notwendige Bedingung für die Analyse großer unstrukturierter Datenmengen. Andererseits stellen sie für Absolvent:innen eine wichtige Kompetenz auf dem Arbeitsmarkt dar.

Neben einer Softwareausbildung sind weitere Kompetenzen vonnöten, um souverän und eigenständig mit neuen Daten umgehen zu können. Da ein Großteil der relevanten Daten im Internet gespeichert ist, benötigen Studierende ebenso grundsätzliches Wissen über zentrale Webtechnologien wie HTML, XML, JSON oder APIs (vgl. Kapitel 3.1.1). Für ambitionierte Projekte sind außerdem Kompetenzen im Bereich der Verarbeitung und –speicherung großer Datenmengen notwendig. Bedeutende Konzepte sind unter anderem Datenbanksysteme zur effizienten Speicherung großer Datenmengen sowie Parallelisierungsalgorithmik zur beschleunigten Verarbeitung. Grundlegende Fähigkeiten in diesen Bereichen legen ein wichtiges Fundament für die selbständige tiefere Auseinandersetzung. Dieses ermöglicht den Einsatz moderner Analysetechniken auch jenseits des Text Minings beispielsweise zur Netzwerk- oder Clusteranalyse.

Da eine umfangreiche Erweiterung des bestehenden Methodencurriculums aufgrund limitierter zeitlicher Ressourcen nicht möglich ist, stellt die Integration einer Softwareausbildung in die bestehenden Methodikmodule eine Möglichkeit dar, den Zusatzaufwand im Rahmen zu halten (MUNZERT 2014, 2018). Skriptbasierte Programmiersprachen sind generisch einsetzbar. Daher

können diese anstelle unterschiedlicher bestehender Softwareanwendungen, z.B. SPSS oder Stata, vermittelt werden. Entsprechend besteht die Möglichkeit sukzessive eine generisch einsetzbare Open-Source-Software (z.B. Python) als Standardwerkzeug in die (quantitativ-orientierte) Methodenlehre aufzunehmen. Sogar qualitative Auswertungen sind mit generischen Programmiersprachchen möglich (THIEM und DUSA 2013).

Neben den bekannten Vorteilen von Open-Source-Software erleichtert ein Grundverständnis von Programmiersprache die selbständige Erarbeitung fortgeschrittener Analysetechniken. Das Erlernen von Programmiersprachen ist zweifelsohne mit einigen Anlaufkosten verbunden. Allerdings existieren vielfältige Möglichkeiten, mittels Kursen geführt die ersten Erfahrungen zu sammeln. Exemplarisch zu nennen sind interaktive Videokurse (z.B. DataCamp) oder Tutorials, die den Einstieg erleichtern, eigenständiges Lernen vereinfachen und in die Lehre eingebunden werden können. Darüber hinaus ist NLP-Forschung stark von einem Open-Source-Gedanken geprägt, sodass in einigen Fällen keine eigenständige Software-Entwicklung stattfinden muss, sondern die Anwendung bestehender Modelle auf individuelle Forschungsfragen ausreicht. Somit werden ML-Anwendungen immer nutzerfreundlicher. Beispielsweise bietet Huggingface verschiedene ML-Applikationen an, die einfach über den Browser ausprobiert werden können.

11.3 Handlungsempfehlungen für die (Hochschul)politik

Aus den bisherigen Darstellungen wird deutlich, dass KI auch außerhalb der Forschung in vielen Lebensbereichen einen enormen technologischen Fortschritt erzeugen kann. Entsprechend sollte politisch sichergestellt werden, dass Auswirkungen, Nutzungsmöglichkeiten sowie ethische und sicherheitstechnische Implikationen von KI intensiv evaluiert werden und KI-Kompetenzen weiter ausgebaut werden können. An dieser Stelle nimmt akademische Forschung und Lehre eine zentrale Rolle ein, sodass diese Bereiche weiter gestärkt werden sollten, um langfristig in aktuelle KI-Entwicklungen involviert zu sein.

In der jüngeren Vergangenheit ist jedoch ein gegenläufiger Trend zu beobachten, sodass die Privatisierung von KI-Kompetenzen immer weiter zunimmt. Bedingt durch den explosionsartigen Anstieg der benötigten Rechenleistung um das 300.000-fache innerhalb der letzten Dekade ist im selben Zeitraum der Anteil der großskaligen wissenschaftlichen KI-Experimente von über 60 % auf nahezu 0 % gefallen (GANGULI et al. 2022). Die technische und infrastrukturelle Ausstattung wissenschaftlicher Einrichtungen stellt damit ein massives Hemmnis für die KI-Entwicklung dar. Wissenschaftler:innen können zwar über APIs auf die großen, privaten Modelle zurückgreifen. Jedoch rücken diese zunehmend hinter Bezahlschranken, sodass eine freie und offene wissenschaftliche Auseinandersetzung mit der Thematik immer weiter erschwert wird. Angesichts der erwartbaren sozialen, gesellschaftlichen und wirtschaftlichen Auswirkungen, die mit einer KI-

Transformation einhergehen können, kann diese Entwicklung nicht politisch gewollt sein. An dieser Stelle ist auch die Entwicklung eigener umfassender Sprachmodelle anzustreben. Diese bilden zunehmend das Rückgrat nachgelagerter Webanwendungen und Smartphone-Applikationen. Vor diesem Hintergrund nimmt die Relevanz solcher Modelle für soziale und gesellschaftliche weiter zu, sodass Politik und Wissenschaft auf der Höhe der Zeit bleiben sollte.

Eine zunehmende Privatisierung von KI-Forschung ist ebenso im wissenschaftlichen Output in Form von Artikeln zu beobachten. Auf der wichtigsten computerwissenschaftlichen Konferenz 2019 (Conference and Workshop on Neural Information Processing Systems) kamen 167 vorgestellte Artikel aus dem Hause Google (11,7 %), während die am zweit häufigsten vertretene Institution, die Stanford Universität, 82 Beiträge stellte (5,7 %). Entsprechend ist in der öffentlichen Forschung ein Brain-Drain in Richtung Privatwirtschaft zu beobachten, welcher eine noch stärkere Abkopplung der öffentlichen Forschung von aktuellen KI-Entwicklungen zur Folge haben könnte (JUROWETZKI et al. 2021). Gleichzeitig sind die populärsten Deep Learning Frameworks Tensorflow (GoogleBrain) und PyTorch (Meta AI) ebenfalls von amerikanischen Technologieunternehmen entwickelt worden, sodass das Forschungsfeld historisch eine starke privatwirtschaftliche Fokussierung aufweist.

Hochschulen und Politik sollten also gemeinschaftlich Lösungen finden, um KI-Forschenden eine attraktive Perspektive bieten zu können. Neben konkurrenzfähigen Gehältern und Forschungsfinanzierungen gilt es auch technisch-infrastrukturelle Voraussetzungen zu schaffen, die eine moderne KI-Forschung ermöglichen. An dieser Stelle könnten auch interdisziplinäre Forschungsprogramme beispielsweise mit Unternehmen eine Möglichkeit darstellen, um moderne KI-Forschung zu ermöglichen. Darüber hinaus sollte intensiver evaluiert werden, aus welchen Beweggründen Forschende den Wissenschaftsbetrieb in Richtung Privatwirtschaft verlassen.

12 Fazit

Abschließend soll das folgende Kapitel die eingangs formulierten Forschungsfragen beantworten und auf Basis der Synthese der Ergebnisse Perspektiven für weitere Forschung eruieren. Das übergeordnete Ziel dieser Arbeit war es die Techniken des NLP und Web Minings auf (wirtschafts)geographische Fragestellungen anzuwenden und somit zu prüfen, inwiefern die Methoden perspektivisch einen Mehrwert für wirtschaftsgeographische Forschung darstellen. Da die Methoden in der Geographie bisher kaum zur Anwendung kommen, wurden im ersten Teil der Arbeit zentrale Begrifflichkeiten erläutert, Verfahren des Web Scrapings und NLP vorgestellt und die Entwicklung des Forschungsfelds dargelegt. Im zweiten Teil der Arbeit wurde der Einsatz der vorgestellten Werkzeuge und Verfahren konkret anhand von wirtschaftsgeographischen Fallbeispielen demonstriert. Die Fallbeispiele skizzierten jeweils vollständige Forschungsprozesse von Datengrundlagen und –erhebungen bis hin zur Ergebnisinterpretation. Sie illustrierten ferner die eigenständige Entwicklung und den Einsatz unterschiedlicher überwachter und unüberwachter NLP-Verfahren, behandelten multilinguale Texte und stellten Möglichkeiten zur Quer- und Längsschnittsbetrachtung vor.

Die empirischen Ergebnisse wurden im dritten Teil der Arbeit innerhalb einer Synthese diskutiert. Aus dieser Diskussion heraus wurden übergeordnet Unterschiede zwischen den vorgestellten Verfahren und den tradierten Methoden der empirischen Sozialforschung beleuchtet. Da sich aus der neu entstehenden Diversität der Untersuchungsmethoden ebenfalls neue Forschungsmöglichkeiten ergeben, wurden in diesem dritten Teil ebenfalls Beispiele für integrierende Forschungsdesigns skizziert und daraus Handlungsempfehlungen abgeleitet.

12.1 Beantwortung der Forschungsfragen

Im folgenden Teil der Arbeit werden die Forschungsfragen angelehnt an den Web Mining Prozess abschließend beantwortet.

(1) Datenzugang

> *Wie kann das offene Webrepositorium CommonCrawl (CC) als Datengrundlage für empirische geographische Untersuchungen genutzt werden?*

Die Datenbeschaffung ist für die Wissenschaft von zentraler Bedeutung. Sie stellt die Basis für die weitere Analyse und Interpretation von empirischen Ergebnissen dar. In der jüngeren Vergangenheit hat sich das Internet zu einer enorm umfassenden Datenbasis entwickelt. Unter dem Begriff Big Data werden seit einigen Jahren Implikationen, Herausforderungen und Potentiale diskutiert. Insbesondere die feinkörnige Analyse vieler Einzelbeobachtungen können demnach zu deutlich umfassender empirischer Evidenz führen. Trotz der Möglichkeiten, die das Internet als Daten-

grundlage für empirische Forschung bereithält, bestehen bislang nur wenige systematische Ansätze zur umfassenden Nutzung von Webdaten. Dies lässt sich unter anderem mit mangelnden und wenig etablierten bestehenden Datenzugängen begründen, sodass Webforschung bereits an dieser Stelle auf große Barrieren stößt.

Um diese Barrieren leichter überwinden zu können, bietet die gemeinnützige Organisation CommonCrawl Foundation seit 2011 Zugang zu Webdaten mit dem übergeordneten Ziel Webforschung zu demokratisieren (NAGEL 2021). Das CC bietet Webdaten in vier verschiedenen Formaten an, sodass zunächst zu prüfen ist, welches Format für die jeweilige Forschungsfrage geeignet ist. Im Rahmen dieser Arbeit hat sich die Verwendung des URL-Index als geeignete Variante erwiesen. Dieser enthält Informationen zu Millionen von Domains und ermöglicht es diese vorab nach Sprache bzw. TLD zu filtern. Um Zugang zu den Inhalten der Domains zu erlangen, ist jedoch ein eigenständiger Abruf dieser notwendig. Zwar bietet das CC auch Formate inklusive der Webinhalte an. Jedoch ist nicht garantiert, dass die gesuchten Inhalte auch in den zufälligen Crawls enthalten sind. Außerdem stellt bei Verwendung älterer Crawls ein erneuter Abruf sicher, dass die Webseite zum Untersuchungszeitpunkt noch aktiv ist.

Die deskriptive Analyse der enthaltenen deutschsprachigen Domains zeigte, dass bereits mit wenigen Crawls ein Großteil der insgesamt enthaltenen Domains bezogen werden kann. Der betrachtete Umfang dieser Arbeit von 18 Crawls, die einen Zeitraum von zwei Jahren abdecken, bot somit Zugang zu fast zehn Millionen Domains. Damit ist das CC eine enorm umfassende Datengrundlage. Allerdings bestehen insbesondere für den deutschsprachigen Bereich keine tiefgreifenderen Untersuchungen, inwiefern das CC die tatsächliche Grundgesamtheit aller deutschsprachigen Domains repräsentiert. Daher sollten veredelte Daten, die aus dem CC extrahiert wurden, nochmals hinsichtlich Repräsentativität und Selektivität validiert werden.

(2) Datenabruf

> *Wie können systematisch Webmassendaten für Forschungszwecke aus dem Internet abgerufen werden?*

Neben dem grundsätzlichen Zugang potentiell relevanter Domains stellt das Web Scraping den zweiten zentralen Baustein moderner Web Mining-Vorhaben dar, um die Inhalte der jeweiligen Domains abzurufen. Dies ist nötig, um Domains kategorisieren zu können oder um Veränderungen der Inhalte im Zeitverlauf analysieren zu können. Um systematisch und performant Webseiten abrufen zu können, wurde in dieser Arbeit das Scrapy Framework verwendet (SCRAPY COMMUNITY 2022b). Dieses wurde explizit für umfassende Web Scraping-Vorhaben entwickelt und bietet viele vorinstallierte Möglichkeiten, um Scraping- und Crawlingprozesse effizient koordinieren zu können. So ist das Framework in der Lage mehrere Abrufe parallel zu tätigen, Inhalte zu

verarbeiten und diese abzuspeichern. Darüber hinaus ist Scrapy nahezu beliebig individualisierbar, sodass eigene Regeln, Heuristiken oder Verarbeitungsschritte implementiert werden können. In Kombination mit dem URL-Index des CC ermöglicht Scrapy eine umfassende Webforschung.

(3) Datenselektion- und vorverarbeitung

Wie können Unternehmensdomains identifiziert und georeferenziert werden?

Für die Wirtschaftsgeographie sind räumlich verortete Informationen zu Unternehmen eine wichtige Datenquelle für weitere Analysen. Daher stellt die Geokodierung von Domains eine zentrale Voraussetzung für weitere geographische Untersuchungen dar. Da Domains per se keine Ortsangaben zu entnehmen sind, sind zusätzliche Verarbeitungsschritte notwendig, um Adressdaten aus Webseitentexten zu extrahieren und diese zu geokodieren.

