Entwicklung eines Public Opinion Mining Systems zur Analyse der öffentlichen Meinung zum Thema Digitalisierung

Bachelorarbeit

Hochschule der Medien Stuttgart
Fachbereich Information und Kommunikation
Studiengang Wirtschaftsinformatik und digitale Medien
7. Semester

Erstbetreuer: Prof. Dr. David Klotz
Zweitbetreuer: Prof. Dr. Martin Engstler

Sommersemester 2020

Vorgelegt von: Felix Bieswanger
Matrikelnummer: 34379

Stuttgart, 03.08.2020

Ehrenwörtliche Erklärung

Name:	Bieswanger	Vorname:	Felix
Matrikel-Nr.:	34379	Studiengang:	Wirtschaftsinformatik

Hiermit versichere ich, Felix Bieswanger, ehrenwörtlich, dass ich die vorliegende Bachelorarbeit mit dem Titel: *„Entwicklung eines Public Opinion Mining Systems zur Analyse der öffentlichen Meinung zum Thema Digitalisierung“* selbständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§ 24 Abs. 2 Bachelor-SPO (7-Semester) der Hochschule der Medien) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Stuttgart, 03. August 2020

.....	.....
Ort, Datum	Unterschrift

Kurzfassung

Die Auswirkungen der Digitalisierung beeinflussen die Entwicklung der gesamten globalen Gesellschaft und werden daher stark in den Medien diskutiert. Von hohem Interesse ist dabei die Meinung der Bevölkerung zu Themen der Digitalisierung, da diese die Adaption, der aus der Digitalisierung entstehenden Innovationen, maßgeblich beeinflusst. Die Bevölkerung des deutschsprachgien Raums bekommt häufig die Eigenschaft zugewiesen, technikskeptisch zu sein und im Vergleich zum anglo-amerikanischen Sprachraum eher die Risiken zu sehen. In dieser Arbeit wird aufgrund dessen ein Public Opinion Mining System entwickelt, das automatisiert und computergestützt die Stimmungsrichtung eines Sprachraums erfasst und in einer Kennzahl ausdrückt, mit der die Wahrnehmung der Digitalisierung verglichen werden kann. Es kann hoch signifikant gezeigt werden, dass im deutschen Sprachraum negativer über die Digitalisierung gesprochen wird als im anglo-amerikanischen Sprachraum. Dabei ist die Tonalität des deutschen Sprachraums zwar leicht positiv, jedoch nur halb so positiv als die des anglo-amerikanischen Sprachraums.

Schlagworte: **Digitalisierung, Opinion Mining, Topic Modelling, Webscraping, öffentliche Meinung**

Abstract

The effects of digitization influence the development of the entire global society and are therefore strongly discussed in the media. The opinion of the population towards digitization is of great interest, as this has a decisive influence on the adaptation of the innovations resulting from digitization. People in the German-speaking world are often said to be skeptical of technology and, compared to the Anglo-American world, are more likely to see the risks. Therefore, in this thesis a public opinion mining system is developed, which automatically and computer-aided captures the sentiment of a linguistic area and expresses it in a key figure, with which the perception of digitization in different language areas are compared. This work demonstrates in a highly significant manner, that the German language area talks more negatively about digitization than in the Anglo-American language area. While the tonality of the German language area is slightly positive, it is nevertheless just half as positive as the Anglo-American language area.

Keywords**: Digitalization, Opinion Mining, Public Opinion, Topic Modelling, Webscraping**

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

LDA	Latent Dirichlet Allocation
MAE	Median Absolute Error
ML	Machine Learning
USA	United Staates of America

1 Einleitung

Dieses erste Kapitel führt in das Thema dieser Arbeit ein und definiert die Forschungsfrage. Zudem wird der Aufbau der Arbeit erklärt, indem ein Überblick über die folgenden Kapitel erstellt wird.

1.1 Themenrelevanz

Die signifikante Zunahme an Rechenleistung und die Menge an gesammelten Daten hat in der Vergangenheit eine Vielzahl an digitalen Innovationen geführt und ist als Digitalisierung (auch digitale Wende) bekannt (vgl. Bendel, 2018). Durch den rasanten Anstieg an technologischen Möglichkeiten kommt es zu immer kürzeren Entwicklungszeiten (vgl. Hamidian & Kraijo, 2013, S.13). Diese hohe Dynamik hat Auswirkungen auf Wirtschaft, Politik, Gesellschaft und sorgt für einen tiefgreifenden Wandel in der Industrie und Gesellschaft (vgl. Bundesministerium für Wirtschaft und Energie, 2020; Gillen & Wambach, 2018, S.160f).

Ein Thema mit derartiger Tragweite für die gesamte Gesellschaft ergibt einen hohen Kommunikations- und Disskusionsbedarf, wodurch die Digitalisierung längst ein fester Bestandteil des öffentlichen Diskurses ist und aktuell sehr stark von den allgemeinen und fachbezogenen Medien diskutiert wird (vgl. Gadatsch, 2017, S.193; vgl. Kröhling, 2017, S.23; vgl. Lenz, 2019, S.149). Auf der einen Seite wird mit der Darstellung von Chancen ein positives Stimmungsbild gezeichnet. Dabei wird unter anderem auf die verbesserte Lebensqualität eingegangen, die durch Innovationen der Digitalisierung ermöglicht wird. Auf der anderen Seite werden die Risiken beleuchtet. Dazu zählt beispielsweise die Angst vieler Menschen aus ihren angestammten Berufen verdrängt zu werden, da im Vergleich zu klassischen Arbeitgebern jene mit digitalem Geschäftsmodell meist erheblich weniger Ressourcen für die Erzielung eines ähnlichen Marktwert benötigen (vgl. Gillen & Wambach, 2018, S.161ff).
Dabei ist vielerorts die Behauptung verbreitet, dass die deutsche Öffentlichkeit latent technikskeptisch sei und sich eher auf die Risiken der Digitalisierung fixiere (vgl. Zeller et al., 2010, S.504). Demgegenüber besteht der subjektive Eindruck, dass die mediale Darstellung englischsprachiger Quellen ein deutlich optimistischeres Bild

zeichne. In Zusammenhang besteht ein Erkenntnisinteresse woraus die Forschungsfrage für diese Arbeit abgeleitet werden kann:

Wie wird das Thema Digitalisierung im deutschen Sprachraum im Vergleich zum anglo-amerikanischen Sprachraum behandelt?

Die von der Digitalisierung geschaffenen Möglichkeiten und der damit verbundene technische Wandel, erzeugen eine Vielzahl von Innovationen. Als Voraussetzung für die Durchsetzung von Innovationen gilt meist nicht allein der technische Fortschritt, sondern zudem vielmehr die gesellschaftliche Akzeptanz (vgl. Zeller et al, 2010, S.503; vgl. Granig, 2007, S.10). Erkenntnisse über die Meinung der Bevölkerung zum Thema Digitalisierung sind daher vor allem für politische Entscheidungsträger hochgradig relevant, da die gewonnen Erkenntnisse als Frühwarnsystem eingesetzt werden können, um mögliche Fehlentwicklungen des technologischen Wandels zu vermeiden und notwendige Gegenmaßnahmen abzuleiten (vgl. Zerfaß und Volk, 2019, S.195f.).
Auch im Kontext vieler internationaler Unternehmen mit digitalem Geschäftsmodell ist die Beantwortung der Fragestellung von Interesse, da diese durch die Ergebnisse dieser Arbeit ein besseres Verständnis für die Bedürfnisse ihrer Kunden erlangen. Beispielsweise wäre dieses Wissen bei der Entscheidung zur Wahl eines Absatzmarktes neuer digitaler Produkte im strategischen Management relevant, in dem der Sprachraum (und speziell die Länder welche darin enthalten sind) mit der positiveren Einstellung zur Digitalisierung bevorzugt wird oder mehr Ressourcen als Überzeugungsleistung im kritischeren Sprachraum eingeplant werden.

Verstärkt wird der Bedarf dieser Arbeit, da Studien der öffentlichen Meinung hinsichtlich des technischen Wandels der Bundesrepublik fast vollständig fehlen (vgl. Störk-Biber et al., 2020, S. 22f). Anhand der Forschungsfrage trägt diese Arbeit damit einen Beitrag zu dem Teil der Sozialforschung bei, die sich mit der Wahrnehmung der Digitalisierung befasst.

1.2 Aufbau der Arbeit

Nachfolgend ist in Abbildung 1 der Aufbau dieser Arbeit dargestellt. In Bezug auf die Merkmale der empirischen Forschungsmethoden ist im Hinblick auf die Forschungsfrage eindeutig, dass sich ein quantitativer Forschungsansatz empfielt (vgl. Tausendpfund, 2018, S.18). Dabei orientiert sich der Aufbau der Arbeit an das allgemeine Modell von Raithel für den Forschungsablauf für quantitative Studien (vgl. Raithel, 2006, S.24ff).

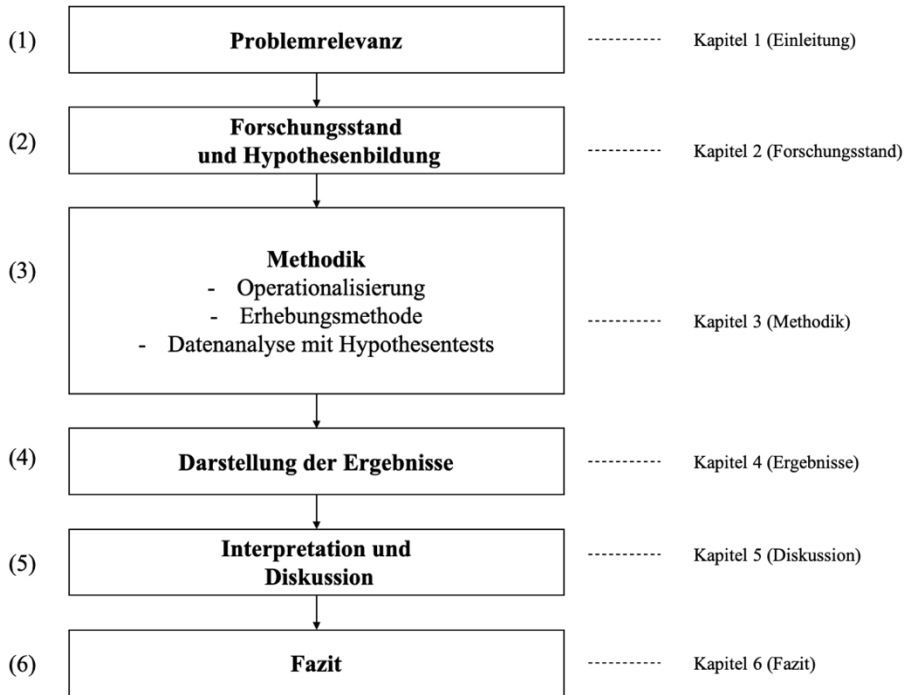


Abbildung 1: Aufbau der Arbeit

Quelle: Raithel, 2006, S.24ff (geändert)

Im Kapitel 1.1 (Einführung) der Einleitung wurde bereits die Relevanz des Themas beschrieben und daraufhin die Forschungsfrage aufgestellt. Nachfolgend wird im Kapitel 2 (Forschungsstand) der Forschungsstand beschrieben und darauf aufbauend die Hypothesenbildung durchgeführt. Im Kapitel 3 (Methodik) wird zuerst die Hypothese operationalisiert und anschließend wird das Erhebungsinstrument beschrieben, mit dem die empirische Messung der Variablen der Hypothesen gelingen soll. Im letzten Schritt werden die erhobenen Daten mit Hilfe von einem statistischen Hypothesentests analysiert. Im Anschluss folgt in Kapitel 4 (Ergebnisse) die Darstellung der Ergebnisse und woraufhin in Kapitel 5 (Diskussion) die Diskussion und Interpretation derer folgt. Schließlich wird im Kapitel (6) ein Fazit gezogen.

2 Stand der Forschung und Theoretische Grundlagen

In diesem Kapitel wird zuerst wird der Stand der Forschung der vorgestellten Forschungsfrage aufgearbeitet. Daraus werden anschließend Forschungslücken identifiziert und das konkrete Forschungsinteresse in dieser Arbeit spezifiziert. Anschließend werden noch theoretische Grundlagen zum Verständnis des Public Opinion Mining gelegt.

2.1 Forschungsstand

In der Literatur gibt derzeit nur wenige aktuelle Arbeiten, die sich mit der öffentlichen Meinung zum Thema Digitalisierung beschäftigen (vgl. Störk-Biber et al., 2020, S.22). Die Aktualität der Studien ist dabei von besonderer Relevanz, da sich die Meinung der Öffentlichkeit schnell ändern kann. Nichtsdestotrotz liegen einige wenige Arbeiten vor und werden nachfolgend betrachtet. Die gefundenen Studien sind nicht jedoch nicht direkt mit der Forschungsfrage vergleichbar, da konkrete Länder und nicht Sprachräume untersucht wurden. Allerdings können die Befunde sehr gut als Annäherung verwendet werden und Aufschluss über bisherige Lösungsansätze geben.

Störk-Biber et al. präsentieren in Ihrer Arbeit die Ergebnisse des TechnikRadars der Deutschen Akademie der Technikwissenschaften und Körber-Stiftung. Mittels einer jährlichen Befragung der deutschen Bevölkerung, erstellt das TechnikRadar eine Analyse zur Haltung gegenüber innovativen Technologien. Die Studie ergab, dass die Meinung in Deutschland zum Thema sehr zwiegespalten sei und die Technikbegeistung vieler befragten stehe der Skepsis um den Fortschritt gegenüber (vgl. Störk-Biber et al., 2020, S.25). Jedoch betonen die Autoren, dass es keine Hinweise auf eine allgemeine Ablehnung von Technik in Deutschland gebe (vgl. Störk-Biber et al., 2020, S.24).
Des Weiteren stellt das Vodafone Institut für Gesellschaft und Kommunikation die Wahrnehmung der Digitalisierung mit den Studien zur Reihe “Die unterschiedliche Wahrnehmung der Digitalisierung in Europa, Asien und den USA” in den internationalen Vergleich (vgl. Vodafone Institut für Gesellschaft und Kommunikation, 2018; vgl. Vodafone Institute for Society and Communications, 2019). In dieser Studienreihe befragten die Autoren Paus et. al 9.000 Menschen aus

neun Ländern mittels einer standardisierten Online-Umfrage im vergangenen Jahr 2019 (teilweise spät 2018). Die von Störk-Biber et al. berichtete Skepsis wird hier erneut widergespiegelt. Nur wenige deutsche Bürger glauben ihre Regierung sei der Herausforderung gewachsen. Amerikanische Bürger auf der anderen Seite sein hinsichtlich dessen positiver eingestellt. Darüber hinaus zeigten sie, dass Deutschland im Vergleich zu allen untersuchten Ländern am pessimistischsten gegenüber der Digitalisierung eingestellt sei und im Vergleich zu den USA oder Großbritannien Sprachraums einen geringeren wahrgenommen Digitalisierungsgrad aufweisen (vgl. Vodafone Institute for Society and Communications, 2018, S.3). Mit den vorgestellten Arbeiten liegt ein Hinweis vor, dass Länder des deutschen Sprachraums kritischer zur Digitalisierung sind als Länder des anglo-amerikanischen Sprachraums.

Die identifizierten Studien in der Literatur bedienten sich derselben Forschungsmethodik. Die Durchführung einer demoskopischen Umfrage ist in der Literatur die gängige Methode in der Meinungsforschung zur Erhebung der öffentlichen Meinung (vgl. Schweiger, 2019, S.116; vgl. Petersen, 2008, S.370). Die Umsetzung und Auswertung dieser ist nur jedoch mit hohem Zeitaufwand möglich. Daher gibt es alternative Ansätze, um die Meinung der Öffentlichkeit abzubilden. Die individuelle Meinungsbildung sei nach Schweiger vor allem durch die Wahrnehmung externer Informationsquellen geprägt (vgl. Schweiger, 2017, S.116). Die Presse gilt dabei als wichtigste externe Informationsquelle, die als Informationsfunktion ein Teil der öffentlichen Meinungs- und Willensbildung in einer Demokratie gilt (vgl. Raupp & Vogelgesang, 2009, S. 17; vgl. Kelly, 2009, S.34; vgl. Schweiger, 2017, S. 131; vgl. Früh, 2017, S.201). Daher sind Meinungen oft weniger das Abbild der persönlichen Wirklichkeitserfahrung, sondern vielmehr das durch die Medien erzeugte Bild der Realität (vgl. Hanni & Hermann, 1996, S.1). Mc Geogor führt in seiner Arbeit ebenfalls auf, dass Nachrichten die Meinung der Bevölkerung über die Berichtserstattung wiederspiegeln würden (vgl. McGregor, 2019, S.2). In diesem Zusammenhang ist die Presse sowohl als ein Abbild der öffentlichen Meinung, als auch als meinungsbildende Instanz zu sehen (vgl. Zeller et al, 2010, S.503). Anhand dieses alternativen Ansatz untersuchte Zeller et al. bereits, die ob es Tendenzen in der der Berichterstattung zur Digitalisierung in den deutschen Printmedien gäbe (vgl. Zeller et al., 2010). Die konkreten Ergebnisse der Arbeit sind jedoch veraltet und

aufgrund dessen nicht während der Einordnung der bisherigen Arbeiten zur Forschungssfrage genannt. Da in der Arbeit von Zeller at al. Daten manuell ausgewertet wurden besteht vor allem in großen Textmengen die Schwierigkeit, die Aktualität der Daten bei gleichzeitig aufwändiger Konzeptionsphase und Durchführung zu vereinen. Automatisierte, computeunterstützende Verfahren kommen dieser Problematik zugute, da sie in der Lage sind die im Internet entstehende Informationsflut zu verarbeiten, bei der eine manuelle Auswertung unmöglich scheint (vgl. Früh, 2017, S.199; vgl. Neuberger, 2004, S.6). Verstärkt wird diese Relevanz automatisierter und computeunterstützender Verfahren, da durch die Nutzung des Internets die Veröffentlichung seiner Meinung gegenüber den traditionellen Medien, die selbst von der Digitalisierung betroffen sind, deutlich vereinfacht wurde (vgl. Neuberger, 2004, S.7). Seit circa 10 bis 15 Jahren verlagert sich daher die Meinungsäußerungen immer stärker in Richtung der Online-Angebot von Zeitungen (vgl. Petring, 2016, S.372). Während sich die Forschung auf die Analyse von Rezensionen konzentriert, halten Scholz et al. sogar die Arbeit mit Nachrichtendaten als interessanter, da im Gegensatz zu Rezensionen keine offensichtliche Gesamtwertung bereits vorliege (vgl. Scholz et al., 2012, S.259). Hiernach sei nach Schweiger die Nutzung der online Medien als eine weitere Variante der Öffentlichkeitsmessung valide (vgl. Schweiger, 2017, S.117).