Dieses Forschungsziel greift die Vorarbeiten der beiden vorherigen Arbeitsschritte auf. Auf Basis der gewonnenen Domains aus dem URL-Index sollten mittels Scrapy Webseiten abgerufen und deren Inhalte verarbeitet werden. Konkret sollten aus der Gesamtheit deutschsprachiger Domains Unternehmensdomains identifiziert und lokalisiert werden. Hierzu wurde im Programmcode von Scrapy definiert, dass auf den Startseiten der Domains nach Webseiten gesucht werden sollte, die auf eine Impressumswebseite verweisen. Diese wurden abgerufen und der Text aus den Webseiten extrahiert. Anschließend wurde ein individuell trainiertes NER-Modell eingesetzt, um Adressinformationen aus den Impressen auszulesen. Insbesondere in Deutschland stellt das Impressum eine geeignete Möglichkeit dar, um Informationen zu Webseitenbetreiber:innen zu erhalten. Um Adressen in einzelne Bestandteile gliedern zu können, sind NER-Modelle gut geeignet. Zu extrahierende Entitäten können individuell festgelegt und das Modell entsprechend trainiert werden.

Die extrahierten Bestandteile der Adressen wurden anschließend geokodiert und die extrahierten Domains kartographisch dargestellt. Die geographische Beleuchtung offenbarte signifikante Verteilungsunterschiede. So konnten Unterschiede zwischen Stadt und Land, zentralen und peripheren Gemeinden sowie zwischen Gemeinden der ehemaligen DDR und BRD festgestellt werden. Darüber hinaus besteht eine höchstsignifikante Korrelation zwischen der Anzahl der identifizierten Unternehmensdomains und der Anzahl der Unternehmen auf Gemeindeebene. Der Vergleich mit bestehenden georeferenzierten Analysen von Unternehmenswebseiten zeigte in der Absolutanzahl identifizierter Domains ebenfalls große Überschneidungen. Auf Basis dieser Ergebnisse kann die Variable *Webseitenanteil von Unternehmen* als neuer Indikator zur Messung des Digitalisierungsgrads von Unternehmen betrachtet werden.

Das entwickelte Verfahren lässt sich prinzipiell auch in anderen Kontexten anwenden, um Webinhalte systematisch verorten zu können. Neben der Erweiterung der Untersuchungseinheiten

durch die Integration von weiteren regionalen Akteuren (z.B. öffentliche Einrichtungen, Forschungseinrichtungen, Vereine) könnten potentiell auch Unternehmen im Ausland über die beschriebene Heuristik identifiziert und geographisch verortet werden.

(4) Datenanalyse

Wie kann NLP eingesetzt werden, um Unternehmenswebseiten nach Technologienutzung zu klassifizieren?

Über die geographische Verortung von Unternehmensdomains hinaus war es Ziel dieser Forschungsfrage ein Verfahren zu entwickeln, das in der Lage ist, automatisiert Texte zu klassifizieren. Die Kombination aus räumlich fein aufgelösten Daten in Form von geokodierten Domains und inhaltlich frei definierbaren Filterschritten durch die Verarbeitung des Webseitentexts ermöglichen es raumbezogener Forschung, neue Perspektiven einzunehmen. Für die Wirtschaftsgeographie stellen solche granularen Daten hinsichtlich spezifischer Technologienutzungen eine wertvolle Datenquelle dar, um die Hintergründe regionaler Wirtschaftsentwicklung besser verstehen zu können.

Im Rahmen von Fallstudie 2 wurden weitere Inhalte der Unternehmenswebseiten bezogen und deren Texte für die weitere Verarbeitung aufbereitet. Anschließend wurden die Texte der Unternehmenswebseiten hinsichtlich der KI-Nutzung des Unternehmens klassifiziert. Ein besonderes Augenmerk dieser Fallstudie lag auf dem Training und dem Einsatz eines Textklassifikationsmodells auf Basis der Transformerarchitektur. Dieses war nach abgeschlossenem Training in der Lage, über 85 % der Absätze des Testdatensatzes korrekt zu klassifizieren. Der Textklassifikation wurde eine Stichwortsuche vorgeschaltet, welche eine effiziente Möglichkeit darstellte, um die feiner zu klassifizierende Textmenge drastisch zu reduzieren.

Abschließend kann festgehalten werden, dass Textklassifikationsmodelle bereits mit überschaubaren Mengen an Trainingsdaten in der Lage sind, Text verlässlich zu kategorisieren. Besonders hervorzuheben ist dabei die Fähigkeit multilingualer vortrainierter Sprachmodelle. Einerseits können mit diesen Modellen mehrsprachtige Korpora analysiert werden ohne, dass mehrere Modelle trainiert werden müssen. Andererseits sorgt das Transfer Learning (vgl. Kapitel 4.8) dafür, dass bereits mit wenigen Trainingsbeispielen robuste und performante Vorhersagen getroffen werden können. Dieses Vorgehen lässt sich prinzipiell auf unterschiedliche Fragestellungen übertragen und ermöglicht so die Generierung neuer Untersuchungsvariablen auf Basis von unstrukturiertem Webseitentext.

Inwiefern können mittels NLP Themen innerhalb großer Textkorpora modelliert werden?

Neben der gezielten Filterung und Klassifikation von Dokumenten können NLP-Verfahren eingesetzt werden, um Themen in Textkorpora zu identifizieren und zu analysieren. Angesichts der

enormen Menge digitalisierter Texte können diese Verfahren helfen, Textmengen semantisch zu erschließen, die menschliche Verarbeitungskapazitäten bei weitem übersteigen. Insbesondere die Analyse großer Textmengen im Zeitverlauf bieten neue Möglichkeiten, wissenschaftliche Evidenz zu unterschiedlichen Veränderungsprozessen zu beleuchten. Daher zielte die Forschungsfrage darauf ab, Verfahren zur Exploration abstrakter Themen in großen Textkorpora anzuwenden.

In Fallstudie 3 wurde hierzu ein unüberwachtes NLP-Verfahren eingesetzt. Dieses weist Dokumenten eines Textkorpus automatisch Themen zu. Im Unterschied zu den überwachten Lernverfahren, die in den ersten beiden Fallstudien eingesetzt wurden, benötigen unüberwachte Verfahren keine Trainingsdaten. Entsprechend handelt es sich hierbei eher um ein explorativ-induktives Vorgehen, welches Einblicke in die Themen deren Verteilung und Verwandtschaft großer Textkorpora gibt. Exemplarisch wurde in Fallstudie 3 die Themenverteilung ausgewählter wirtschaftsgeograhischer Literatur untersucht. Das verwendete Topic Modeling-Verfahren basiert ebenfalls auf der in Kapitel 4.6 vorgestellten Transformerarchitektur. Das verwendete Python-Paket BERTopic beinhaltet eine Vielzahl an Möglichkeiten, um Textkorpora semantisch zu explorieren (GROOTENDORST 2022). Neben der Möglichkeit, unterschiedliche Dimensionsreduktions- sowie Clusteralgorithmen einzusetzen, können Themen fusioniert und aufgespalten, Ähnlichkeiten zwischen Themen bestimmt und Hierarchien aus diesen generiert werden. Darüber hinaus stellt die Möglichkeit, Themenveränderungen über Zeit analysieren zu können ein wichtiges Werkzeug zur Beobachtungen von Diskursveränderungen dar.

Über die Analyse von wissenschaftlichen Publikationen hinaus bietet das Verfahren vielfältige Anwendungsmöglichkeiten. Beispielsweise können Themen innerhalb von Nachrichtenartikeln, Webseiteninhalten oder grauer Literatur dynamisch modelliert werden. Die daraus ableitbaren Indikatoren können wiederum verwendet werden, um größere gesellschaftliche, politische oder soziale Veränderungsprozesse zu erklären. Speziell für die Wirtschaftsgeographie können Längsschnittsbetrachtungen regional verankerter Textdokumente einen zusätzlichen Erklärungsgehalt liefern beispielsweise, um unterschiedliche Entwicklungspfade von Regionen besser verstehen und quantifizieren zu können. Dabei kann die Ableitung von quantitativer Indikatorik aus qualitativen Textdaten heraus eine wertvolle Ergänzung zu etablierten Strukturindikatoren darstellen.

(5) Dateninterpretation

> *Wie können Verfahren des Web Mining und NLP in klassische wirtschaftsgeographische Forschungsdesigns eingebunden werden?*

Während die vorherigen Forschungsfragen im Rahmen der einzelnen Fallstudien beantwortet wurden, ergibt sich die Antwort auf diese Forschungsfrage aus der Synthese der Fallstudien heraus. Für viele Forschungsfragen in der Wirtschaftsgeographie sind inhaltlich, räumlich und zeitlich fein aufgelöste Daten notwendige Bedingung, um verlässliche Aussagen über die komplexen

Wechselwirkungen zwischen im Raum verankerten Akteuren und deren Outputs treffen zu können. Die Fallstudien dieser Arbeit zeigten, dass sowohl räumlich als auch inhaltlich granulare Daten mittels Text Mining gewonnen werden können. In Fallstudie 2 wurde darüber hinaus demonstriert, wie sich diese Daten in ein klassisch quantitatives Forschungsdesign einbinden lassen.

Insofern können mittels Web Mining und NLP Datensätze erzeugt werden, welche mit tradierten Methoden der empirischen Sozialforschung kaum operationalisierbar sind. Diese Möglichkeit stellt den zentralen Mehrwert dieser Verfahren für die Wirtschaftsgeographie dar. Die neu erzeugten Indikatoren können anschließend unterschiedlich in komplexere Forschungsdesigns eingebunden werden. Einerseits in rein quantitativen Forschungsprojekten, welche klassische Datenquellen (z.B. Sekundärstatistiken, Patentdaten, Befragungsdaten) mit Webindikatoren ergänzen. Wie Fallstudie 2 zeigte, können durch das koordinatenscharfe Material an ökonometrische Betrachtung noch weitere mikrogeographische Analysen anschließen. Andererseits ergeben sich Möglichkeiten zur Kopplung von Text Mining Bausteinen mit qualitativen Elementen im Rahmen von Mixed-Methods-Designs. Die besondere Stärke dieser Verbindung liegt darin begründet, dass umfassendes Textmaterial computerbasiert vorstrukturiert und anschließend mittels qualitativen Untersuchungen näher ergründet werden kann.

Zusammenfassend lässt sich daher festhalten, dass die Integration der im Rahmen dieser Arbeit untersuchten Methoden in die wirtschaftsgeographische Forschungspraxis eine Vielzahl neuer Perspektiven eröffnet. Bestehende quantitative Arbeiten greifen datenseitig häufig auf Sekundärstatistiken, Patent- und Publikationsdaten zurück, um Regionen strukturell vergleichen zu können. Die Möglichkeit mittels Text Mining qualitative Texte auch quantitativ fassen zu können, bieten besonders für die Wirtschaftsgeographie neue Perspektiven. So fehlt es einschlägiger Forschung bis dato an Ansätzen, die die Kontextualisierung des Untersuchungsgegenstandes und die Detailschärfe als relevante Stärken qualitativer Forschungsdesigns mit der Generalisierbarkeit und externen Validität quantitativer Verfahren verbinden.

Im Kontext raumbezogener Forschung schlägt sich diese Beobachtung zumeist auch auf die geographische Betrachtungsebene nieder. Entsprechend können qualitative Designs einzelne Akteure in den Fokus nehmen, während quantitative Ansätze häufig auf räumlich aggregierte Daten zurückgreifen müssen. Administrative Daten verschleiern jedoch viele individuelle Entwicklungen einzelner Akteur:innen, sodass keine differenzierte Exploration latenter Triebkräfte sozioökonomischer Wandlungspfade stattfinden kann. Vielmehr sind administrative Gebietseinheiten starre Konstrukte, weswegen auch die zugehörigen Daten lediglich mathematische Mittelwerte darstellen (FELDMAN und LOWE 2015; FELDMAN et al. 2015). Da Text Mining sowohl im Sinne induktiver als auch deduktiver Ansätze verwendet werden kann, entsteht durch die gemeinschaftliche Nutzung der Datengrundlage Text eine engere Verknüpfung qualitativer und quantitativer

Denkrichtungen. Des Weiteren ermöglichen Mikrodaten eine einheitlichere Akteursorientierung zwischen qualitativen und quantitativen Bausteinen.

Trotz der dargestellten Potentiale ist das Forschungsfeld noch jung und wenig ausdifferenziert, sodass weitere methodische und inhaltliche Forschung darauf abzielen sollte, ein profundereres Verständnis für die Implikationen, Grenzen und Möglichkeiten der Methodik zu schaffen. Konkreter weiterer Forschungsbedarf wird daher im nächsten Kapitel vorgestellt.

12.2 Ausblick und weiterer Forschungsbedarf

Aus einer methodenexplorierenden Arbeit wie dieser ergibt sich eine Vielzahl an Möglichkeiten wie weitere Forschung an die vorliegenden Ergebnisse anknüpfen kann. Insbesondere aus inhaltlicher Sicht eröffnet das Methodenset einhergehend mit den generierten Datensätzen neue Forschungsperspektiven für die Wirtschaftsgeographie.

Zunächst sollte im Rahmen aufsetzender Forschung der generierte Datensatz zu deutschen Unternehmensdomains tiefgreifender validiert werden. Dies kann beispielsweise durch eine Harmonisierung des Webdatensatzes mit bestehenden Datentöpfen gelingen. Ein Ansatzpunkt bietet das deutsche Handelsregister. Die vorliegende Datengrundlage erlaubt es, dieses mit den identifizierten Unternehmensdomains zu vereinen. Hieraus lassen sich weitere Informationen zu den Unternehmen ableiten, z.B. das Gründungsdatum, das Stammkapital, die Geschäftsführer:innenstruktur oder der Unternehmensgegenstand. Weitere Matchingmöglichkeiten bestehen mit Patentdaten. Diese spielen in innovationsbezogener Forschung eine zentrale Rolle und könnten somit einen erheblichen Mehrwert darstellen, indem sich die Informationen aus den jeweiligen Datentöpfen wechselseitig ergänzen.