2.2 Forschungsinteresse

In Folge des vorgestellten Forschungsstands kann eine Forschungslücke in Bezug auf die Methodik identifiziert werden. Zwar wurden bereits Studien zur unterschiedlichen Wahrnehmung der Digitalisierung durchgeführt, jedoch nur mit Hilfe von manuellen Methoden. Aus dem Forschungsstand geht hervor, dass die Nutzung von automatischen Mitteln für die Erzeugung von Stimmungsbildern der öffentlichen Meinung ein geeigneter Ansatz ist. Dabei kritisieren Lemke & Wiedemann, dass automatische Verfahren in der Sozialforschung bisher nicht sehr anerkannt seien, obwohl durch die Digitalisierung immer mehr Texte maschinenlesbar vorliegen. Gleichzeitig stelle es jedoch SozialwissenschaftlerInnen vor die große Herausforderung Analysen durch eigenständige Programmierung umzusetzen (vgl. Lemke & Wiedemann, 2016, S.1; vgl. Lemke & Wiedemann, 2016, S.4).

An dieser Problematik knüpft diese Arbeit an. Die Neuheit der Arbeit besteht also darin, eine diziplinfremde Forschungsfrage der Sozialwissenschaft mit Hilfe der Mittel der Wirtschaftsinformatik zu bearbeiten.

Aus diesem Zusammenhang resultierend besteht das Ziel dieser Arbeit, unter Anwendung eines automatisierten, computerunterstützenden Systems nachzuweisen, dass im deutschen Sprachraum negativer über die Digitalisierung gesprochen wird als im anglo-amerikanischen Sprachraum. Damit ergibt sich die Hypothese H1:

“Im deutschen Sprachraum wird mit einer negativeren Tonalität über die Digitalisierung gesprochen als im anglo-amerikanischen Sprachraum."

Weiterhin steht hiernach auch die Nullhypothese H0 fest:

“Im deutschen Sprachraum wird mit der gleichen Tonalität über die Digitalisierung gesprochen wie im anglo-amerikanischen Sprachraum”

Dabei liegt der Fokus dieser Arbeit nicht auf der Validierung der Annahme, dass die Online Medien zur Abbildung der öffentlichen Meinung geeignet sei. Eine solche Arbeit wäre eher der Sozialwissenschaft zuzuordnen. Vielmehr liegt der Kern dieser Arbeit die Stimmungsrichtung von Quellen zum Thema Digitalisierung festzustellen. Damit ist diese Arbeit nach Petz der wissenschaftlichen Disziplin der Wirtschaftsinformatik zuzuordnen (vgl. Petz, 2019, S.7).

2.3 Grundlagen

In diesem Kapitel werden die Grundlagen und Stand der Forschung hinsichtlich der computergestützten Abbildung der öffentlichen Meinung gelegt.

2.3.1 Begriffe und Definition

Das computergestützte Analysieren von Meinungen aus Texten ist in der Literatur unteranderem von Petz, Liu, Pang & Lee, Scholz und Kim et al. durch den Begriff *Opinion Mining* geprägt (vgl. Y. Kim et al., 2014; vgl. Liu, 2012; vgl. Pang & Lee, 2008; vgl. Petz, 2019; vgl. Scholz, 2011).
Allerdings werden die Begriffe *Sentiment Analyse* und *Opinion Mining* als Synonym verwendet, wie beispielsweise in der Arbeit on Liu (vgl. Liu, 2012, S.7). Übereinstimmend zu Petz wird die Gleichstellung der Begriffe in dieser Arbeit abgelehnt (vgl. Petz, 2019, S.22). Die Sentiment Analyse wird als der Teil der Quantifizierung der Stimmungsrichtung verstanden.
Zudem ist auch der Begriff *Sentiment Klassifikation* in der Literatur verbreitet. Dieser wird von Turney und Yang & Huang verwendet, da in ihren Arbeiten die zu untersuchenden Daten in die Klassen Positiv und Negativ (manchmal auch Neutral) eingeordnet werden soll (vgl. Turney, 2002; vgl. Yang & Huang, 2019). Dabei wird jedoch die Stärke der Stimmungsrichtung vernachlässigt.
Der Begriff *Opinion Mining* wird ferner von Li & Gao,Shang et al. und Kim & Kim mit dem Zusatz Public erweitert, da analog zum Forschungsziel in dieser Arbeit auch in Ihren Arbeiten die Meinung der Öffentlichkeit durch das Opinion Mining repräsentiert werden soll (Kim und Kim 2014) (Shang u. a. 2015) (Li und Gao 2013).

Somit wird in dieser Arbeit der Begriff *Public Opinion Mining* verwendet und definiert die computerunterstützte Analyse und systematische Aufbereitung der Meinung der Öffentlichkeit.

2.3.2 Einordnung in andere Forschungsbereiche

Dabei bedient sich das Public Opinion Mining Methoden anderer Forschungsbereiche und wird nachfolgend eingeordnet. In Abbildung 2 sind die Beziehungen zwischen den relevanten verwandten Forschungsbereiche dargestellt.

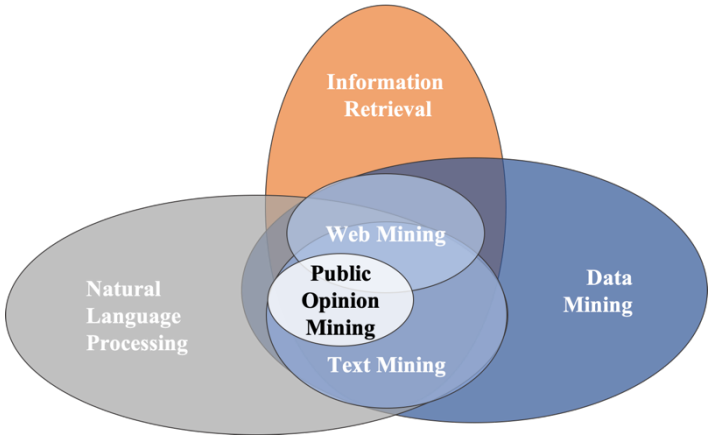


Abbildung 2: Venn Diagramm mit den Beziehungen der Forschungsbereiches

Als allumfassendes Forschungsgebiet des Public Opinion Mining wird zuerst das Forschungsgebiet *Data Mining* genannt, dass sich im Allgemeinen mit dem Generieren von Wissen aus Daten beschäftigt (vgl. Runkler, 2015, S.2). Dabei können die Datenmengen zwar aus unterschiedlichen Datentypen bestehen, jedoch wird mit dem Begriff Data Mining eher die Analyse von strukturierten Daten assoziiert (vgl. Hippner & Rentzmann, 2006, S.287).

Daher kommen spezifischer für das Public Opinion Mining jene Data Mining Methoden zum Einsatz, die Erkenntnisse aus unstrukturierten Textdaten hervorbringen, dem sogenannten *Text Mining*. Hipper und Rentzmann erklären, dass das Verstehen der sprachlich wiedergegebenen Information in unstrukturierten Text-Daten die Kernfunktion des Text Mining sei (vgl. Hippner & Rentzmann, 2006, S.287; vgl. Liu & Zhang, 2012, S.2). Dabei ist das wesentliche Vorgehensmodell und verwendete Algorithmen ähnlich zum Data Mining (vgl. Petz, 2019, S.27). Eine typische Aufgabenstellung für das Public Opinion Mining ist dabei durch die Analyse der Ähnlichkeit Beziehungen zwischen den Dokumenten herzustellen. Dabei werden Klassifizierungsalgorithmen für gelabelte Daten (Supervised Learning) und Clustering Algorithmen für nicht gelabelte Daten (Unsupervised Learning) verwendet (vgl. Miner et al., 2012, S.36).

Damit Machine Learning (ML) Algorithmen auf Text angewendet werden können, müssen die Textdaten zuvor transformiert werden. Die meist verbreitete Darstellung von textuellen Daten wird bag-of-words genannt. Dabei werden zuerst aus allen unterschiedlichen Wörtern aller Dokumente ein Vokabular aufgebaut. Anschließend wird ein mulitdimensionalier Vektor für jedes Dokument erstellt, für den die Häufigkeit jedes Wort des Vokabulars in dem jeweiligen Dokument als Dimension notiert wird. Anhand dieser Vektoren sind ML Algorithmen erneut einsetzbar, jedoch treten dabei viele Datenlücken auf, da die Anzahl der Wörter des Lexicons sehr viel größer ist als die Anzahl der Worte eins Dokuments. Dadurch bekommen die meisten Dimensionen im Vektor den Wert null zugeordnet (engl. sparsity) (vgl. Aggarwal, 2015, S.430f). Durch diesen Effekt sind ML Modelle mit Textdaten sehr rechenaufwändig.