Zu einer umfassenden Validierung gehört ebenfalls eine qualitative Betrachtung. Somit können Klassifikationen oder Themenbeschreibungen, die mittels Text Mining gewonnen wurden, gezielt an Fallstudien qualitativ nachvollzogen werden. Erkenntnistheoretisch stellt dieses Vorgehen einerseits ein spannendes Untersuchungsdesign dar, welches qualitative und quantitative Bausteiene synergetisch verbindet. Andererseits stellt eine detaillierte Befragung identifizierter Akteure (z.B. Unternehmen) und deren Multiplikatoren (z.B. Industrie- und Handelskammern, Wirtschaftsförderungen) einen notwendigen Arbeitsschritt dar, um die quantitativen Ergebnisse tiefer zu validieren und die Grenzen der Methodik auszuloten. Diese Untersuchungen leisten somit einen wichtigen Beitrag, um Implikationen, Grenzen und Charakteristika der präsentierten Methoden besser verstehen zu können.

Auf Basis dessen kann der Untersuchungsbereich erweitert werden. Beispielsweise können analog zu dem in Kapitel 6 beschriebenen Vorgehen Domains der öffentlichen Hand (z.B. Gemeinde-

oder Kreisdomains), von Vereinen oder wissenschaftlichen Einrichtungen identifiziert und verortet werden. Diese leisten beispielsweise in Theoriekonzepten zu regionalen Innovationssystemen einen wichtigen Beitrag und sollten daher perspektivisch in die Betrachtung aufgenommen werden. Außerdem gilt es zu prüfen, inwiefern Möglichkeiten bestehen, mittels der dargelegten Methodik, Unternehmen international zu identifizieren. Eine dahingehende Erweiterung des Datensatzes würde nochmals neue Forschungsmöglichkeiten eröffnen, beispielsweise zum internationalen Vergleich.

Des Weiteren bietet der im Rahmen dieser Arbeit generierte Datensatz einen Ausgangspunkt für Untersuchungen der Hyperlinkstruktur der Unternehmen. Aus diesen Verlinkungen lassen sich Netzwerke erzeugen, welche anschließend mit gängigen Verfahren der Netzwerkanalyse beleuchtet werden können. Auch an dieser Stelle sollte eine qualitative Betrachtung vorgeschaltet werden, um ein besseres Verständnis für die Gründe von Hyperlinks auf Unternehmenswebseiten zu erlangen. Gegebenenfalls könnte über die Hyperlinkstruktur eine innovative Möglichkeit bestehen, Relationen zwischen Unternehmen und anderen Akteuren abzuleiten.

Darüber hinaus können die Inhalte der Unternehmensdomains in regelmäßigen Intervallen abgerufen werden. Ein so entstehender Paneldatensatz bietet ebenfalls neue Möglichkeiten zur Analyse von (Innovations)dynamiken, Diffusions- und Disseminationsprozessen sowie Adaptionsvorgängen. Insbesondere die Möglichkeit, Webseiten in nahezu beliebig kurzen Intervallen abzufragen, kann zu einem deutlich profunderen Verständnis von zeitlichen Veränderungen beitragen. Ergänzend hierzu kann ein fortlaufendes Monitoring des CC dabei helfen, neue Unternehmensdomains in den bestehenden Datensatz zu integrieren. Inwiefern das CC geeignet ist, um Unternehmensgründungen zu erkennen, könnte ebenfalls im Rahmen weiterer Forschungsprojekte analysiert werden. Neben wissenschaftlichen Betrachtungen können diese Analysen eine wichtige Entscheidungsgrundlage für Politiker:innen sein, um räumlich und zeitliche granulare Informationen über die Unternehmenslandschaft zu gewinnen.

Auch auf methodischer Seite ergeben sich neue Anknüpfungspunkte. Hinsichtlich des Webseitenabrufs gilt es zu evaluieren, welche Inhalte in welchem Umfang abgerufen werden müssen, um ein möglichst vollständiges Profil der Domain ableiten zu können. Bestehende Studien beziehen nahezu ungefiltert Webseiteninhalte, sodass auch ein großer Teil irrelevanter Inhalte abgerufen wird. An dieser Stelle benötigt es neue Verfahren, die gezieltes, priorisierendes Webcrawling ermöglichen. Ein Ausgangspunkt wäre z.B. die Semantik der Links der jeweiligen Domains. Überwachte Klassifikationsverfahren könnten dazu beitragen enorme Effizienzgewinne beim Webseitenabruf sowie der weiteren Prozessierung auszulösen. Ein solches Verfahren könnte prinzipiell auch in der Lage sein, bereits beim Abruf einer Webseite die Inhalte grob semantisch zu fassen.

Eine solche Annotierung des Datensatzes in Themenbereiche (z.B. Karriere, Produkte, Historie, Partner) könnte ebenfalls aufsetzende Analysen deutlich beschleunigen.

Ferner deuten neueste Entwicklungen im Bereich NLP an, dass eine weitere Leistungssteigerung von Klassifikationsverfahren zu erwarten ist. Einerseits werden immer weniger Trainingsdaten benötigt, da das grundlegende Sprachverständnis vortrainierter Modelle immer weiter zunimmt. Andererseits sind ebenfalls deutliche Steigerungen der Vorhersagegenaugikeit erwartbar. Im Bereich der Wirtschaftsgeographie kann eine fortlaufende Evaluierung der Eignung dieser Tools dabei helfen perspektivisch auch sehr unscharfe, weiche und komplexe Konzepte aus Webseitentexten zu extrahieren. Beispielhaft zu nennen sind Konzepte wie regionale Embeddedness, Corporate Social Responsibility, verschiedene Dimensionen von Nähe oder regionale Kooperationsintensitäten.

13 Literaturverzeichnis

ABBASI, O.; ALESHEIKH, A. & SHARIF, M. (2017): Ranking the City: The Role of Location-Based Social Media Check-Ins in Collective Human Mobility Prediction. In: ISPRS International Journal of Geo-Information 6(5): 136–149.

ABBASIHAROFTEH, M.; KINNE, J. & KRÜGER, M. (2021): The Strength of Weak and Strong Ties in Bridging Geographic and Cognitive Distances. In: Leibniz-Zentrum für Europäische Wirtschaftsforschung 21: 21–49.

AGHION, P.; JONES, B. & JONES, C. (2017): Artificial Intelligence and Economic Growth. Cambridge: National Bureau of Economic Research.

AGRAWAL, A. & COCKBURN, I. (2003): The anchor tenant hypothesis: exploring the role of large, local, R&D-intensive firms in regional innovation systems. In: International Journal of Industrial Organization 21(9): 1227–1253.

AGRAWAL, A.; COCKBURN, I.; GALASSO, A. & OETTL, A. (2014): Why are some regions more innovative than others? The role of small firms in the presence of large labs. In: Journal of Urban Economics 81: 149–165.

AGRAWAL, A.; MCHALE, J. & OETTL, A. (2019): Finding Needles in Haystacks: Artificial Intelligence and. In: GOLDFARB, A.; GANS, J. & AGRAWAL, A. (Hrsg.): The economics of artificial intelligence: An agenda. Chicago: University of Chicago Press: 149–174.

AI, Q.; YANG, L.; GUO, J. & CROFT, W. B. (2016): Analysis of the Paragraph Vector Model for Information Retrieval. In: Association for Computing Machinery (Hrsg.): Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. New York: ACM.

ANGELOV, D. (2020): Top2Vec: Distributed Representations of Topics. In: arXiv preprint arXiv: 2008.09470: 1–25.

ANSELIN, L. (1995): Local Indicators of Spatial Association-LISA. In: Geographical Analysis 27(2): 93–115.

ARORA, S. K.; YOUTIE, J.; SHAPIRA, P.; GAO, L. & MA, T. (2013): Entry strategies in an emerging technology: a pilot web-based study of graphene firms. In: Scientometrics 95(3): 1189–1207.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; DEL SER, J.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCIA, S.; GIL-LOPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R. & HERRERA, F. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. In: Information Fusion 58: 82–115.

ARTS, S.; HOU, J. & GOMEZ, J. C. (2021): Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. In: Research Policy 50(2): 104–144.

ASKITAS, N. & ZIMMERMANN, K. F. (2009): Google Econometrics and Unemployment Forecasting. In: Applied Economics Quarterly 55(2): 107–120.

AUDRETSCH, D. B. & FELDMAN, M. P. (1996): R&D spillovers and the geography of innovation and production. In: The American economic review 86(3): 630–640.

BA, J. L.; KIROS, J. R. & HINTON, G. E. (2016): Layer Normalization. In: arXiv preprint arXiv:1607.06450: 1–14.

BAEVSKI, A.; EDUNOV, S.; LIU, Y.; ZETTLEMOYER, L. & AULI, M. (2019): Cloze-driven Pretraining of Self-attention Networks. In: arXiv preprint arXiv:1903.07785: 1–10.

BAHDANAU, D.; CHO, K. & BENGIO, Y. (2015): Neural Machine Translation by Jointly Learning to Align and Translate. In: arXiv preprint: 1409.0473v7: 1–15.

BALAKRISHNAN, V. & LLOYD-YEMOH, E. (2014): Stemming and lemmatization: A comparison of retrieval performances. In: Lecture Notes on Software Engineering 2(3): 174–179.

BALDINI, N.; GRIMALDI, R. & SOBRERO, M. (2007): To patent or not to patent? A survey of Italian inventors on motivations, incentives, and obstacles to university patenting. In: Scientometrics 70(2): 333–354.

BALSMEIER, B.; ASSAF, M.; CHESEBRO, T.; FIERRO, G.; JOHNSON, K.; JOHNSON, S.; LI, G.-C.; LÜCK, S.; O'REAGAN, D.; YEH, B.; ZANG, G. & FLEMING, L. (2018): Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. In: Journal of Economics & Management Strategy 27(3): 535–553.

BARBARESI, A. (2021): Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Stroudsburg, PA, USA: Association for Computational Linguistics.

BARCAROLI, G.; SCANNAPIECO, M. & SUMMA, D. (2016): On the use of internet as a data source for official statistics: a strategy for identifying enterprises on the web. In: Rivista italiana di economia, demografia e statistica 70(4): 20–41.

BATHELT, H. & GLÜCKLER, J. (2003): Toward a relational economic geography. In: Journal of Economic Geography 3: 117–144.

BATHELT, H. & LI, P. (2020): Building Better Methods in Economic Geography. In: Zeitschrift für Wirtschaftsgeographie 64(3): 103–108.

BATHELT, H.; MALMBERG, A. & MASKELL, P. (2004): Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. In: Progress in Human Geography 28(1): 31–56.

BATHELT, H. & TURI, P. (2011): Local, global and virtual buzz: The importance of face-to-face contact in economic interaction and possibilities to go beyond. In: Geoforum 42(5): 520–529.

BAUR, N. & BLASIUS, J. (2022): Methoden der empirischen Sozialforschung. In: BAUR, N. & BLASIUS, J. (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS: 1–32.

BBSR (2020): Raumbeobachtung - Downloads. Raumtypen: Besiedelung und Lage. www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/downloads/downloadsReferenz2.html (04.05.2022).

BELTAGY, I.; LO, K. & COHAN, A. (2019): SciBERT: A Pretrained Language Model for Scientific Text. In: arXiv preprint arXiv:1903.10676: 1–6.

BENGIO, Y.; DUCHARME, R. & VINCENT, P. (2000): A neural probabilistic language model. In: Advances in Neural Information Processing Systems 13: 1–7.

BENGIO, Y.; SIMARD, P. & FRASCONI, P. (1994): Learning long-term dependencies with gradient descent is difficult. In: IEEE Transactions on Neural Networks 5(2): 157–166.

BENNEWORTH, P. & HOSPERS, G.-J. (2007): The New Economic Geography of Old Industrial Regions: Universities as Global — Local Pipelines. In: Environment and Planning C: Government and Policy 25(6): 779–802.

BENNEWORTH, P.; JONES, G. A. & PINHEIRO, R. (Hrsg.) (2012): Universities and regional development. A critical assessment of tensions and contradictions. International studies in higher education. 0. Aufl. New York, N.Y.: Routledge.

BENSBERG, F. & BUSCHER, G. (2016): Job Mining als Analyseinstrument für das Human-Resource-Management. In: HMD Praxis der Wirtschaftsinformatik 53(6): 815–827.

BERĘSEWICZ, M.; LEHTONEN, R.; REIS, F.; DI CONSIGLIO, L. & KARLBERG, M. (2018): An overview of methods for treating selectivity in big data sources. In: Eurostat Statistical Working Paper: 1–114.

BERSCH, J.; GOTTSCHALK, S.; MUELLER, B. & NIEFERT, M. (2014): The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany. In: ZEW Discussion Papers 14(104): 1–17.

BETHLEHEM, J. (2010): Selection Bias in Web Surveys. In: International Statistical Review 78(2): 161–188.

BEYER, M. A. & DO LANEY (2012): The importance of 'big data': a definition. In: CT: Gartner: 2014–2018.

BIANCHI, F.; TERRAGNI, S. & HOVY, D. (2021): Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In: arXiv preprint arXiv:2004.03974: 1–8.

BIANCHINI, S.; MÜLLER, M. & PELLETIER, P. (2020): Deep Learning in Science. In: arXiv preprint arXiv:2009.01575: 1–67.

BIEMANN, C.; HEYER, G. & QUASTHOFF, U. (2022): Wissensrohstoff Text - Konzepte, Algorithmen, Ergebnisse. Eine Einführung in das Text Mining. Springer eBook Collection. 2. Aufl. Wiesbaden: Springer.

BLAZQUEZ, D. & DOMENECH, J. (2018): Web Data Mining for monitoring business export orientation. In: Technological and Economic Development of Economy 24(2): 406–428.

BLEI, D. M.; ANDREW Y., N. & MICHAEL, I. J. (2003): Latent dirichlet allocation. In: Journal of Machine Learning Research 3: 993–1022.

BLEI, D. M. & LAFFERTY, J. D. (2007): A correlated topic model of Science. In: The Annals of Applied Statistics 1(1): 17–35.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A. & MIKOLOV, T. (2017): Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics 5: 135–146.

BOLLIER, D. & CHARLES, M. F. (2010): The promise and peril of big data. Washington: Aspen Institute.

BOUMA, G. (2009): Normalized (Pointwise) Mutual Information in Collocation Extraction. In: Proceedings of GSCL 30: 31–40.

BOYD, D. & CRAWFORD, K. (2012): Critical questions for Big Data. In: Information, Communication & Society 15(5): 662–679.