In Bezug auf die Herkunft der zu analysierenden Daten lässt sich das Forschungsgebiet noch weiter spezifizieren. Während im Data Mining und Text Mining davon ausgegangen wird, dass die zu untersuchenden Daten bereits in einer Datenbank vorliegen, werden im sogenannten *Web Mining* die Daten des Internets analysiert. Für die Erforschung der Web Struktur mittels semi-strukturierten Hyperlinks wird in der Literatur der Begriff *Web Structure Mining* verwendet. Die dabei verwendeten Graphanalyse Algorithmen lassen sich dem Data Mining zuordnen (vgl. Stoffel, 2009, S.9).
Allerdings viel relevanter für das Public Opinion Mining ist das *Web Content Mining*, bei dem die Inhalte der Webseiten erforscht werden (vgl. Walther, 2001, S.16). Für die Datensammlung im Web Mining kommen daher sognannte Webscraper zum Einsatz. Dabei gilt es den relevanten Text von irrelevanten Zusatzinformation wie Werbungen zu unterscheiden (vgl. Aggarwal, 2015, S.433). Vereinfacht geht der Scraper von einer Reihe von Startseiten aus und verwendet dann die darin enthaltenen Links, um andere Seiten abzurufen und Informationen zu extrahieren. Die Links in diesen Seiten werden wiederum extrahiert, und die entsprechenden nächsten Seiten besucht. Der Prozess wiederholt sich, bis eine ausreichende Anzahl von Seiten besucht oder ein anderes Ziel erreicht ist (vgl. B. Liu, 2011. S.312).

Eng mit der Datenbeschaffung verbunden ist das Forschungsgebiet *Information Retrieval* (IR), das die Informationssuche in großen Sammlungen von unstrukturierten Daten beschreibt (vgl. Manning et al., 2008, S.1). Die traditionelle IR geht davon aus, dass die grundlegenden Informationseinheit ein Dokument ist, die in einer Sammlung eine Textdatenbank bilden. Angewendet auf das Internet können diese Informationseinheiten auch Web-Seiten sein (vgl. B. Liu, 2011, S.211). Nach Petz zählen zu den Aufgaben des IR typischerweise „die Suche nach Dokumenten auf Basis von Suchbegriffen, die Filterung von gefundenen Dokumenten sowie die Klassifizierung von Dokumenten nach Inhalt“ (Petz, 2019, S.30).

Das Verstehen von Text stellt für Maschinen eine sehr viel komplexere Aufgabe im Vergleich zu Menschen dar. Aus diesem Grund kommen letztlich aus dem Forschungsgebiet *Natural Language Processing* Methoden zum Einsatz, die versuchen die fehlende Datenstruktur in Texten wiederherzustellen (vgl. Allahyari et al., 2017, S.1; vgl. Rajman & Vesely, 2004, S.7). Die maschinelle Verarbeitung natürlicher Sprache ist der Disziplin *Natural Language Processing* (NLP) zuzuordnen, welches als Bindeglied zwischen Informatik und Linguistik steht (vgl. Carstensen et al., 2010, S.1). Der Begriff NLP ist als synonym und englische Übersetzung des Begriffs der *Computer Linguistik* (CL) anzusehen (vgl. Carstensen et al., 2010, S.2; Miner et al., 2012, S.32). Die Herstellung von Strukturen stellt vor allem in der Datenvorbeitungsphase (engl. preprocessing) von Text Mining Algorithmen eine essenzielle Rolle.

Dies kann auch als Prozess gesehen werden, bei dem zuerst mit Hilfe der lexikalischen Analyse (auch Tokenisierung) Texte in Segmente (Wörter oder Sätze) zerteilt werden soll. Nachfolgend wird durch die syntaktische Analyse versucht grammatikalischen Strukturen abzubilden. Diese bilden die Basis für die semantische Analyse, die schließlich auf Erkenntnisse der tatsächlichen Bedeutung des natürliches Textes hofft (vgl. Petz, 2019, S.28).

Aus der CL stammt ferner der Begriff *Korpus*, der für die strukturierte Datenhaltung in empirischen Untersuchungen steht. Ein Korpus ist eine maschinenlesbare Sammlung von gesprochenen oder schriftlichen Äußerungen, die digital erfasst und für eine linguistische Aufgabe aufbereitet wurden und damit die Analyse vereinfachen. Typischerweise bestehen die Korpora aus drei Bestandteilen. Im Kern des Korpus

stehen die Sprachdaten, die in digitaler Form abgespeichert wurden. Diese werden mit Annotationen angereichert, wie beispielsweise aus der Segmentierung in textstrukutelle Einheiten wie Überschrift, Hauptteil und Fußnoten. Ferner wird der Korpus mit Metadaten wie das Enstehungsdatum oder die Textgattung erweitert (vgl. Carstensen et al., 2010, S.482f).

2.3.3 Quantifizierung einer Meinung

Kim und Hovoy stellten in ihrer Arbeit den ersten Ansatz zur computergestützten Abbildung von Meinungen vor. Sie definieren eine Meinung sei durch das Quadrupel (Thema, Meinungsinhaber, Behauptung, Sentiment) abbildbar (Kim & Hovy, 2004). Diese Definition wird später von Liu überarbeitet. Nach Liu wird eine Meinung durch das Quintupel (Entität, Aspekt, Meinungsinhaber, Zeitpunkt, Sentiment) ausgedrückt. Der Zeitpunkt der Meinungsäußerung sei Teil der Definition, sodass die Elemente zusammengehörig zu betrachten sind. Eine Meinung enthalte ferner auch eine Entität, die Produkte, Services, Personen, Ereignisse, Organisationen oder Themen repräsentiert. Eigenschaften einer Entität als seien Aspekt definiert. Beispielsweise kann ein Computer als Entität und die Rechenleistung als Aspekt betrachtet werden. Zusammengefasst werden daher Entität und Aspekt als Meinungsziel bezeichnet. Im zuvor genannten Beispiel wäre das Meinungsziel folglich die Rechenleistung eines Computers. Der Meinungsinhaber sei die Person (oder Organisation), welche die Meinung vertritt. Schließlich halte das Sentiment den zugehörigen Tonalitätswert fest (vgl. Liu, 2012, S.19).

Die Sentiment Analyse, also die tatsächliche Bestimmung der Tonalität eines Textes, kann wie jedes ML Problem in Supervised und Unsupervised Learning (zu Deutsch Überwachtes und Unüberwachtes Lernen) abstrahiert werden.
Die Sentiment Analyse mit Supervised Learning Methoden ist stärker in der Literatur verbreitet (vgl. Petz, 2019, S.58). Zudem zeigen Chaovalit & Zhou in ihrer Arbeit, dass die Sentiment Analyse mit Supervised Learning genauere Ergebnisse ermögliche (vgl. Chaovalit & Zhou, 2005, S.112f). Konzeptionell werden mit Supervised Learning Algorithmen Modelle trainiert, die bei gegebenen Input Daten ein gewünschtes Ergebnis hervorbringen (vgl. Q. Liu & Wu, 2012, S. 192). Um Modelle auf Daten anwenden zu können, bei der das Ergebnis noch nicht bekannt ist, müssen diese Modelle zuerst mit einer großen Menge gelabelter Daten trainiert werden. Auf die Sentiment Analyse angewendet werden dabei Texte als Daten mit dem Tonalitätwert als Label verwendet. Hauptsächlich werden dabei (Film oder Produkt) Rezensionen verwendet, bei denen bereits vorhandene Metadaten einen Hinweis in Richtung des Sentiments des Textes geben (Bajpai et al., 2019; Pang et al., 2002, Lui, 2012, S.31).

Pang & Lee führen daher auf, dass der Einsatz von Supervised Learning Methoden für das Public Opinion Mining weniger geeignet sei, da nur für die wenigsten Domänen Trainingsdaten zur Verfügung stehen und eine manuelle Generation mit enormen Zeitaufwand verbunden sei (Pang & Lee, 2008; S.25). Lui betont zudem, dass Supervised Learning Modelle sehr empfindlich auf den Fachbereich reagieren würden, auf dem es trainiert wurde. Daraufhin würden Transfer Learning Modelle in der Anwendung auf unbekannte Fachbereiche oft schlecht abschneiden. Zudem kommt dass Wörtern in unterschiedlichen Fachbereichen unterschiedliche Bedeutungen haben können (vgl. B. Liu, 2012, S.38).
Aus diesem Grund sind Unsupervised Learning Modelle für das Public Opinion Mining interessanter und werden nachfolgend näher betrachtet. Diese sind im Vergleich zum Supervised Learning nicht so sehr verbreitet (vgl. Petz, 2019, S.65).

Es können in der Literatur Drei Analyseebenen identifiziert werden. Die höchste Abstraktionsstufe ist dabei die Dokumentenebene, bei der versucht wird die Stimmungsrichtung eines ganzen Dokuments zu bestimmen. Es wird dabei angenommen, dass das gesamte Dokument nur eine Meinung zu einer Entität enthält. Analog dazu wird auf Satzebene angenommen, dass ein Satz nur eine Meinung zu einer Entität enthält. Am spezifischsten kann die Problemstellung auf Aspektebene betrachtet werden. Hierbei wird versucht das oben aufgestellte Quintupel zu befüllen. Dieser Ansatz resultiert in der höchsten Komplexität, da alle Fünf Elemente des Quintuple aus dem Text extrahiert werden müssen (vgl. B. Liu, 2012, S.10f).