BRAMWELL, A. & WOLFE, D. A. (2008): Universities and regional economic development: The entrepreneurial University of Waterloo. In: Research Policy 37(8): 1175–1187.

BRANTS, S.; DIPPER, S.; EISENBERG, P.; HANSEN-SCHIRRA, S.; KÖNIG, E.; LEZIUS, W.; ROHRER, C.; SMITH, G. & USZKOREIT, H. (2004): TIGER: Linguistic Interpretation of a German Corpus. In: Research on Language and Computation 2(4): 597–620.

BREFCZYNSKI, J. A. & DEYOE, E. A. (1999): A physiological correlate of the 'spotlight' of visual attention. In: Nature Neuroscience 2(4): 370–374.

BRESCHI, S. (2001): Knowledge Spillovers and Local Innovation Systems: A Critical Survey. In: Industrial and Corporate Change 10(4): 975–1005.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P. & AMODEI, D. (2020): Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33: 1877–1901.

BRYNJOLFSSON, E. & MITCHELL, T. (2017): What can machine learning do? Workforce implications. In: Science 358(6370): 1530–1534.

BUBENHOFER, N. (2017): Kollokationen, n-Gramme, Mehrworteinheiten. In: KERSTEN, S. R.; WENGELER, M. & ZIEM, A. (Hrsg.): Handbuch Sprache in Politik und Gesellschaft. Berlin, Boston: De Gruyter: 69–93.

BULKELEY, H. & NEWELL, P. (2015): Governing climate change. Routledge global institutions series. 2. Aufl. Abington: Routledge.

BUNDESAGENTUR FÜR ARBEIT (2022): Studiensuche. web.arbeitsagentur.de/studiensuche (22.08.2022).

BUNDESMINISTERIUM FÜR BILDUNG UND FORSCHUNG (BMBF) (2011): Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung von Forschungs- und Entwicklungsvorhaben aus dem Bereich der eHumanities. www.bmbf.de/foerderungen/bekanntmachung-643.html (18.08.2022).

BUNDESMINISTERIUM FÜR BILDUNG UND FORSCHUNG (BMBF) (2020): BuFI - Liste der Einrichtungen. www.bundesbericht-forschung-innovation.de/de/Liste-der-Einrichtungen-1790.html (21.09.2022).

BUNDESMINISTERIUM FÜR BILDUNG UND FORSCHUNG (BMBF) (2022): Digital Humanities. www.geistes-und-sozialwissenschaften-bmbf.de/de/Digital-Humanities-1710.html (18.08.2022).

CAMERON, M. P.; BARRETT, P. & STEWARDSON, B. (2016): Can Social Media Predict Election Results? Evidence From New Zealand. In: Journal of Political Marketing 15(4): 416–432.

CARABANTES, M. (2020): Black-box artificial intelligence: an epistemological and critical analysis. In: AI & SOCIETY 35(2): 309–317.

CARNEIRO, H. A. & MYLONAKIS, E. (2009): Google trends: a web-based tool for real-time surveillance of disease outbreaks. In: Clinical Infectious Diseases 49(10): 1557–1564.

CASTELVECCHI, D. (2016): Can we open the black box of AI? In: Nature 538(7623): 20–23.

CHARTON, F.; HAYAT, A. & LAMPLE, G. (2021): Learning advanced mathematical computations from examples. In: arXiv preprint arXiv:2006.06462: 1–25.

CHATTERJEE, R. P.; RAY, C. & BAG, R. (2017): A Comparative Study on Latest Substring Association Rule Mining and Hidden Markov Model. In: 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE). IEEE.

CHESHIRE, J. & BATTY, M. (2012): Visualisation Tools for Understanding Big Data. In: Environment and Planning B: Planning and Design 39(3): 413–415.

CHO, K.; VAN MERRIENBOER, B.; GULCEHRE, C.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H. & BENGIO, Y. (2014): Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: arXiv preprint arXiv: 1406.1078: 1–15.

CHO, Y. H. & KIM, J. K. (2004): Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. In: Expert Systems with Applications 26(2): 233–246.

CHOI, H. & VARIAN, H. A. (2012): Predicting the Present with Google Trends. In: Economic Record 88: 2–9.

CLARK, K.; KHANDELWAL, U.; LEVY, O. & MANNING, C. D. (2019): What Does BERT Look At? An Analysis of BERT's Attention. In: arXiv preprint arXiv:1906.04341: 1–11.

CLARK, K.; LUONG, M.-T.; MANNING, C. D. & LE V, Q. (2018): Semi-Supervised Sequence Modeling with Cross-View Training. In: arXiv preprint arXiv:1809.08370: 1–17.

CLEMENS, K. (2015): Geocoding with openstreetmap data. In: GEOProcessing 10: 1–2.

COCKBURN, I. M.; HENDERSON, R. & & STERN, S. (2019): The impact of artificial intelligence on innovation: An exploratory analysis. In: GOLDFARB, A.; GANS, J. & AGRAWAL, A. (Hrsg.): The economics of artificial intelligence: An agenda. Chicago: University of Chicago Press: 115–146.

COHEN, J. (1992): Statistical Power Analysis. In: Current Directions in Psychological Science 1(3): 98–101.

COLLOBERT, R. & JASON, W. (2008): A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning: 1–8.

COLLOBERT, R.; WESTON, J.; BOTTOU LEON; KARLEN, M.; KAVUKCUOGLU, K. & KUKSA, P. (2011): Natural language processing (almost) from scratch. In: Journal of Machine Learning Research 2011(12): 2493–2537.

COMMONCRAWL (2022): Common Crawl. commoncrawl.org/ (13.09.2022).

COOKE, P. (2001): Regional Innovation Systems, Clusters, and the Knowledge Economy. In: Industrial and Corporate Change 10(4): 945–974.

COOKE, P.; HEIDENREICH, M. & BRACZYK, H.-J. (2004): Regional innovation systems. The role of governance in a globalized world. 2. Aufl. London, New York: Routledge.

COX, D. R. & SNELL, E. J. (2018): Analysis of Binary Data. 2. Aufl. Boca Raton: Routledge.

DAAS, P. & VAN DER DOEF, S. (2021): Using Website texts to detect Innovative Companies. In: Center for Big Data Statistics Working Paper 21(1): 1–20.

DAELEMANS, W. & HOSTE, V. (2002): Evaluation of machine learning methods for natural language processing tasks. In: LREC 2002 third international conference on language resources and evaluation. European Language Resources Association (ELRA).

DAI, A. M.; OLAH, C. & LE, Q. V. (2015): Document Embedding with Paragraph Vectors. In: arXiv preprint arXiv:1507.07998: 1–8.

DAVENPORT, T. H. & PRUSAK, L. (1998): Working knowledge. How Organizations Manage what They Know. Boston, Mass: Harvard Business School Press.

DAVIES, A.; VELIČKOVIĆ, P.; BUESING, L.; BLACKWELL, S.; ZHENG, D.; TOMAŠEV, N.; TANBURN, R.; BATTAGLIA, P.; BLUNDELL, C.; JUHÁSZ, A.; LACKENBY, M.; WILLIAMSON, G.; HASSABIS, D. & KOHLI, P. (2021): Advancing mathematics by guiding human intuition with AI. In: Nature 600(7887): 70–74.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K. & HARSHMAN, R. (1990): Indexing by latent semantic analysis. In: Journal of the American Society for Information Science 41(6): 391–407.

DELGADO, M.; PORTER, M. E. & STERN, S. (2016): Defining clusters of related industries. In: Journal of Economic Geography 16(1): 1–38.

DENNY, M. J. & SPIRLING, A. (2018): Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. In: Political Analysis 26(2): 168–189.

DEVLIN, J.; CHANG, M.-W.; LEE, K. & TOUTANOVA, K. (2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: arXiv preprint arXiv:1810.04805: 1–16.

DEVRIENDT, L.; DERUDDER, B. & WITLOX, F. (2008): Cyberplace and Cyberspace: Two Approaches to Analyzing Digital Intercity Linkages. In: Journal of Urban Technology 15(2): 5–32.

DICKEN, P. (1998): Global shift. Transforming the world economy. 3. Aufl. New York, London: Guilford Press.

DIENG, A. B.; RUIZ, F. J. R. & BLEI, D. M. (2020): Topic Modeling in Embedding Spaces. In: Transactions of the Association for Computational Linguistics 8: 439–453.

DUMM, S. & NIEKLER, A. (2016): Methoden, Qualitätssicherung und Forschungsdesign. In: LEMKE, M. & WIEDEMANN, G. (Hrsg.): Text Mining in den Sozialwissenschaften. Wiesbaden: Springer VS: 89–116.

DUSHYANT, M.; RATHOD, B. & KHANA, S. (2013): A Review on Emerging Trends of Web Mining and It's Applications. In: International Journal of Engineering, Development and Research 8: 98–102.

EINAV, L. & LEVIN, J. (2014): The Data Revolution and Economic Analysis. In: Innovation Policy and the Economy 14(1): 1–24.

EKBIA, H.; MATTIOLI, M.; KOUPER, I.; ARAVE, G.; GHAZINEJAD, A.; BOWMAN, T.; SURI, V. R.; TSOU, A.; WEINGART, S. & SUGIMOTO, C. R. (2015): Big data, bigger dilemmas: A critical review. In: Journal of the Association for Information Science and Technology 66(8): 1523–1545.

ELMAN, J. L. (1990): Finding Structure in Time. In: Cognitive Science 14(2): 179–211.

ESPEHOLT, L.; AGRAWAL, S.; SØNDERBY, C.; KUMAR, M.; HEEK, J.; BROMBERG, C.; GAZEN, C.; HICKEY, J.; BELL, A. & KALCHBRENNER, N. (2021): Skillful Twelve Hour Precipitation Forecasts using Large Context Neural Networks. In: arXiv preprint arXiv:2111.07470: 1–34.

ETZKOWITZ, H. & LEYDESDORFF, L. (1997): Universities and the Global Knowledge Economy: A Triple Helix of University-Industry Relations.

EUROSTAT (2022): Digitale Wirtschaft und Gesellschaft. ec.europa.eu/eurostat/de/web/digital-economy-and-society/data/database (25.08.2022).

FAN, J.; SAMWORTH, R. & WU, Y. (2009): Ultrahigh dimensional feature selection: beyond the linear model. In: The Journal of Machine Learning Research 10: 2013–2038.

FELDMAN, M. (2003): The Locational Dynamics of the U.S. Biotech Industry: Knowledge Externalities and the Anchor Hypothesis. In: Industry and Innovation 10(3): 201–224.

FELDMAN, M.; KENNEY, M. & LISSONI, F. (2015): The New Data Frontier. In: Research Policy 44(9): 1629–1632.

FELDMAN, M. & LOWE, N. (2015): Triangulating regional economies: Realizing the promise of digital data. In: Research Policy 44(9): 1785–1793.

FELDMAN, S. (Hrsg.) (2004): The thirteenth International World Wide Web Conference alternate track papers & posters. New York: Association for Computing Machinery.

FENG, Z.; GUO, D.; TANG, D.; DUAN, N.; FENG, X.; GONG, M.; SHOU, L.; QIN, B.; LIU, T.; JIANG, D. & ZHOU, M. (2020): CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In: arXiv preprint arXiv:2002.08155: 1–12.

FÉVOTTE, C. & IDIER, J. (2011): Algorithms for Nonnegative Matrix Factorization with the β-Divergence. In: Neural Computation 23(9): 2421–2456.

FIROOZEH, N.; NAZARENKO, A.; ALIZON, F. & DAILLE, B. (2020): Keyword extraction: Issues and methods. In: Natural Language Engineering 26(3): 259–291.

FLICK, U. (2011): Triangulation. Eine Einführung. Wiesbaden: Springer.

FLORIDI, L. (2012): Big Data and Their Epistemological Challenge. In: Philosophy & Technology 25(4): 435–437.

FRENKEN, K.; VAN OORT, F. & VERBURG, T. (2007): Related Variety, Unrelated Variety and Regional Economic Growth. In: Regional Studies 41(5): 685–697.

FRIEDL, J. E. F. (2006): Mastering Regular Expressions. Peking, Cambridge, Farnham, Köln, Paris, Sebastol, Taipei, Tokyo: O'Reilly Media, Inc.

FROMHOLD-EISEBITH, M. & WERKER, C. (2013): Universities' functions in knowledge transfer: a geographical perspective. In: The Annals of Regional Science 51(3): 621–643.

FUJII, H. & MANAGI, S. (2018): Trends and priority shifts in artificial intelligence technology invention: A global patent analysis. In: Economic Analysis and Policy 58: 60–69.

FULLER, M. (2008): Software studies. A lexicon. Leonardo books. Cambridge, Mass.: MIT Press.

GANGULI, D.; HERNANDEZ, D.; LOVITT, L.; ASKELL, A.; BAI, Y.; CHEN, A.; CONERLY, T.; DASSARMA, N.; DRAIN, D.; ELHAGE, N.; EL SHOWK, S.; FORT, S.; HATFIELD-DODDS, Z.; HENIGHAN, T.; JOHNSTON, S.; JONES, A.; JOSEPH, N.; KERNIAN, J.; KRAVEC, S.; MANN, B.; NANDA, N.; NDOUSSE, K.; OLSSON, C.; AMODEI, D.; BROWN, T.; KAPLAN, J.; MCCANDLISH, S.; OLAH, C.; AMODEI, D. & CLARK, J. (2022): Predictability and Surprise in Large Generative Models. In: ACM Conference on Fairness, Accountability, and Transparency (Hrsg.): 2022 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM: 1747–1764.

GARECHANA, G.; RÍO-BELVER, R.; BILDOSOLA, I. & SALVADOR, M. R. (2017): Effects of innovation management system standardization on firms: evidence from text mining annual reports. In: Scientometrics 111(3): 1987–1999.

GEHRING, J.; AULI, M.; GRANGIER, D.; YARATS, D. & N. DAUPHIN, Y. (2017): Convolutional Sequence to Sequence Learning. In: International Conference on Machine Learning: 1243–1252.

GENTZKOW, M.; KELLY, B. & TADDY, M. (2019): Text as Data. In: Journal of Economic Literature 57(3): 535–574.