Für die Sentiment Analyse, also die tatsächliche Bestimmung der Tonalität, greifen Autoren wie Ding et al., Kim & Hovy und Taboada et al. im Bereich des Unsupervised Learning in Bezug auf alle Analyseebenen auf Wörtbuch-basierte Ansätze zurück (vgl. Ding et al., 2008; vgl. Kim & Hovy, 2004; vgl. Taboada et al., 2011). Dabei wird zuerst ein Wörtbuch mit Sentimentwörten aufgebaut und anschließend der Tonalitätswert einer Texteinheit mit Hilfe einer Funktion beziehungsweise Algorithmus bestimmt. Sentimentwörter sind Wörter innerhalb des Wörterbuchs, die mit einem Tonalitätswert annotiert sind. Für die Berechnung der Stimmungsrichtung einer Texteinheit greift die Funktion dabei auf die Sentimentwörter des Wörterbuchs zu. (vgl. Pang & Lee, 2008, S.27).

Einen Überblick von dem beispielhaften Aufbau eines solchen Algorithmus auf Aspektebene geben Ding et al. in ihrer Arbeit. Zuerst sollen alle Sentimentwörter des Satzes den im Wörterbuch definierten Tonalitätswert zugewiesen bekommen. Anschließend werden sogenannte Sentiment Shifter angewendet, die aus Worten bestehen, welche die Stimmungsrichtung verändern. Ein grundlegendes Beispiel ist das Wort *nicht*, dass wenn vorangestellt die Stimmungsrichtung des darauffolgenden Sentimentworts umkehrt (beispielsweise nicht gut). Adversative Konjunktionen wie zum Beispiel mit *aber* oder *jedoch* soll nach Ding et al. auch berücksichtigen werden, indem die Stimmungsrichtung nach der Konjunktion gegenteilig zu der Stimmungsrichtung vor der Konjunktion zu werten ist (vgl. Ding et al., 2008). Zum Schluss sollen alle Stimmungsrichtungen zu einem Wert aggregiert werden. Dabei sollen Sentimentwörter, die näher am Aspekt liegen eine höhere Gewichtung bekommen, wie jene die weiter entfernt sind (vgl. Ding et al., 2008). Andere Autoren wie Hu & Liu gehen wiederrum sehr simpel vor und berechnen lediglich den Durschschnitt der identifizierten Tonaliäten (vgl. Hu & Liu, 2004).

Die Generierung der zugrundeliegenden Wörterbücher kann auf der einen Seite manuell mit Hilfe menschlicher Annotation erfolgen. Diese Vorgehensweise ist unvermeidlich mit enormen Zeitaufwand verbunden, dafür aber mit hoher Qualität (vgl. Petz, 2019, S.38). Kim & Hovy stellen auf der anderen Seite einen automatischen Ansatz vor. Sie schlagen vor mit einer kleinen Zahl von manuell annotierten Sentimentwörten zu starten, der sogenannten Seed Liste. Für die Begriffe der Seed Liste sollen automatisch Synonyme bzw. Antonyme aus Wörterbüchern wie WordNet gefunden und die Liste erweitert werden. Dieser Prozess soll solange wiederholt werden, bis keine neuen Wörter mehr zu identifizieren sind (vgl. S.-M. Kim & Hovy, 2004).

3 Methodik

In diesem Kapitel wird die Methodik vorgestellt, mit dem durch Überprüfung der Hypothese die Beantwortung der Forschungsfrage gelingen soll. Es wird ein quantitativer Ansatz entwickelt, der die wesentliche Leistung der Demoskopie automatisiert und mit Hilfe der Mittel der Wirtschaftsinformatik umsetzt. Unter Demoskopie wird das Abbilden der öffentlichen Meinung aus Einzelmeinungen von Bürgern in leicht verständlichen Kennzahlen verstanden (Hippner & Rentzmann, 2006; Miner et al., 2012). Diese Kennzahlen werden aus den Ergebnissen von Text-Mining Verfahren erzeugt und daraufhin Zusammenhänge zwischen der Tonalität des deutschen und dem anglo-amerikanischen Sprachraums sichtbar machen (vgl. Wiedemann & Niekler, 2016, S.64).

Zusammenfassend lässt sich der quantitative Ansatz in drei Schritten beschreiben. Im ersten Schritt werden die Variablen der Hypothese operationalisiert (siehe Kapitel 3.1 Operationalisierung). Anschließend wird eine Messung der Indikatoren unter Verwendung der spezifischen Erhebungsmethode durchgeführt (siehe Kapitel 3.2). Im letzten Schritt werden die erhobenen Daten durch einen statistischen Hypothesentest ausgewertet und somit eine Verifizierung der Hypothese ermöglicht (siehe Kapitel 3.3 Datenanalyse).

3.1 Operationalisierung

Im ersten Schritt der Methodik wird die Operationalisierung durchgeführt, welche den Vorgang beschreibt, die Variablen der Hypothesen in messbare Merkmale zu überführen (vgl. Raithel, 2006, S.8). Abgeleitet aus der Hypothese ist die Variable in dieser Arbeit die Tonalität eines Sprachraums zum Thema Digitalisierung.

Für die empirische Überprüfung der Hypothese müssen der theoretischen Ebene konkret messbare Sachverhalte zugeordnet werden. Dafür werden Indikatoren anhand der Literatur identifiziert, die diese Variable durch einen beobachtbaren Sachverhalt repräsentieren (vgl. Tausendpfund, 2018, S.107f). Dieser Zusammenhang ist in Abbildung 3 visualisiert. In dieser Arbeit wird die Stimmungrichtung der Online-Angebote von Nachrichten und Zeitschriften je Sprachraum als Indikatoren herangezogen, da diese als Informationsfunktion massiv die Meinungsbildung beeinflussen und so als Abbild der Öffentlichkeit verwendet werden kann (siehe Kapitel 2.1 Forschungsstand).

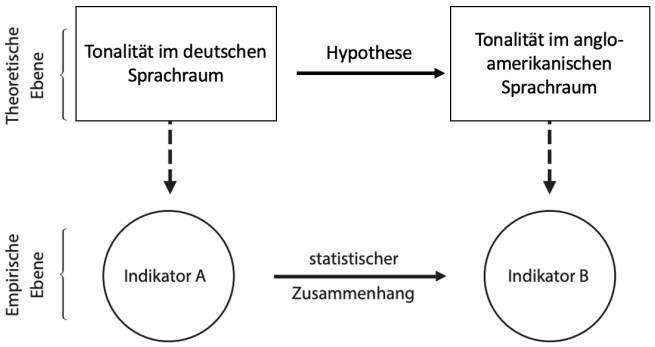


Abbildung 3: Zusammenhänge der Operationalisierung

Quelle: Tausendpfund, 2018, S.108 (geändert)

Auf die generelle Machbarkeit der Nutzung von Nachrichtenartikel für die Sentimentanalyse deuten die Arbeiten von Bleich & van der Veen, Shirsat et al und Shapiro et al. hin (vgl. Bleich & van der Veen, 2018; vgl. Shapiro et al., 2020; vgl. Shirsat et al., 2017).

3.2 Erhebungsmethode

Nachdem im vergangenen Kapitel die Indikatoren definiert wurden, wird im zweiten Schritt die Messung der Indikatoren durchgeführt. Die Messung erfolgt anhand der Erhebungsmethode und wird in diesem Kapitel ausführlich beschrieben.

Im Gegensatz zur standardisierten Umfrage, die in bisherigen Forschungsarbeiten eingesetzt wurden, wird in dieser Arbeit ein Public Opinion Mining System als Erhebungsmethode implementiert. Aus dem Forschungsstand geht hervor, dass diese Systeme für automatisierte die Abbildung der öffentlichen Meinung geeignet sind.

Die vorgeschlagene Vorgehensweise in dieser Arbeit ist an den von Fayyad et al. vorgestellten Prozess für "Knowledge Discovery in Databases" (Synonym für Data Mining) angelehnt, der den Gesamtprozess zur Mustererkennung in großen Datenmengen bezeichnet (vgl. Fayyad et al., 1996). Dieser allerdings etwas angepasst werden, vor allem, da in Bezug auf die international ausgerichtete Forschungsfrage Anforderungen hinzukommen. Ferner wird der Aufbau einer wissenschaftlichen Arbeit berücksichtigt, indem der Prozess lediglich als Mittel der Datenerhebung eingesetzt wird. Damit wird die Generierung von Wissen aus den gefundenen Mustern in das Kapitel 3.3 Datenanalyse in Zusammenhang mit der Interpretation der Ergebnisse in Kapitel 5 Diskussion ausgelagert.

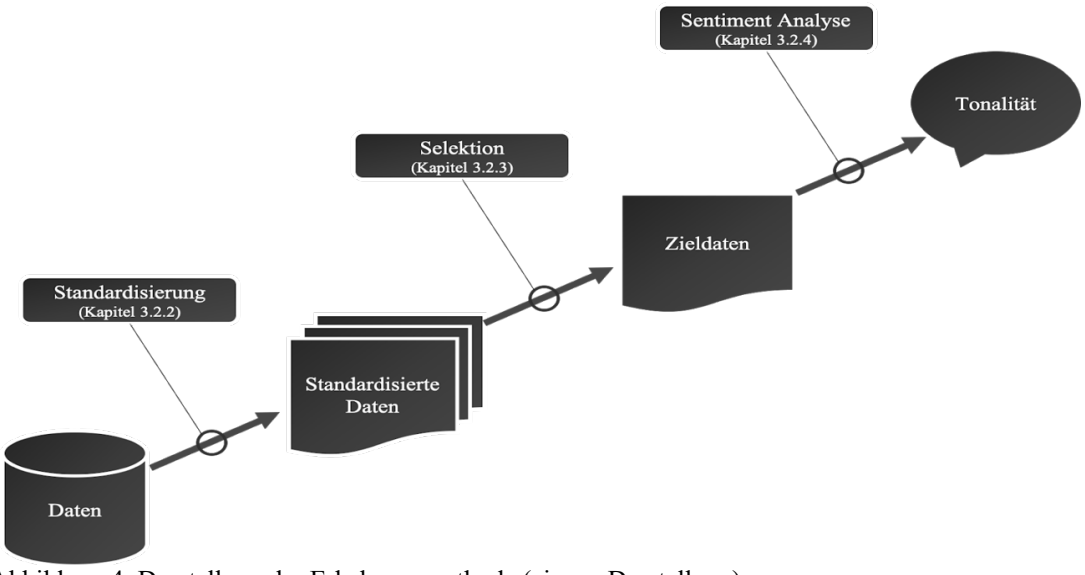


Abbildung 4: Darstellung der Erhebungsmethode (eigene Darstellung)

Quelle: inspiriert durch Fayyad et al., 1996, S.41

In Abbildung 4 sind die einzelnen Schritte der Erhebungsmethode dargestellt und dient als Leitfaden der folgenden Kapitel. Im ersten Schritt werden dabei relevante Quellen identifiziert und anschließend die Daten mittels Webscraping erhoben (siehe Kapitel 3.2.2 Datensammlung). Um während der späteren Datenanalyse eine Vergleichbarkeit zwischen den Sprachräumen zu schaffen, werden die Daten als nächstes standardisiert (siehe Kapitel 3.2.3 Standardisierung). Daraufhin wird die Datengesamtheit auf das untersuchungsrelevante Material eingeschränkt (siehe Kapitel 3.2.4 Datenselektion). Abschließend werden die Daten mittels Sentiment Analyse quantifiziert (siehe Kapitel 3.2.5 Sentiment Analyse)

3.2.1 Architektur

Bevor allerdings die weitere Methodik näher beschrieben wird, soll in diesem Abschnitt die Architektur des Systems skizziert werden. Die Architektur des vorgeschlagenen Systems wird in Abbildung 5 dargestellt und ist in dieser Arbeit von besonderer Relevanz, da sie die Beantwortung des Forschungsziels ermöglicht.

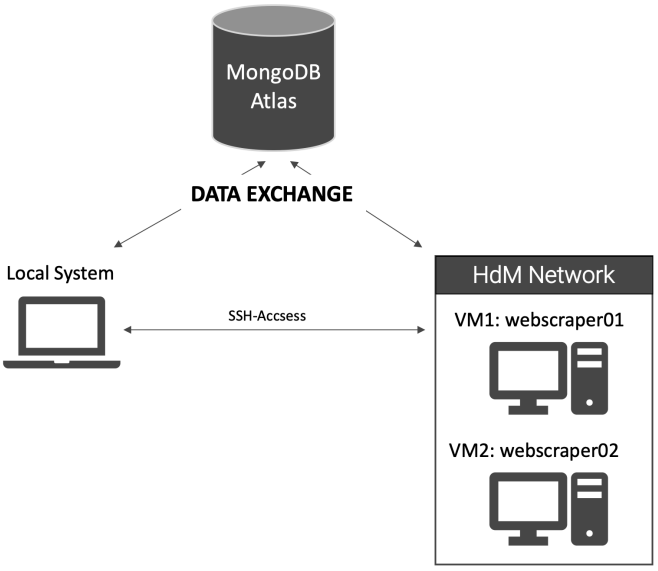


Abbildung 5: Architektur des Systems

Alle Komponenten dieser Arbeit werden mit Hilfe von Python-Scripts implementiert, welche parallel auf mehreren Servern ablaufen, um den Berechnungsaufwand zu verteilen. Hierfür kommen eine lokale Maschine und zusätzlich zwei virtuelle Maschinen im Netzwerk der Hochschule der Medien zum Einsatz, die über einen SSH-Zugriff bedient werden. Eine Aufteilung der Rechenlast ergibt vor allem Kontext des Forschungsziels einige relevant Vorteile. Einerseits kann so die Datensammlung per Webscraper durchgängig über einen größeren Zeitraum in einem isolierten System laufen. Andererseits erzeugen die gesammelten Textdaten durch ihre Menge und hohe Dimensionalität enormen Rechenaufwand.

Damit die implementieren Analysefunktion miteinander interagieren können, muss eine einheitliche Datenbasis geschaffen werden. Aufgrund der von Wei-ping et al. und Gyorodi et al. vorgestellten verbesserten Performance in der Bearbeitung von unstrukturierten Texten, gegenüber traditionellen sequenziellen Dankenbanken, wird auch in dieser Arbeit eine NoSQL-Datenbank für die Speicherung der Daten verwendet (vgl. Gyorodi et al., 2015; vgl. Wei-ping et al., 2011). Der Begriff *NoSQL*

steht für “non sequal” und beschreibt eine Art von Datenbank, die kein vordefiniertes Datenbankschema für die Speicherung der Daten benötigt (vgl. Meier, 2018, S.9). Konkret wurde eine MongoDB ausgewählt, da diese Daten dokumentenorientiert in Form von JSON-Objekten speichert und infolgedessen eine hohe Kompatibilität mit der gewählten Programmiersprache Python für die Implementierung von Quellcode vorliegt.
Weiterhin ist die Verfügbarkeit der Datenbank in Zusammenhang mit den Webscrapern von großer Bedeutung, da andernfalls wertvolle Daten während der Datenerhebung verloren gehen würden. Um diese hohe Verfügbarkeit zu gewährleisten wurde die MongoDB auf der Google Cloud Platform gehostet.

Für die Synchronisierung des Quellcodes auf die verschiedenen Systeme wird das Versionierungssystem Git mit einem GitHub Repository verwendet.

3.2.2 Datensammlung

In diesem Abschnitt wird die Datenerhebung dargestellt, welche zuerst auf die Quellenselektion eingeht und anschließend vorstellt wie die Daten der Quellen gesammelt und gespeichert werden.

3.2.2.1 Quellselektion

Für die Auswahl der Datenquellen werden zunächst Qualitätsblätter herangezogen, da diese als Leitmedien gelten und für die weitere Generierung von Inhalten in anderen Medien genutzt werden (vgl. Wilke, 1999, S.60). Mit diesem Effekt kann durch das Analysieren eines Qualitätblatts potentiell der Inhalt vieler verschiedeneren Quellen abgebildet werden und so ein breites Meinungssprekturm erfasst werden. Die Auswahl wird ferner mit Online-Zeitschriften angereichert, bei denen vermutet werden kann, dass die Digitalisierung in Beiträgen thematisiert wird.
In Bezug auf die Sprachräume wurden die USA und Großbritannien als Repräsentation für den anlgo-amerikanischen Sprachraums und Deutschland für die Repräsentation des deutschen Sprachraums gewählt.

Für die Suche nach geeigneten Quellen wurden die Kriterien Suchfunktion, Verfügbarkeit und Scrapability entwickelt, um enorme Menge an Webseiten systematisch Analysieren zu können. Diese werden nachfolgend beschrieben.
Das Kriterium der Suchfunktion beschreibt den Zustand, ob die Webseite eine Möglichkeit zur Verfügung stellt, um die Gesamtheit der Artikel mittels Stichwörter zu durchsuchen. Andernfalls würde eine große Menge von nicht untersuchungsrelevanten Artikeln gesammelt werden, die nicht zum Thema Digitalisierung verfasst wurden und zu einem späteren Zeitpunkt wieder identifiziert und entfernt werden müssten. Dieses Kriterium verringert zwar die Anzahl der potentiellen Quellen, verringert auf der anderen Seite aber auch das Datenvolumen des gesamten Korpus und folglich den Berechnungsaufwand.
Weiterhin wird mit dem Kriterium Verfügbarkeit überprüft, ob von der Suchfunktion der jeweiligen Seite eine ausreichend große Menge gefunden wird. Für deutsche Webseiten wurde das Stichwort "*Digitalisierung*" und für englische Webseiten die Stichwörter "*Digitisation*", "*Digitization*", "*Digitalisation*" und "*Digitalization*"

verwendet. Der Schwellwert wird auf die Anzahl von 500 Artikel festgelegt und alle Webseiten, die ihn unterschreiten nicht länger betrachtet.

Nicht zuletzt wird mit dem Kriterium Scrapability überprüft, inwiefern sich die Webseite eignet, um Artikel mit automatisierten Webscrapern herunterzuladen. Einige Quellen, wie die us-amerikanische Zeitung "*The Washington Post*", erfordern eine Authentifizierung, da der Abruf der Artikelvollansicht nur mit einem gültigen Abonnement möglich ist. Andere, darunter die britische Zeitung "*Financial Times*", schützen ihre Webseite mit automatisierten Turing Tests, sogenannt "Completely Automated Public Turing Test To Tell Computers and Humans Apart"-Tests (CAPTCHA). Diese sind vorerst nur effizient von Menschen lösbar und werden verwendet, um Webseiten vor Bots zu schützen. (vgl. von Ahn et al., 2004). Es muss je Quellseite überprüft werden, ob Möglichkeiten mittels Webscraper gefunden werden können, um an die Artikelvollansicht zu gelangen und schließlich den untersuchungsrelevanten Artikeltext speichern zu können.