GÉRON, A. (2022): Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3. Aufl. Zürich: O'Reilly Media, Incorporated.

GINSBERG, J.; MOHEBBI, M. H.; PATEL, R. S.; BRAMMER, L.; SMOLINSKI, M. S. & BRILLIANT, L. (2009): Detecting influenza epidemics using search engine query data. In: Nature 457(7232): 1012–1014.

GLAESER, E. L.; KERR, W. R. & PONZETTO, G. A. (2010): Clusters of entrepreneurship. In: Journal of Urban Economics 67(1): 150–168.

GLENISSON, P.; GLÄNZEL, W.; JANSSENS, F. & MOOR, B. (2005): Combining full text and bibliometric information in mapping scientific disciplines. In: Information Processing & Management 41(6): 1548–1572.

GÖK, A.; WATERWORTH, A. & SHAPIRA, P. (2015): Use of web mining in studying innovation. In: Scientometrics 102(1): 653–671.

GOLDBERG, Y. (2017a): A Primer on Neural Network Models for Natural Language Processing. In: Journal of Artificial Intelligence Research 57: 345–420.

GOLDBERG, Y. (2017b): Neural Network Methods for Natural Language Processing 10. Toronto: Morgan & Claypool Publishers.

GONZALEZ, A. G. (2006): The software patent debate. In: Journal of Intellectual Property Law & Practice 1(3): 196–206.

GOODCHILD, M. F. (2013): The quality of big (geo)data. In: Dialogues in Human Geography 3(3): 280–284.

GOODFELLOW, I.; BENGIO, Y. & COURVILLE, A. (2016): Deep Learning. Cambridge: MIT Press.

GRAUPE, D. (2013): Principles of artificial neural networks. Advanced series in circuits and systems v. 7. 3. Aufl. Singapur, Hackensack: World Scientific Pub. Co.

GRAVE, E.; BOJANOWSKI, P.; GUPTA, P.; JOULIN, A. & MIKOLOV, T. (2018): Learning Word Vectors for 157 Languages. In: arXiv preprint arXiv:1802.06893: 1–5.

GRAVES, A. & SCHMIDHUBER, J. (2005): Framewise phoneme classification with bidirectional LSTM and other neural network architectures. In: Neural networks the official journal of the International Neural Network Society 18(5-6): 602–610.

GREENACRE, Z. A. (2016): The Importance of Selection Bias in Internet Surveys. In: Open Journal of Statistics 6(3): 397–404.

GRIFFITHS, T. L. & STEYVERS, M. (2004): Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America: 5228–5235.

GROOTENDORST, M. (2021): MaartenGr/KeyBERT: BibTeX. Zenodo.

GROOTENDORST, M. (2022): BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv.

GROOTENDORST, M. P. (2023): FAQ - BERTopic. maartengr.github.io/BERTopic/faq.html (11.01.2023).

HÄDER, M. (2010): Empirische Sozialforschung. Wiesbaden: VS Verlag für Sozialwissenschaften.

HAIN, D.; JUROWETZKI, R. & SQUICCIARINI, M. (2022): Mapping Complex Technologies via Science-Technology Linkages; The Case of Neuroscience -- A transformer based keyword extraction approach. In: arXiv preprint arXiv:2205.10153: 1–33.

HAKLAY, M. (2010): How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. In: Environment and Planning B: Planning and Design 37(4): 682–703.

HALL, B. H. (2022): Patents, innovation, and development. In: International Review of Applied Economics: 1–26.

HAN, E.-H.; KARYPIS, G. & KUMAR, V. (2001): Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. In: CARBONELL, J. G. & SIEKMANN, J. (Hrsg.): Knowledge Discovery and Data Mining - PAKDD 2001: 5th Asia-Pacific Conference, Hong Kong, China, April 16-18, 2001: Proceedings. Lecture Notes in Artificial Intelligence 2035. New York: Springer: 53–65.

HARRIS, Z. S. (1954): Distributional Structure. In: Word 10(2-3): 146–162.

HAUNSCHILD, R.; BORNMANN, L. & MARX, W. (2016): Climate Change Research in View of Bibliometrics. In: PLOS ONE 11(7): 0160393.

HE, P.; LIU, X.; GAO, J. & CHEN, W. (2021): DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In: arXiv preprint arXiv: 2006.03654: 1–23.

HELBICH, M. (2012): Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata. In: Proceedings of GI_Forum: 24–33.

HELLER & PARTNER (2022): Werksviertel – werksviertel münchen. werksviertel.de/ (17.10.2022).

HELLMANZIK, C. & SCHMITZ, M. (2017): Taking gravity online: The role of virtual proximity in international finance. In: Journal of International Money and Finance 77: 164–179.

HELSLEY, R. W. & STRANGE, W. C. (2002): Innovation and Input Sharing. In: Journal of Urban Economics 51(1): 25–45.

HERNÁNDEZ, B.; JIMÉNEZ, J. & MARTÍN, M. J. (2009): Key website factors in e-business strategy. In: International Journal of Information Management 29(5): 362–371.

HÉROUX-VAILLANCOURT, M.; BEAUDRY, C. & RIETSCH, C. (2020): Using web content analysis to create innovation indicators—What do we really measure? In: Quantitative Science Studies 1(4): 1601–1637.

HIDALGO, C. A.; BALLAND, P.-A.; BOSCHMA, R.; DELGADO, M.; FELDMAN, M.; FRENKEN, K.; GLAESER, E.; HE, C.; KOGLER, D. F.; MORRISON, A.; NEFFKE, F.; RIGBY, D.; STERN, S.; ZHENG, S. & ZHU, S. (2018): The Principle of Relatedness. In: MORALES, A. J.; GERSHENSON, C.; BRAHA, D.; MINAI, A. A. & BAR-YAM, Y. (Hrsg.): Unifying Themes in Complex Systems IX: Proceedings of the Ninth International Conference on Complex Systems. Springer Proceedings in Complexity. Cham: Springer International Publishing; Imprint: Springer: 451–457.

HIDALGO, C. A. & HAUSMANN, R. (2009): The building blocks of economic complexity. In: Proceedings of the National Academy of Sciences of the United States of America 106(26): 10570–10575.

HINTON, G. E. (1992): How neural networks learn from experience. In: Scientific American 267(3): 144–151.

HIRSCHBERG, J. & MANNING, C. D. (2015): Advances in natural language processing. In: Science 349(6245): 261–266.

HOCHREITER, S. (1998): The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6(2): 107–116.

HOCHREITER, S. & SCHMIDHUBER, J. (1997): Long short-term memory. In: Neural Computation 9(8): 1735–1780.

HOU, Y. & WANG, Q. (2021): A bibliometric study about energy, environment, and climate change. In: Environmental Science and Pollution Research 28(26): 34187–34199.

HOWARD, J. & RUDER, S. (2018): Universal Language Model Fine-tuning for Text Classification. In: arXiv preprint arXiv:1801.06146: 1–12.

HOWELLS, J. R. L. (2002): Tacit Knowledge, Innovation and Economic Geography. In: Urban Studies 39(5-6): 871–884.

HU, Z.; FANG, S. & LIANG, T. (2014): Empirical study of constructing a knowledge organization system of patent documents using topic modeling. In: Scientometrics 100(3): 787–799.

HUANG, L.; CHEN, K. & ZHOU, M. (2020): Climate change and carbon sink: a bibliometric analysis. In: Environmental Science and Pollution Research 27(8): 8740–8758.

HUGGING FACE (o.J.): all-mpnet-base-v2. huggingface.co/sentence-transformers/all-mpnet-base-v2 (05.10.2022).

HUGGING FACE (2022a): Models. huggingface.co/models (09.03.2022).

HUGGING FACE (2022b): xlm-roberta-base. huggingface.co/xlm-roberta-base (19.10.2022).

IGLESIAS, J. A.; TIEMBLO, A.; LEDEZMA, A. & SANCHIS, A. (2016): Web news mining in an evolving framework. In: Information Fusion 28: 90–98.

JAHANBIN, K. & RAHMANIAN, V. (2020): Using twitter and web news mining to predict COVID-19 outbreak. In: Asian Pacific Journal of Tropical Medicine 13(8): 378.

JAIN, S. & WALLACE, B. C. (2019): Attention is not Explanation. In: arXiv preprint arXiv:1902.10186: 1–16.

JIANG, H.; QIANG, M. & LIN, P. (2016): A topic modeling based bibliometric exploration of hydropower research. In: Renewable and Sustainable Energy Reviews 57: 226–237.

JIANG, S.; PANG, G.; WU, M. & KUANG, L. (2012): An improved K-nearest-neighbor algorithm for text categorization. In: Expert Systems with Applications 39(1): 1503–1509.

JIVANI, A. G. (2011): A comparative study of stemming algorithms. In: International Journal of Computer Applications in Technology 2(6): 1930–1938.

JOACHIMS, T. (1998): Text categorization with Support Vector Machines: Learning with many relevant features. In: NÉDELLEC, C. & ROUVEIROL, C. (Hrsg.): Machine Learning: ECML-98: ECML 1998. Lecture notes in computer science. Berlin, Heidelberg: Springer: 137–142.

JONES, K. S. (1994): Natural Language Processing: A Historical Review. In: Current issues in computational linguistics: in honour of Don Walker: 3–16.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P. & MIKOLOV, T. (2016): Bag of Tricks for Efficient Text Classification. In: arXiv preprint arXiv: 1607.01759: 1.5.

JOZEFOWICZ, R.; VINYALS, O.; SCHUSTER, M.; SHAZEER, N. & WU, Y. (2016a): Exploring the Limits of Language Modeling. In: arXiv preprint arXiv:1602.02410: 1–11.

JOZEFOWICZ, R.; VINYALS, O.; SCHUSTER, M.; SHAZEER, N. & WU, Y. (2016b): Exploring the Limits of Language Modeling.

JOZEFOWICZ, R.; ZAREMBA, W. & SUTSKEVER, I. (2015): An Empirical Exploration of Recurrent Network Architectures. In: BACH, F. & BLEI, D. (Hrsg.): Proceedings of the International Conference on International Conference on Machine Learning. New York: ACM: 2342–2350.

JUMPER, J.; EVANS, R.; PRITZEL, A.; GREEN, T.; FIGURNOV, M.; RONNEBERGER, O.; TUNYASUVUNAKOOL, K.; BATES, R.; ŽÍDEK, A.; POTAPENKO, A.; BRIDGLAND, A.; MEYER, C.; KOHL, S. A. A.; BALLARD, A. J.; COWIE, A.; ROMERA-PAREDES, B.; NIKOLOV, S.; JAIN, R.; ADLER, J.; BACK, T.; PETERSEN, S.; REIMAN, D.;

CLANCY, E.; ZIELINSKI, M.; STEINEGGER, M.; PACHOLSKA, M.; BERGHAMMER, T.; BODENSTEIN, S.; SILVER, D.; VINYALS, O.; SENIOR, A. W.; KAVUKCUOGLU, K.; KOHLI, P. & HASSABIS, D. (2021): Highly accurate protein structure prediction with AlphaFold. In: Nature 596(7873): 583–589.

JUN, S.-P.; YOO, H. S. & CHOI, S. (2018): Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. In: Technological Forecasting and Social Change 130: 69–87.

JURAFSKY, D. & MARTIN, J. H. (2019): Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3. Aufl. Stanford: Prentice Hall.

JUROWETZKI, R.; HAIN, D.; MATEOS-GARCIA, J. & STATHOULOPOULOS, K. (2021): The Privatization of AI Research(-ers): Causes and Potential Consequences -- From university-industry interaction to public research brain-drain? In: arXiv preprint arXiv:2102.01648: 1–36.

KATZ, J. S. & COTHEY, V. (2006): Web indicators for complex innovation systems. In: Research Evaluation 15(2): 85–95.

KAVVADIAS, S.; DROSATOS, G. & KALDOUDI, E. (2020): Supporting topic modeling and trends analysis in biomedical literature. In: Journal of Biomedical Informatics 110: 103574.

KAYSER, V. & BLIND, K. (2017): Extending the knowledge base of foresight: The contribution of text mining. In: Technological Forecasting and Social Change 116: 208–215.

KELLE, U. (2014): Mixed Methods. In: BAUR, N. & BLASIUS, J. (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer: 153–166.

KENEKAYORO, P.; BUCKLEY, K. & THELWALL, M. (2014): Automatic classification of academic web page types. In: Scientometrics 101(2): 1015–1026.

KERR, W. R. & ROBERT-NICOUD, F. (2020): Tech Clusters. In: Journal of Economic Perspectives 34(3): 50–76.

KESKAR, N. S.; MCCANN, B.; VARSHNEY, L. R.; XIONG, C. & SOCHER, R. (2019): CTRL: A Conditional Transformer Language Model for Controllable Generation. In: arXiv preprint arXiv:1909.05858: 1–18.

KEßLER, C. (2017): Extracting central places from the link structure in Wikipedia. In: Transactions in GIS 21(3): 488–502.

KHURANA, D.; KOLI, A.; KHATTER, K. & SINGH, S. (2022): Natural language processing: state of the art, current trends and challenges. In: Multimedia Tools and Applications 82(3): 3713–3744.

KIM, J. Y. & CHOONG, K. L. (2016): An Empirical Analysis of Requirements for Data Scientists Using Online Job Postings. In: International Journal of Software Engineering and Its Applications 10(4): 161–172.

KIM, S.-B.; HAN, K.-S.; RIM, H.-C. & MYAENG, S. H. (2006): Some Effective Techniques for Naive Bayes Text Classification. In: IEEE Transactions on Knowledge and Data Engineering 18(11): 1457–1466.

KIM, T. J. (2020): COVID-19 news analysis using news big data: Focusing on topic modeling analysis. In: the Journal of the Korea Contents Association 20(4): 457–466.

KINNE, J. & AXENBECK, J. (2020): Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. In: Scientometrics 125(3): 2011–2041.

KINNE, J. & LENZ, D. (2021): Predicting innovative firms using web mining and deep learning. In: PLOS ONE 16(4): e0249071.

KINNE, J. & RESCH, B. (2018): Analyzing and Predicting Micro-Location Patterns of Software Firms. In: ISPRS International Journal of Geo-Information 7(1): 1–21.

KITCHIN, R. (2013): Big data and human geography. In: Dialogues in Human Geography 3(3): 262–267.