Sprachraum	**Name der Quelle**	**Scrapability**	**Suchfunktion**	**Verfügbarkeit**	**Auswahl**
en (us)	The Wall Street Journal	Möglich	Vorhanden	nicht Ausreichend	Nein
en (us)	The New York Times	Möglich (sogar API)	Vorhanden	Ausreichend	Ja
en (us)	Los Angeles Times	Möglich	Vorhanden	Ausreichend	Ja
en (us)	The Washington Post	nicht performant mit http-requests realisierbar, authentication benötigt JavaScript	Vorhanden	Ausreichend	Nein
en (us)	Forbes	Möglich	Vorhanden	Ausreichend	Ja
en (uk)	Financial Times	Möglich (Anmeldedaten)	Vorhanden	Ausreichend	Ja
en (uk)	The Sunday Times	unbekannt	Nicht Vorhanden	keine Information	Nein
en (uk)	The Daily Telegraph	unbekannt	Nicht Vorhanden	keine Information	Nein
de	Süddeutsche Zeitung	Möglich	Vorhanden	Ausreichend	Ja
de	Frankfurter Allgemeine Zeitung	Nicht Möglich	Vorhanden	Ausreichend	Nein
de	Zeit Online	Möglich (Anmeldedaten)	Vorhanden	Ausreichend	Ja

Tabelle 1: Ergebnisse der Analysekriterien der Webseitevorauswahl

In Tabelle 1 sind die Ergebnisse der Analyse je Quelle und Kriterium festgehalten. Nur wenn alle drei der vorgestellten Kriterien zutreffen, kann ein Webscraper für die

Quelle zur Erreichung des Forschungsziels implementiert werden. Schließlich fiel die Auswahl auf die Zeitungen "*The New York Times",* "*Los Angeles Times*", "*Financial Times*" und die Zeitschrift "*Forbes*" für den englischen Sprachraum. Für den deutschen Sprachraum wurden die Zeitungen "*Süddeutsche Zeitung*" und "*Zeit Online*" selektiert.

3.2.2.2 Webscraper

Die Aufgabe der Webscraper besteht darin den initialen Korpus für diese Arbeit zu schaffen. Im Zuge dessen wird versucht automatisiert die Sprachdaten (hier Artikeltext), die Überschrift und einige Metadaten zu extrahieren und anschließend persistent für weitere Untersuchung abzuspeichern. Besonders relevant bei den Metadaten ist das Veröffentlichungsdatum des Artikels, da es bei der Datenselektion (siehe Kapitel 3.2.4 Datenselektion) als Analysekriterium verwendet wird.

Die Webscraper werden in dieser Arbeit mittels Python Skripte und der Library BeautifulSoup implementiert, die häufig für diesen Zweck eingesetzt wird (vgl. Lawson, 2015, S.26). Der Aufbau und vor allem die technische Umsetzung der Suchfunktion aller ausgewählten Webseiten ist ähnlich aufgebaut, sodass sich der Aufbau der Webscraper nur wenig voneinander unterscheiden. Vorteilhaft ist, dass alle potentiell relevanten Zielseiten zentralisiert als Suchergebnis angezeigt werden. Als Suchbegriff werden die bei der Quellselektion genannten Begriffe verwendet.

3.2.3 Standardisierung

Anhand der Forschungsfrage und den selektierten Quellen wird klar, dass ein Teil des gesammelten Datensatzes in Form von deutschen Texten und ein anderer Teil in Form von englischen Texten vorliegen wird. Dabei Eine multilinguale Sentiment Analyse ist eine große Herausforderung, vor allem, da die zur Verfügung stehenden Python Libraries zur Sentiment Analyse nicht für multilinguale Korpora ausgelegt sind (siehe Kapitel 3.2.5 Sentiment Analyse). Diese werden nur für die Verwendung von einer Sprache entwickelt, sodass die Sprache der Artikel innerhalb des Korpus vereinheitlicht werden muss.
Zusätzlich ist die Standardisierung der Texte vor der Durchführung der Datenselektion (Kapitel 3.2.3) unerlässlich, da Algorithmen verwendet werden, bei denen ein multilingualer Korpus von Nachteil ist.

Das Forschungsgebiet der Sentimentanalyse wurde hauptsächlich unter Verwendung englischer Texte vorangetrieben, womit ein Großteil der Ressourcen nur für die englische Sprache zur Verfügung stehen (vgl. Petz, 2019, S.5; vgl. Liu, 2012, S.53). Vereinzelt stehen jedoch Ressourcen für die Sentiment Analyse von deutschen Texten zur Verfügung. Diese könnten nach der Datenbereinigung genutzt werden, um Qualitätsverluste während einer Übersetzung zu vermeiden. Das Ergebnis der Unsupervised Sentiment Analyse hängt wie in den Grundlagen (siehe Kapitel 2.3.3 Quantifizierung einer Meinung) aufgezeigt stark von dem verwendetet Wörterbuch ab, das zur Bestimmung verwendet wird. Wenn nun je Sprache unterschiedliche Python Libraries mit unterschiedlichen Wörterbüchern für die Bestimmung der Tonalität eingesetzt werden würden, könnte man die Vergleichbarkeit der Ergebnisse in Frage stellen. Denn es kann nicht verifiziert werden, ob derselbe Artikel von beiden Libraries dieselbe Tonalität zugewiesen bekommen würde. Somit würde die Verifizierung der Hypothese stark durch die unterschiedliche Art der Messung beeinflusst werden und nicht ausschließlich durch die Tonalität je Sprachraum. Aufgrund der dargestellten Problematik wird dieser Ansatz für diese Arbeit ausgeschlossen und Texte hinsichtlich der Sprache standardisiert.

Mohammad et al. und Araujo et al. zeigen, wie trotz Qualitätsverluste die automatische Übersetzung der Texte ins englische kompetitive Resultate gegenüber dem Stand der

Technik in anderen Sprachen hervorbringen kann (vgl. Araujo et al., 2016, S.1144f; vgl. Mohammad et al., 2016, S.125). Auch Denecke nutzt in Ihrer Arbeit die Übersetzung von Texten ins Englische als Herangehensweise für die Sentimentanalyse eines multilingualen Korpus (vgl. Denecke, 2008).

Daher werden in dieser Arbeit deutsche Texte mittels der Google Translate API übersetzt, um die stärker entwickelten Ressourcen für das Englische nutzen zu können. Die Übersetzungsmaschine von DeepL ist im Vergleich zwar akurater als die von Google Translate, kann jedoch aufgrund von zu hoher Nutzungskosten nicht zum Einsatz kommen (vgl. Petereit, 2020).

3.2.4 Datenselektion

Die Artikel werden während der Datenerhebung (siehe 3.2.2 Datensammlung) automatisch abgespeichert. Um die Forschungsfrage beantworten zu können, wird aus dem initialen Korpus (Datengesamtheit) ein untersuchungsrelevanter Korpus (Teilmenge) gebildet. In diesem Kapitel wird die aus zwei Schritten bestehende Datenselektion beschreiben.

3.2.4.1 Zeitliche Selektion

Zum einen werden Artikel anhand ihres Veröffentlichungsdatum selektiert, da die Digitalisierung mit einer rasanten Dynamik voranschreitet und sich damit die Meinung der Öffentlichkeit zum Thema Digitalisierung im Zeitverlauf ändern kann. So wäre es nicht sinnvoll aktuelle Artikel mit zehn Jahre altern Artikeln zu vergleichen. Als zeitliche Eingrenzung werden für die Erstellung des untersuchungsrelevanten Korpus im ersten Schritt nur jene Artikel selektiert, die nach dem 1. Januar 2019 veröffentlicht wurden.

3.2.4.2 Inhaltliche Selektion

Während der Datensammlung werden nur solche Artikel gespeichert, die von der Suchmaschine der jeweiligen Quellseite zu den Stichwörtern angezeigt wurden. Da aber die Funktionsweise der Suchmaschinen und Qualität der Suchergebnisse unbekannt ist, muss der Inhalt des Artikels weiter analysiert werden. Dabei ist durch manuelle Vorprüfung aufgefallen, dass die Suchmaschinen auch solche Artikel listen, die lediglich einmal das Wort Digitalisierung enthalten aber zu einem anderen Thema geschrieben wurden. Da die Stimmungsrichtung in solchen Artikeln nicht auf das in der Forschungsfrage definierte Thema abzielt, würden solche daher das Ergebnis verfälschen und damit die Aussagekraft verringern. Es wäre also nicht zielführend themenfremde Artikel in die Beantwortung der Forschungsfrage einfließen zu lassen. Um die Datenqualität und somit die Aussagekräftigkeit der Ergebnisse zu erhöhen, sollen im zweiten Schritt aus dem resultierenden Koprus der zeitlichen Selektion nur jene Artikel selektiert werden, die die Digitalisierung thematisieren.

Da bei der Datenerhebung eine große Menge von Artikeln gespeichert werden, wäre es nur mit erheblichem Zeitaufwand möglich jeden einzelnen zu lesen und manuell die

Themenstellung zu erfassen. Eine manuelle Vorgehensweise scheint also nicht effizient genug und im Rahmen dieser Arbeit nicht tragbar.
Für die automatische Identifizierung von Themen in großen Mengen von Artikeln wurden in der Literatur bereits Ansätze entwickelt. Die sogenannten Topic Modelling Ansätze sind statistische Methoden, welche die Wörter der Texte analysieren, um die darin enthaltenen Themen zu erkennen und entsprechend zuzuweisen (vgl. Petz, 2019, S.91f). Sie versuchen so semantische Cluster mit Sinnzusammenhang in Textkollektionen zu identifizieren und werden oft dafür eingesetzt die im Internet entstehende Informationsflut zu bewältigen (vgl. Khalifa et al., 2013, S.51). In dieser Arbeit kann ein solcher Ansatz also für die inhaltliche Artikelselektion verwendet werden, um nur Artikel mit der Themenstellung Digitalisierung zu extrahieren.