KITCHIN, R. (2014): Big Data, new epistemologies and paradigm shifts. In: Big Data & Society 1(1): 1-12.

KLINGER, J.; MATEOS-GARCIA, J. & STATHOULOPOULOS, K. (2018): Deep learning, deep change? Mapping the development of the Artificial Intelligence General Purpose Technology. In: arXiv preprint arXiv:1808.06355: 1–26.

KOEHN, P. (2020): Neural machine translation. 1. Aufl. New York: Cambridge University Press.

KOSALA, R. & BLOCKEEL, H. (2000): Web mining research. In: ACM SIGKDD Explorations Newsletter 2(1): 1–15.

KOWSARI; MEIMANDI, J.; HEIDARYSAFA; MENDU; BARNES & BROWN (2019): Text Classification Algorithms: A Survey. In: Information 10(4): 1–69.

KRÜGER, M.; KINNE, J.; LENZ, D. & RESCH, B. (2020): The Digital Layer: How Innovative Firms Relate on the Web. In: ZEW-Centre for European Economic Research Discussion Paper(20-003): 1-14.

KRÜGER, R. (2021): Die Transformer-Architektur für Systeme zur neuronalen maschinellen Übersetzung -eine popularisierende Darstellung. In: trans-kom 14(2): 278–324.

KUCKARTZ, U. (2010): Einführung in die computergestützte Analyse qualitativer Daten. Wiesbaden: VS Verlag für Sozialwissenschaften.

KUCKARTZ, U. (2014): Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren. Wiesbaden: Springer.

KÜHNEMANN, H. (2021): Anwendungen des Web Scraping in der amtlichen Statistik. In: AStA Wirtschafts- und Sozialstatistisches Archiv 15(1): 5–25.

KUMAR, S. & KUMAR, R. (2021): A Study on Different Aspects of Web Mining and Research Issues. In: IOP Conference Series: Materials Science and Engineering 1022(1): 1–10.

LAMPERT, T.; KROLL, L. E. & DUNKELBERG, A. (2007): Soziale Ungleichheit der Lebenserwartung in Deutschland. Robert Koch-Institut.

LAMPLE, G.; BALLESTEROS, M.; SUBRAMANIAN, S.; KAWAKAMI, K. & DYER, C. (2016): Neural Architectures for Named Entity Recognition. In: arXiv preprint arXiv: 1603.01360: 1–11.

LANEY, D. (2001): 3D data management: Controlling data volume, velocity and variety. In: META group research note: 1–12.

LARSEN, V. H. & THORSRUD, L. A. (2019): The value of news for economic developments. In: Journal of Econometrics 210(1): 203–218.

LE, Q. & MIKOLOV, T. (2014): Distributed Representations of Sentences and Documents. In: International Conference on Machine Learning 32(2): 1188–1196.

LE, V. Q.; JAITLY, N. & HINTON, G. E. (2015): A Simple Way to Initialize Recurrent Networks of Rectified Linear Units. In: arXiv preprint arXiv:1504.00941: 1–7.

LECUN, Y.; BENGIO, Y. & HINTON, G. (2015): Deep learning. In: Nature 521(7553): 436–444.

LEE, J.; YOON, W.; KIM, S.; KIM, D.; KIM, S.; SO, C. H. & KANG, J. (2020): BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In: Bioinformatics 36(4): 1234–1240.

LEMKE, M. & WIEDEMANN, G. (2016): Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse. Wiesbaden: Springer.

LENZ, D. & WINKER, P. (2020): Measuring the diffusion of innovations with paragraph vector topic models. In: PLOS ONE 15(1): e0226685.

LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A. & CHENG, T. (2016): Geospatial big data handling theory and methods: A review and research challenges. In: ISPRS Journal of Photogrammetry and Remote Sensing 115: 119–133.

LI, Y.; ARORA, S.; YOUTIE, J. & SHAPIRA, P. (2018): Using web mining to explore Triple Helix influences on growth in small and mid-size firms. In: Technovation 76-77: 3–14.

LIU, J.; LI, J.; LI, W. & WU, J. (2016a): Rethinking big data: A review on the data quality and usage issues. In: ISPRS Journal of Photogrammetry and Remote Sensing 115: 134–142.

LIU, L.; TANG, L.; DONG, W.; YAO, S. & ZHOU, W. (2016b): An overview of topic modeling and its current applications in bioinformatics. In: SpringerPlus 5(1): 1608–1630.

LIU, P.; QIU, X. & HUANG, X. (2016c): Recurrent Neural Network for Text Classification with Multi-Task Learning. arXiv.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L. & STOYANOV, V. (2019): RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: arXiv preprint arXiv: 1907.11692: 1–13.

LOPEZ, A. (2008): Statistical machine translation. In: ACM Computing Surveys 40(3): 1–49.

LOUREIRO, D.; BARBIERI, F.; NEVES, L.; ANKE, L. E. & CAMACHO-COLLADOS, J. (2022): TimeLMs: Diachronic Language Models from Twitter. In: arXiv preprint arXiv:2202.03829: 1–10.

LUCAS, R. E. (1988): On the mechanics of economic development. In: Journal of Monetary Economics 22(1): 3–42.

LUND, K. & BURGESS, C. (1996): Producing high-dimensional semantic spaces from lexical co-occurrence. In: Behavior Research Methods, Instruments, & Computers 28(2): 203–208.

LUNDVALL, B.-Å. (2012): National Systems of Innovation. Toward a Theory of Innovation and Interactive Learning. Anthem Other Canon series. Cambridge: Cambridge University Press.

MA, D.; SANDBERG, M. & JIANG, B. (2015): Characterizing the Heterogeneity of the OpenStreetMap Data and Community. In: ISPRS International Journal of Geo-Information 4(2): 535–550.

MACKENBACH, J. (2006): Health inequalities: Europe in profile. Rotterdam: Erasmus MC.

MADELIN, M.; GRASLAND, C.; MATHIAN, H.; SANDERS, L. & VINCENT, J.-M. (2009): Das "MAUP": Modifiable Areal Unit - Problem oder Fortschritt? In: Informationen zur Raumentwicklung 10(11): 645–660.

MAGHDID, H. S. (2019): Web News Mining Using New Features: A Comparative Study. In: IEEE Access 7: 5626–5641.

MANDERSCHEID, K. (2019): Text Mining. In: Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer: 1103–1116.

MANNING, C.; SURDEANU, M.; BAUER, J.; FINKEL, J.; BETHARD, S. & MCCLOSKY, D. (2014): The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of

the Association for Computational Linguistics: System Demonstrations. Stroudsburg, PA, USA: Association for Computational Linguistics.

MARSHALL, A. (1890): Principles of Economics. London: Macmillan.

MAYERL, J. (2015): Bedeutet ‚Big Data' das Ende der sozialwissenschaftlichen Methodenforschung? https://soziopolis. de/beobachten/wissenschaft/artikel/bedeutet-big-data-das-ende-der-sozialwissenschaftlichen-methodenforschung (2015) (05.09.2022).

MAYRING, P. (2010): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 12. überarbeitete Aufl. Weinheim: Beltz.

MCCANN, B.; KESKAR, N. S.; XIONG, C. & SOCHER, R. (2018): The Natural Language Decathlon: Multitask Learning as Question Answering. In: arXiv preprint arXiv:1806.08730: 1–23.

MCINNES, L.; HEALY, J. & MELVILLE, J. (2020): UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. In: arXiv preprint arXiv: 1802.03426: 1–63.

MEDAD, A.; GAIO, M.; MONCLA, L.; MUSTIÈRE, S. & LE NIR, Y. (2020): Comparing supervised learning algorithms for Spatial Nominal Entity recognition. In: AGILE: GIScience Series 1: 1–18.

MEIJERS, E. & PERIS, A. (2019): Using toponym co-occurrences to measure relationships between places: review, application and evaluation. In: International Journal of Urban Sciences 23(2): 246–268.

MELO, T. de & FIGUEIREDO, C. M. S. (2021): Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach. In: JMIR Public Health and Surveillance 7(2): e24585.

MENNELL, S. (1975): Ethnomethodology and the new Methodenstreit. In: Acta Sociologica 18(4): 287–302.

MIKOLOV, T.; CHEN, K.; CORRADO, G. & DEAN, J. (2013a): Efficient Estimation of Word Representations in Vector Space. In: arXiv preprint arXiv: 1301.3781: 1–12.

MIKOLOV, T.; JOULIN, A.; CHOPRA, S.; MATHIEU, M. & RANZATO, M. (2015): Learning Longer Memory in Recurrent Neural Networks. In: arXiv preprint arXiv:1412.7753: 1–7.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. & DEAN, J. (2013b): Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 1–9.

MILLER, H. J. (2010): The data avalanche is here. Shouldn't we be digging. In: Journal of Regional Science 50(1): 181–201.

MILLER, H. J. & GOODCHILD, M. F. (2015): Data-driven geography. In: GeoJournal 80(4): 449–461.

MINER, G.; JOHN ELDER, IV; FAST, A.; HILL, T.; NISBET, R. & DELEN, D. (2012): Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Cambridge: Academic Press.

MIRTSCH, M.; KINNE, J. & BLIND, K. (2021): Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis. In: IEEE Transactions on Engineering Management 68(1): 87–100.

MITCHELL, R. (2018): Web scraping with Python. Collecting more data from the modern web. Second edition. Peking, Boston, Farnham, Tokio: O'Reilly.

MIZUNO, T.; OHNISHI, T. & WATANABE, T. (2017): Novel and topical business news and their impact on stock market activity. In: EPJ Data Science 6(1): 1–14.

MONTANI, I.; HONNIBAL, M. & VAN LANDEGHEM, S. (2022): explosion/spaCy: v3.4.1: Fix compatibility with CuPy v9.x. Zenodo.

MOODY, C. E. (2016): Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. In: arXiv preprint arXiv:1605.02019: 1–16.

MOONEY, P. & MINGHINI, M. (2017): A review of OpenStreetMap data. In: FOODY, G.; SEE, L.; FRITZ, S.; MOONEY, P.; OLTEANU-RAIMOND, A.-M.; Costa Fonte; Cidalia & ANTONIOU, V. (Hrsg.): Mapping and the citizen sensor. London: Ubiquity Press: 37–59.

MORAN, P. A. P. (1950): Notes on Continuous Stochastic Phenomena. In: Biometrika 37(1/2): 17–24.

MORETTI, E. (2012): The new geography of jobs. Boston, New York: Mariner Books; Houghton Mifflin Harcourt.

MORETTI, F. (2000): Conjectures on world literature. In: New left review 54(1): 1–27.

MUNZERT, S. (2014): Big Data in der Forschung! Big Data in der Lehre? Ein Vorschlag zur Erweiterung der bestehenden Methodenausbildung. In: Zeitschrift für Politikwissenschaft 24(1-2): 205–220.

MUNZERT, S. (2018): Auf dem Weg zu einer fundierten Softwareausbildung in der Politikwissenschaft. In: BLÄTTE, A.; BEHNKE, J.; SCHNAPP, K.-U. & WAGEMANN, C. (Hrsg.): Computational Social Science: Die Analyse von Big Data. Schriftenreihe der Sektion Methoden der Politikwissenschaft der Deutschen Vereinigung für Politikwissenschaft. Baden-Baden: Nomos: 379–403.

MUNZERT, S. & NYHUIS, D. (2020): Die Nutzung von Webdaten in den Sozialwissenschaften. In: WAGEMANN, C.; GOERRES, A. & SIEWERT, M. B. (Hrsg.): Handbuch Methoden der Politikwissenschaft. Wiesbaden: Springer VS: 373–397.

MUTHUKADAN, B. (2022): Selenium with Python — Selenium Python Bindings 2 documentation. selenium-python.readthedocs.io/ (29.11.2022).

NAGEL, S. (2021): From Web Graphs to Prioritizing Web Crawls. indico.cern.ch/event/1006978/contributions/4539477/attachments/2325769/3962907/os-sym2021-sn-web-graphs-crawling.pdf (01.12.2022).

NANDY, A.; DUAN, C.; GOFFINET, C. & KULIK, H. J. (2022): New Strategies for Direct Methane-to-Methanol Conversion from Active Learning Exploration of 16 Million Catalysts. In: JACS Au 2(5): 1200–1213.

NEFFKE, F.; HENNING, M. & BOSCHMA, R. (2011): How do regions diversify over time? Industry relatedness and the developement of new groth paths in regions. In: Economic geography 87(3): 237–265.

NOMINATIM (2022): Features. www.nominatim.org/ (12.09.2022).

NOTHMAN, J.; RINGLAND, N.; RADFORD, W.; MURPHY, T. & CURRAN, J. R. (2013): Learning multilingual named entity recognition from Wikipedia. In: Artificial Intelligence 194: 151–175.

OLAH, C. (2015): Understanding LSTM Networks. research.google/pubs/pub45500/ (12.07.2022).

OLSSON, F. (2009): A literature survey of active machine learning in the context of natural language processing. Kista: Swedish Institute of Computer Science.

OPENSHAW, S. (1984): The modifiable areal unit problem. In: Concepts and Techniques in Modern Geography 38(41): 60–69.

ORDUN, C.; PURUSHOTHAM, S. & RAFF, E. (2020): Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. In: arXiv preprint arXiv: 2005.03082: 1–19.

ORTEGA, J. L. & AGUILLO, I. F. (2008): Visualization of the Nordic academic web: Link analysis using social network tools. In: Information Processing & Management 44(4): 1624–1633.

PAGE, L.; BRIN, S.; MOTWANI, R. & WINOGRAD, T. (1999): The PageRank Citation Ranking: Bringing Order to the Web. ilpubs.stanford.edu/422 (16.01.2023).

PANICHELLA, A.; DIT, B.; OLIVETO, R.; DI PENTA, M.; POSHYNANYK, D. & LUCIA, A. (2013): How to effectively use topic models for software engineering tasks? An approach based on Genetic Algorithms. In: Institute of Electrical and Electronics Engineers (Hrsg.): 35th International Conference 2013. San Francisco: Institute of Electrical and Electronics Engineers: 1–10.

PAPAGIANNIDIS, S.; SEE-TO, E. W.; ASSIMAKOPOULOS, D. G. & YANG, Y. (2018): Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the Internet age? In: Computers & Operations Research 98: 355–366.

PARK, H. W. (2003): Hyperlink network analysis: A new method for the study of social structure on the web. In: Connections 25(1): 50–62.

PARKES, D. C. & WELLMAN, M. P. (2015): Economic reasoning and artificial intelligence. In: Science 349(6245): 267–272.

PENNINGTON, J.; SOCHER, R. & MANNING, C. (2014): Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics.

PETERS, M.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K. & ZETTLEMOYER, L. (2018): Deep Contextualized Word Representations. In: WALKER, M.; JI, H. & STENT, A. (Hrsg.): Proceedings of the 2018 Conference of the North American Chapter of. Stroudsburg, PA, USA: Association for Computational Linguistics: 2227–2237.

PETERS, M. E.; RUDER, S. & SMITH, N. A. (2019): To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In: arXiv preprint arXiv:1903.05987: 1–8.

PHILIPPS, A. (2018): Text Mining-Verfahren als Herausforderung für die rekonstruktive Sozialforschung. In: Sozialer Sinn 2: 367–387.

POLANYI, M. (2012): The Tacit Dimension. In: PRUSAK, L. (Hrsg.): Knowledge in Organisations. Hoboken: Taylor & Francis: 135–146.

PORTER, M. E. (1990): The competitive advantage of nations. New York: The Free Press.

PREIS, T.; MOAT, H. S. & STANLEY, H. E. (2013): Quantifying trading behavior in financial markets using Google Trends. In: Scientific Reports 3(1): 1–6.

PUCHINGER, C. (2016): Die Anwendung von Text Mining in den Sozialwissenschaften. In: LEMKE, M. & WIEDEMANN, G. (Hrsg.): Text Mining in den Sozialwissenschaften. Wiesbaden: Springer VS: 117–136.

QU, Z.; SONG, X.; ZHENG, S.; WANG, X.; SONG, X. & LI, Z. (2018): Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification. In: IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T. & SUTSKEVER, I. (2018): Improving language understanding with unsupervised learning.

RADINSKY, K. & HORVITZ, E. (2013): Mining the web to predict future events. In: Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13. New York: ACM Press.

RAE, J. W.; BORGEAUD, S.; CAI, T.; MILLICAN, K.; HOFFMANN, J.; SONG, F.; ASLANIDES, J.; HENDERSON, S.; RING, R.; YOUNG, S.; RUTHERFORD, E.; HENNIGAN, T.; MENICK, J.; CASSIRER, A.; POWELL, R.; VAN DEN DRIESSCHE, G.; HENDRICKS, L. A.; RAUH, M.; HUANG, P.-S.; GLAESE, A.; WELBL, J.; DATHATHRI, S.; HUANG, S.; UESATO, J.; MELLOR, J.; HIGGINS, I.; CRESWELL, A.; MCALEESE, N.; WU, A.; ELSEN, E.; JAYAKUMAR, S.; BUCHATSKAYA, E.; BUDDEN, D.; SUTHERLAND, E.; SIMONYAN, K.; PAGANINI, M.; SIFRE, L.; MARTENS, L.; LI, X. L.; KUNCORO, A.; NEMATZADEH, A.; GRIBOVSKAYA, E.; DONATO, D.; LAZARIDOU, A.; MENSCH, A.; LESPIAU, J.-B.; TSIMPOUKELLI, M.; GRIGOREV, N.; FRITZ, D.; SOTTIAUX, T.; PAJARSKAS, M.; POHLEN, T.; GONG, Z.; TOYAMA, D.; D'AUTUME, C. D. M.; LI, Y.; TERZI, T.; MIKULIK, V.; BABUSCHKIN, I.; CLARK, A.; CASAS, D. D. L.; GUY, A.; JONES, C.; BRADBURY, J.; JOHNSON, M.; HECHTMAN, B.; WEIDINGER, L.; GABRIEL, I.; ISAAC, W.; LOCKHART, E.; OSINDERO, S.; RIMELL, L.; DYER, C.; VINYALS, O.; AYOUB, K.; STANWAY, J.; BENNETT, L.; HASSABIS, D.; KAVUKCUOGLU, K. & IRVING, G. (2021): Scaling Language Models: Methods, Analysis & Insights from Training Gopher. In: arXiv preprint arXiv:2112.11446: 1–120.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S. & MATENA, M., … & LIU, P. J. (2019): Exploring the limits of transfer learning with a unified text-to-text transformer. In: arXiv preprint arXiv:1910.10683: 1–67.

RAISI, H.; BAGGIO, R.; BARRATT-PUGH, L. & WILLSON, G. (2018): Hyperlink Network Analysis of a Tourism Destination. In: Journal of Travel Research 57(5): 671–686.

RAJ, M. & SEAMANS, R. (2019): Artificial intelligence, labor, productivity, and the need for firm-level data. In: GOLDFARB, A.; GANS, J. & AGRAWAL, A. (Hrsg.): The economics of artificial intelligence: An agenda. Chicago: University of Chicago Press: 553–565.

RAMESH, A.; PAVLOV, M.; GOH, G.; GRAY, S.; VOSS, C.; RADFORD, A.; CHEN, M. & SUTSKEVER, I. (2021): Zero-Shot Text-to-Image Generation. In: International Conference on Machine Learning: 8821–8831.

RAMMER, C. & ES-SADKI, N. (2022): Using big data for generating firm-level innovation indicators: A literature review. In: ZEW Discussion Papers 22(007): 1–38.

RAMMER, C.; FERNÁNDEZ, G. P. & CZARNITZKI, D. (2022): Artificial intelligence and industrial innovation: Evidence from German firm-level data. In: Research Policy 51(7): 1–40.

RANAEI, S.; SUOMINEN, A. & DEDEHAYIR, O. (2017): A topic model analysis of science and technology linkages: A case study in pharmaceutical industry. In: 2017 IEEE Technology & Engineering Management Conference (TEMSCON). IEEE.

RASCHKA, S. (2020): Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. In: arXiv preprint arXiv:1811.12808: 1–49.

RASCHKA, S.; PATTERSON, J. & NOLET, C. (2020): Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. In: Information 11(4): 1–44.

RAVURI, S.; LENC, K.; WILLSON, M.; KANGIN, D.; LAM, R.; MIROWSKI, P.; FITZSIMONS, M.; ATHANASSIADOU, M.; KASHEM, S.; MADGE, S.; PRUDDEN, R.; MANDHANE, A.; CLARK, A.; BROCK, A.; SIMONYAN, K.; HADSELL, R.; ROBINSON, N.; CLANCY, E.; ARRIBAS, A. & MOHAMED, S. (2021): Skilful precipitation nowcasting using deep generative models of radar. In: Nature 597(7878): 672–677.

REIMERS, N. & GUREVYCH, I. (2019): Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: arXiv preprint arXiv:1908.10084: 1–11.

RICHARDSON, L. (2022): Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation. www.crummy.com/software/BeautifulSoup/bs4/doc/ (29.11.2022).

RODRIGUEZ, P. L. & SPIRLING, A. (2022): Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. In: The Journal of Politics 84(1): 101–115.

ROMER, P. M. (1990): Endogenous Technological Change. In: Journal of Political Economy 98(5): 71-102.

ROUSIDIS, D.; KOUKARAS, P. & TJORTJIS, C. (2020): Social media prediction: a literature review. In: Multimedia Tools and Applications 79(9-10): 6279–6311.

RUDER, S. (2019): Neural transfer learning for natural language processing. Dissertation. Galway.

RUDER, S.; PETERS, M. E.; SWAYAMDIPTA, S. & WOLF, T. (2019a): Transfer Learning in Natural Language Processing. In: Proceedings of the 2019 Conference of the North. Stroudsburg, PA, USA: Association for Computational Linguistics.

RUDER, S.; SØGAARD, A. & VULIĆ, I. (2019b): Unsupervised Cross-Lingual Representation Learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. Stroudsburg, PA, USA: Association for Computational Linguistics.

RÜDIGER, M.; ANTONS, D. & SALGE, T. O. (2017): From Text to Data: On The Role and Effect of Text Pre-Processing in Text Mining Research. In: Academy of Management Proceedings 2017(1): 16353.

RUMELHART, D. E.; MCCLELLAND, J. & PDP RESEARCH GROUP (1988): Parallel distributed processing 1. New York: IEEE.

SAFAVIAN, S. R. & LANDGREBE, D. (1991): A survey of decision tree classifier methodology. In: IEEE Transactions on Systems, Man, and Cybernetics 21(3): 660–674.

SAMUEL, M. O.; TOLULOPE, A. I. & OYEJOKE, O. O. (2019): A Systematic Review of Current Trends in Web Content Mining 1299(1): 1–12.

SCHMID, H. (1999): Improvements in Part-of-Speech Tagging with an Application to German. In: ARMSTRONG, S.; CHURCH, K.; ISABELLE, P.; MANZI, S.; TZOUKERMANN, E. & Yarowsky David (Hrsg.): Natural Language Processing Using Very Large Corpora. Dordrecht: Springer: 13–25.

SCHMIDHUBER, J. (2015): Deep learning in neural networks: an overview. In: Neural networks the official journal of the International Neural Network Society 61: 85–117.

SCHNAPP, K.-U. & BLÄTTE, A. (2018): Epistemologische, methodische und politische Herausforderungen von Big Data. In: BLÄTTE, A.; BEHNKE, J.; SCHNAPP, K.-U. & WAGEMANN, C. (Hrsg.): Computational Social Science: Die Analyse von Big Data. Schriftenreihe der Sektion Methoden der Politikwissenschaft der Deutschen Vereinigung für Politikwissenschaft. Baden-Baden: Nomos: 25–52.

SCHUMPETER, J. (1939): Business Cycles. London.

SCHUSTER, M. & PALIWAL, K. K. (1997): Bidirectional recurrent neural networks. In: IEEE Transactions on Signal Processing 45(11): 2673–2681.

SCHÜTZE, H. (1992): Word space. In: Advances in Neural Information Processing Systems 5: 895–902.

SCHWIERZY, J.; DEHGHAN, R.; SCHMIDT, S.; RODEPETER, E.; STOEMMER, A.; UCTUM, K.; KINNE, J.; LENZ, D. & HOTTENROTT, H. (2022): Technology Mapping Using WebAI: The Case of 3D Printing. In: arXiv preprint arXiv:2201.01125: 1–16.

SCRAPY COMMUNITY (2022a): Generic Spiders — Scrapy 2.6.2 documentation. docs.scrapy.org/en/latest/topics/spiders.html#generic-spiders (19.09.2022).

SCRAPY COMMUNITY (2022b): Scrapy 2.7 documentation — Scrapy 2.7.1 documentation. docs.scrapy.org/en/latest/index.html (07.11.2022).

SENANAYAKE, U.; PIRAVEENAN, M. & ZOMAYA, A. (2015): The Pagerank-Index: Going beyond Citation Counts in Quantifying Scientific Impact of Researchers. In: PloS one 10(8): e0134794.

SEVILLA, J.; HEIM, L.; HO, A.; BESIROGLU, T.; HOBBHAHN, M. & VILLALOBOS, P. (2022): Compute Trends Across Three Eras of Machine Learning. In: arXiv preprint arXiv:2202.05924: 1–25.

SHARMA, S.; SHARMA, S. & ATHAIYA, A. (2020): Activation functions in neural networks. In: International Journal of Engineering Applied Sciences and Technology 4(12): 310–316.

SHAZEER, N.; CHENG, Y.; PARMAR, N.; TRAN, D.; VASWANI, A.; KOANANTAKOOL, P.; HAWKINS, P.; LEE, H.; HONG, M.; YOUNG, C.; SEPASSI, R. & HECHTMAN, B. (2018): Mesh-TensorFlow: Deep Learning for Supercomputers. In: BENGIO, S.; WALLACH, H.; LAROCHELLE, H.; GRAUMAN, K.; CESA-BIANCHI, N. & GARNETT, R. (Hrsg.): Advances in neural information processing systems: 32nd conference on neural information processing sys. La Jolla: Neural Information Processing Systems Foundation.

SHAZEER, N.; MIRHOSEINI, A.; MAZIARZ, K.; DAVIS, A.; LE QUOC; HINTON, G. & DEAN, J. (2017): Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In: arXiv preprint arXiv:1701.06538: 1–19.

SHIH, T.-H. & FAN, X. (2009): Comparing response rates in e-mail and paper surveys: A meta-analysis. In: Educational Research Review 4(1): 26–40.

SIMON, T. & KOLLER, D. (2001): Support vector machine active learning with applications to text classification. In: Journal of Machine Learning Research 2: 45–66.

SJØVAAG, H.; STAVELIN, E.; KARLSSON, M. & KAMMER, A. (2019): The Hyperlinked Scandinavian News Ecology. In: Digital Journalism 7(4): 507–531.

SPACY (2022): German · spaCy Models Documentation. spacy.io/models/de#de_core_news_lg (09.01.2023).

STADT HEIDELBERG (2018): Heidelberg Bahnstadt. Forschungsnahe und wissensbasierte Unternehmen und Hochschulen. www.heidelberg-bahnstadt.de/984262.html (15.09.2022).

STADT HEIDELBERG (2022a): Masterplan Neuenheimer Feld. Fakten zum Areal. www.masterplan-neuenheimer-feld.de/informationen/fakten-zum-areal (15.09.2022).

STADT HEIDELBERG (2022b): Masterplan Neuenheimer Feld. Faktencheck. www.masterplan-neuenheimer-feld.de/informationen/faktencheck (15.09.2022).

STADT HEIDELBERG (2022c): Masterplan Neuenheimer Feld. Zahlen zu den Einrichtungen. www.masterplan-neuenheimer-feld.de/informationen/zahlen-zu-den-einrichtungen (15.09.2022).

STATEVA, G.; BOSCH, O.; WINDMEIJER, D.; MASLANKOWSKI, J.; GIULIO, B. & SCANNAPIECO, M. (2018): Final report. Web scraping Enterprise Characteristics. ec.europa.eu/eurostat/cros/sites/default/files/Wp2_Del2_4.pdf (24.08.2022).

STATISTIK DER BUNDESAGENTUR FÜR ARBEIT (2021): Tabellen, Gemeindedaten der sozialversicherungspflichtig Beschäftigten nach Wohn- und Arbeitsort, Nürnberg, Stichtag 30.6.2021. sta-

tistik.arbeitsagentur.de/SiteGlobals/Forms/Suche/Einzelheftsuche_Formular.html;jsessionid=52D090477F6D4308DF3C02615A572314?nn=1523064&topic_f=beschaeftigung-sozbegemband (04.05.2022).

STATISTISCHES BUNDESAMT (2021a): Rechtliche Einheiten nach zusammengefassten Rechtsformen. www.destatis.de/DE/Themen/Branchen-Unternehmen/Unternehmen/Unternehmensregister/Tabellen/unternehmen-rechtsformen-wz08.html (04.05.2022).

STATISTISCHES BUNDESAMT (2021b): Unternehmen mit einer Website nach Wirtschaftszweigen und Beschäftigtengrößenklassen. www.destatis.de/DE/Themen/Branchen-Unternehmen/Unternehmen/IKT-in-Unternehmen-IKT-Branche/Tabellen/ikti-04-anteil-unternehmen-internetzugang-website.html (27.09.2022).

STEDMAN, R. C.; CONNELLY, N. A.; HEBERLEIN, T. A.; DECKER, D. J. & ALLRED, S. B. (2019): The End of the (Research) World As We Know It? Understanding and Coping With Declining Response Rates to Mail Surveys. In: Society & Natural Resources 32(10): 1139–1154.

STEINKE, I. (2007): Qualitätssicherung in der qualitativen Forschung. In: Kuckartz, U., Dresing, T., & Grunenberg, H (Hrsg.): Qualitative Datenanalyse: computergestützt. Wiesbaden: VS Verlag für Sozialwissenschaften: 176–187.

STICH, C.; TRANOS, E. & NATHAN, M. (2022): Modeling clusters from the ground up: A web data approach. In: Environment and Planning B: Urban Analytics and City Science 50(1): 239980832211081.

STONE, P.; BROOKS, R.; BRYNJOLFSSON, E.; CALO, R.; ETZIONI, O.; HALGER, G. & HIRSCHBERG, J. (2016): Artificial Intelligence and life in 2030: the one hundred year study on artificial intelligence. In: arXiv preprint arXiv:2211.06318: 1–52.

STULPE, A. & LEMKE, M. (2016): Blended Reading. In: LEMKE, M. & WIEDEMANN, G. (Hrsg.): Text Mining in den Sozialwissenschaften. Wiesbaden: Springer VS: 17–61.

SUADAA, L. H. (2014): A survey on web usage mining techniques and applications. In: 2014 International Conference on Information Technology Systems and Innovation (ICITSI). IEEE.

SUOMINEN, A. & TOIVANEN, H. (2016): Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. In: Journal of the Association for Information Science and Technology 67(10): 2464–2476.

SUTSKEVER, I.; VINYALS, O. & LE, Q. V. (2014): Sequence to Sequence Learning with Neural Networks. In: Advances in Neural Information Processing Systems 27: 1–9.

TAN, H.; LI, J.; HE, M.; LI, J.; ZHI, D.; QIN, F. & ZHANG, C. (2021): Global evolution of research on green energy and environmental technologies:A bibliometric study. In: Journal of environmental management 297: 113382.

TANG, R.; PANDEY, A.; JIANG, Z.; YANG, G.; KUMAR, K.; LIN, J. & TURE, F. (2022): What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In: arXiv preprint arXiv:2210.04885: 1–5.

TENNEY, I.; WEXLER, J.; BASTINGS, J.; BOLUKBASI, T.; COENEN, A.; GEHRMANN, S.; JIANG, E.; PUSHKARNA, M.; RADEBAUGH, C.; REIF, E. & YUAN, A. (2020): The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In: arXiv preprint arXiv:2008.05122: 1–12.

THELWALL, M.; VAUGHAN, L. & BJÖRNEBORN, L. (2005): Webometrics. In: Annual Review of Information Science and Technology 39(1): 81–135.

THIEM, A. & DUSA, A. (2013): Qualitative Comparative Analysis with R. A User's Guide. New York: Springer.

TRAJTENBERG, M. (2019): Artificial intelligence as the next GPT: A political-economy perspective. In: GOLDFARB, A.; GANS, J. & AGRAWAL, A. (Hrsg.): The economics of artificial intelligence: An agenda. Chicago: University of Chicago Press: 175–186.

TRANOS, E.; CARRASCAL-INCERA, A. & WILLIS, G. (2022): Using the Web to Predict Regional Trade Flows: Data Extraction, Modeling, and Validation. In: Annals of the American Association of Geographers 112(4): 1–23.

TSENG, Y.-H.; LIN, C.-J. & LIN, Y.-I. (2007): Text mining techniques for patent analysis. In: Information Processing & Management 43(5): 1216–1247.

TURIAN, J.; LEV, R. & YOSHUA, B. (2010): Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics: 1–12.

UYSAL, A. K. & GUNAL, S. (2014): The impact of preprocessing on text classification. In: Information Processing & Management 50(1): 104–112.

UZUN, E. (2020): A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages. In: IEEE Access 8: 61726–61740.

VAJRE, V.; NAYLOR, M.; KAMATH, U. & SHEHU, A. (2021): PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE.

VASHISHTH, S.; UPADHYAY, S.; TOMAR, G. S. & FARUQUI, M. (2019): Attention Interpretability Across NLP Tasks. In: arXiv preprint arXiv:1909.11218: 1–10.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L., GOMEZ, A. N. & POLOSUKHIN, I. (2017): Attention is all you need. In: Advances in Neural Information Processing Systems(30): 1–9.

VAUGHAN, L.; GAO, Y. & KIPP, M. (2006): Why are hyperlinks to business Websites created? A content analysis. In: Scientometrics 67(2): 291–300.

VAUGHAN, L. & WU, G. (2004): Links to commercial websites as a source of business information. In: Scientometrics 60(3): 487–496.

VENUGOPALAN, S. & RAI, V. (2015): Topic based classification and pattern identification in patents. In: Technological Forecasting and Social Change 94: 236–250.

VEREINTE NATIONEN (2015a): Adoption of the Paris agreement. unfccc.int/sites/default/files/english_paris_agreement.pdf (13.12.2022).

VEREINTE NATIONEN (2015b): Transforming our world: the 2030 Agenda for Sustainable Development. sdgs.un.org/2030agenda (13.12.2022).

VERISIGN, INC. (2022): The domain industry brief. www.verisign.com/assets/domain-name-report-Q42021.pdf#page=2 (19.10.2022).

VERNON, R. (1992): International investment and international trade in the product life cycle. In: LETICHE, J. (Hrsg.): International Economic Policies and their Theoretical Foundations. Berkeley: Academic Press: 415–435.

VIG, J. (2019): A Multiscale Visualization of Attention in the Transformer Model. In: arXiv preprint arXiv:1906.05714: 1–6.

VIJAYARANI, S.; ILAMATHI, M. J. & NITHYA, M. (2015): Preprocessing Techniques for Text Mining - An Overview. In: International Journal of Computer Science & Communication Networks 5(1): 7–16.

VIJAYARANI MOHAN (2015): Preprocessing Techniques for Text Mining - An Overview. In: International Journal of Computer Science & Communication Networks(5): 7–16.

VOGEL, P. & HILGENDORF, E. (2019): Web Scraping in der unabhängigen wissenschaftlichen Forschung. Gutachten im Auftrag des Wissenschaftszentrums Berlin für Sozialforschung gGmbH (WZB) – Rat für Sozial- und Wirtschaftsdaten (RatSWD). In: Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement - Mit Gutachten "Web Scraping in der unabhängigen wissenschaftlichen Forschung". RatSWD Output. German Data Forum (RatSWD): 31–56.

VOGELS, T.; GANEA, O.-E. & EICKHOFF, C. (2018): Web2Text: Deep Structured Boilerplate Removal. In: Springer, Cham: 167–179.

WALLACE, E.; TUYLS, J.; WANG, J.; SUBRAMANIAN, S.; GARDNER, M. & SINGH, S. (2019): AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. In: arXiv preprint arXiv:1909.09251: 1–6.

WANG, A.; PRUKSACHATKUN, Y.; NANGIA, N.; SINGH, A.; MICHAEL, J.; HILL, F.; LEVY, O. & BOWMAN, S. R. (2019a): SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In: Advances in Neural Information Processing Systems 32: 1–15.

WANG, A.; SINGH, A.; MICHAEL, J.; HILL, F.; LEVY, O. & BOWMAN, S. R. (2019b): GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: arXiv preprint arXiv:1804.07461: 1–20.

WEBSTER, J. J. & KIT, C. (1992): Tokenization as the initial phase in NLP. In: Proceedings of the 14th conference on Computational linguistics -. Morristown, NJ, USA: Association for Computational Linguistics.

WENZEK, G.; LACHAUX, M.-A.; CONNEAU, A.; CHAUDHARY, V.; GUZMÁN, F.; JOULIN, A. & GRAVE, E. (2019): CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In: arXiv preprint arXiv:1911.00359: 1–9.

WIEDEMANN, G. (2013): Opening up to big data: Computer-assisted analysis of textual data in social sciences. In: Historical Social Research/Historische Sozialforschung 14(2): 332–357.

WIEDEMANN, G. (2016): Text mining for qualitative data analysis in the social sciences. A study on democratic discourse in Germany. Kritische Studien zur Demokratie. Wiesbaden: Springer VS.

WIEGREFFE, S. & PINTER, Y. (2019): Attention is not Explanation. In: arXiv preprint arXiv:1908.04626: 1–12.

WINOGRAD, T. (1972): Understanding natural language. In: Cognitive Psychology 3(1): 1–191.

WOLF, J. (2021): HTML und CSS. Das umfassende Handbuch zum Lernen und Nachschlagen. Inkl. JavaScript, Bootstrap, Responsive Webdesign u. v. m. Rheinwerk Computing. 4. Aufl. Bonn: Rheinwerk Computing.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. von; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; LE SCAO, T.; GUGGER, S.; DRAME, M.; LHOEST, Q. & RUSH, A. M. (2020): HuggingFace's Transformers: State-of-the-art Natural Language Processing. In: arXiv preprint: 1910.03771: 1–8.

WOODS, W. (1977): Lunar rocks in natural English: Explorations in natural language question answering. In: ZAMPOLLI, A. (Hrsg.): Linguistic Structures Processing. Dordrecht: North-Holland Publishing Company: 521–569.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K.; KLINGNER, J.; SHAH, A.; JOHNSON, M.; LIU, X.; KAISER, Ł.; GOUWS, S.; KATO, Y.; KUDO, T.; KAZAWA, H.; STEVENS, K.; KURIAN, G.; PATIL, N.; WANG, W.; YOUNG, C.; SMITH, J.; RIESA, J.; RUDNICK, A.; VINYALS, O.; CORRADO, G.; HUGHES, M. & DEAN, J. (2016): Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In: arXiv preprint arXiv:1609.08144: 1–23.

WYLY, E. (2014): The new quantitative revolution. In: Dialogues in Human Geography 4(1): 26–38.

XIONG, H.; CHENG, Y.; ZHAO, W. & LIU, J. (2019): Analyzing scientific research topics in manufacturing field using a topic model. In: Computers & Industrial Engineering 135: 333–347.

XU, G.; ZHANG, Y. & LI, L. (2011): Web Mining and Social Networking. Techniques and Applications. SpringerLink Bücher 6. 1. Aufl. Boston, MA: Springer US.

YANG, T. I.; TORGET, A. & MIHALCEA, R. (2011): Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities: 96–104.

YANG, Z.; DAI, Z.; YANG, Y.; CARBONELL, J.; SALAKHUTDINOV, R. R.; LE & Q. V. (2019): Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems 32: 1–10.

YAU, C.-K.; PORTER, A.; NEWMAN, N. & SUOMINEN, A. (2014): Clustering scientific documents with topic modeling. In: Scientometrics 100(3): 767–786.

YOON, J.; PARK, H. & KIM, K. (2013): Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. In: Scientometrics 94(1): 313–331.

YOUTIE, J.; HICKS, D.; SHAPIRA, P. & HORSLEY, T. (2012): Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. In: Technology Analysis & Strategic Management 24(10): 981–995.

YUN, J. & GEUM, Y. (2020): Automated classification of patents: A topic modeling approach. In: Computers & Industrial Engineering 147: 106636.

ZAMANI, M. & SCHWARTZ, H. A. (2017): Using Twitter Language to Predict the Real Estate Market. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Stroudsburg, PA, USA: Association for Computational Linguistics.

ZENTRUM FÜR EUROPÄISCHE WIRTSCHAFTSFORSCHUNG (2022): Mannheimer Unternehmenspanel. www.zew.de/forschung/mannheimer-unternehmenspanel (20.04.2022).

ZHANG, Y.; ZHANG, G.; ZHU, D. & LU, J. (2017): Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. In: Journal of the Association for Information Science and Technology 68(8): 1925–1939.

ZHENG, L.; GUHA, N.; ANDERSON, B. R.; HENDERSON, O. & HO, D. E. (2021): When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In: MARANHÃO, J. (Hrsg.): Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. ACM Digital Library. New York: Association for Computing Machinery: 159–168.

ZHU, Y.; LYNETTE WANG, V.; WANG, Y. J. & NASTOS, J. (2020): Business-to-business referral as digital coopetition strategy. In: European Journal of Marketing 54(6): 1181–1203.

ZIELSTRA, D. & ZIPF, A. (2010): A comparative study of proprietary geodata and volunteered geographic information for Germany. In: 13th AGILE international conference in geographic information science: 1–15.

ZOPH, B.; BELLO, I.; KUMAR, S.; DU, N.; HUANG, Y.; DEAN, J.; SHAZEER, N. & FEDUS, W. (2022): ST-MoE: Designing Stable and Transferable Sparse Expert Models. In: arXiv preprint arXiv:2202.08906: 1–38.

ZSCHECH, P.; FLEIßNER, V.; BAUMGÄRTEL, N. & HILBERT, A. (2018): Data Science Skills and Enabling Enterprise Systems. In: HMD Praxis der Wirtschaftsinformatik 55(1): 163–181.

Eidesstattliche Erklärung

Ich erkläre: Ich habe die vorgelegte Dissertation selbstständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Ich stimme einer evtl. Überprüfung meiner Dissertation durch eine Antiplagiat-Software zu. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.

_____ _____

Ort, Datum Unterschrift