Das von Blei et. al (2003) vorgestellte Latent Dirichlet Allocation (LDA) Verfahren liefert für die Problemstellung des Topic Modelling den bekanntesten Ansatz (vgl. Sbalchiero & Eder, 2020, S.1096; vgl. Blei et al., 2003). LDA ist ein generatives probabilistisches Modell, das davon ausgeht, dass jedes Thema eine Verteilung von Wörtern ist und dass jedes Dokument eine Mischung aus einer Menge von Themenwahrscheinlichkeiten besteht (vgl. Blei et al., 2003, S.993). Dabei kommt eine Bag-of-Words Darstellung zu Verwendung, die gemeinsam auftretende Begriffe auf Dokumentenebene auswertet (vgl. Petz, 2019, S.96). Da dabei Worthäufigkeiten eine große Rolle spielen, ist die zu vorige Standardisierung unerlässlich.

Die Funktionsweise des Modells lässt am besten Erklären, in dem die Annahme des Modells von der Vorgehensweise von Erstellung von Texten beschrieben wird. Hierfür wird ein Topic (zu deutsch Thema) als eine Verteilung über ein festes Vokabular definiert. Zum Beispiel beinhaltet das Topic Marketing mit hoher Wahrscheinlichkeit Wörter zum Thema Marketing und das Topic Sicherheit mit hoher Wahrscheinlichkeit Wörter zum Thema Sicherheit. Es wird angenommen, dass diese Topics mit zugehörigem Vokabular vor der Textgeneration erzeugt wurden. Daraus erschließt sich, dass die Anzahl der zu findenden Themen als Parameter mit in das Modell einfließt und ähnlich zum K-Means Clustering (auch zu finden im Bereich Unsupervised Learning) vor der Modellanwendung definiert werden muss (vgl. Aggarwal, 2015, S.162; vgl. Blei, 2011, S.3).

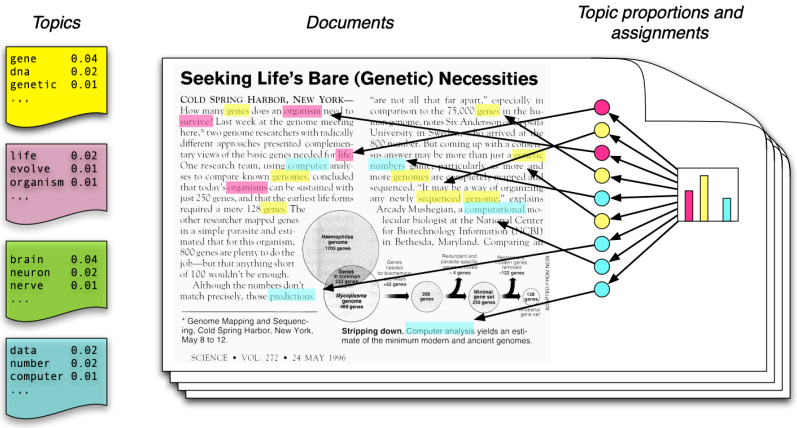


Abbildung 6: Grundprinzip von LDA

Quelle: Blei, 2011, S.3

Die Angenommene Texterstellung nach LDA besteht anschließend aus einem zweistufigen Prozess und ist nachfolgend beschrieben:

1. *Wähle eine Verteilung von Topics*
2. *Für jedes Wort in dem zu erstellenden Text:*
    a. *Wähle ein Topic aus der in (1) gewählten Themenverteilung*
    b. *Wähle ein Wort aus dem korrespondierenden Vokabular*

Das zugrundeliegende statistische Modell spiegelt diesen angenommen Prozess für die Texterstellung wider. Jedes Dokument weist also alle Topics in unterschiedlichem Verhältnis auf, was ein ausschlaggebendes Merkmal des LDA darstellt (1). Darüber hinaus stammt jedes Wort in jedem Dokument aus dem spezifischen Vokabular eines Topics (2b), wobei das ausgewählte Topic aus der Themenverteilung pro Dokument stammt (2a). Dieser generative Prozess wird in Abbildung 6 dargestellt. Das zentrale, rechnerische Problem besteht darin, aus dem Korpus auf die verborgene Themenstruktur zu schließen und kann als die Umkehrung des beschriebenen generativen Prozesses betrachtet werden (vgl. Blei, 2011, S.2ff).

Die gewählte Anzahl an Topics müssen wie oben beschrieben vor Modellanwendung definiert werden und beeinflusst maßgeblich die Ergebnisse. Eine zu kleine Anzahl von Themen würde zu breiten und heterogenen Themen führen. Im Gegenteil würde eine zu große Anzahl, zu spezifische Themen erzeugen. In beiden Fällen wäre es schwierig, die Wortgewichtungen des Themas zu interpretieren und den darin

beschriebenen Inhalt zu erfassen (vgl. Sbalchiero & Eder, 2020, S.1097). Das LDA Modell kann evaluiert werden, indem die erwartete menschliche Interpretierbarkeit betrachtet wird. Die Kennzahl der Themenkoheränz ergibt Aufschluss über diesen Sachverhalt. Dabei beschreibt der Begriff Kohärenz, ob sich eine Reihe von Aussagen oder Fakten gegenseitig unterstützen. Angewandt auf das Topic Modelling erfasst die Themenkoheränz den Grad der semantischen Ähnlichkeit zwischen hoch bewerteten Wörtern des Themenvokabulars (vgl. Lau et al., 2009, S.532).

Da in dieser Arbeit Daten automatisch erhoben werden, ist es unmöglich vor Modellanwendung auf die Anzahl der im Korpus vorkommenden Themen zu schließen. Um den optimalen Wert für diesen Parameter zu finden und so die spätere Interpretation zu vereinfachen, wurde eine einfache Rastersuche zur Hyperparameteroptimierung durchgeführt. In Abbildung 7 ist das Ergebnis der Rastersuche festgehalten. Dabei ist der beschriebene Effekt, dass sowohl eine zu geringe Anzahl an Themen als auch eine zu hohe Anzahl negativ auf die Qualität des LDA Modells auswirkt, entnehmbar.
Für den spezifischen Korpus wird aufgrund des Ergebnisses der Hyperparameteroptierung ein Modell mit 22 Topics verwendet.

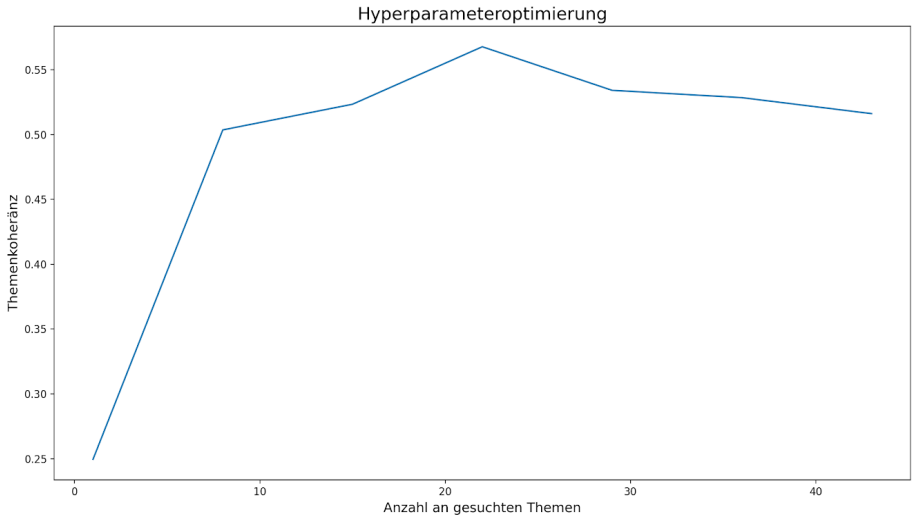


Abbildung 7: Ergebnis der Hyperparameteroptimierung

Das LDA Modell liefert als Ergebnis zum einen, die gefundenen Themen mit themenspezifischen Wortgewichtung des Vokabulars, allerdings keine Bezeichnung für das gefundene Topic. Zum anderen liefert es je Dokument und die Wahrscheinlichkeit, dass das Topic in dem Dokument enthalten ist. Anhand von

diesen Komponenten lässt sich nun eine semi-automatische, themenspezifische Datenselektion durchführen. Zuerst müssen alle jene Topics identifiziert werden, die das Thema Digitalisierung erfassen. Dieser Schritt erfolgt durch menschliche Interpretation der Wortgewichtungen. Hierfür werden die relevantesten 15 Begriffe für das jeweilige Topic ausgegeben werden.
Da die Interpretation der Begriffe zur Themenfeststellung der Topics sehr subjektiv ist und um eine erhöhte Aussagekräftigkeit der Interpretation zu ermöglichen, wurde eine Zwischenstudie durchgeführt. Dabei wurden drei Versuchspersonen unabhängig voneinander gebeten je Topic das für sie beschriebene Thema festzuhalten. Anschließend werden jene Topics selektiert bei den das Thema Digitalisierung interpretiert wurde. Schließlich können so alle Dokumente extrahiert, bei dem die gewählten Topics mit einer hohen Wahrscheinlichkeit in dem Dokument vorkommen. Nur diese werden im Kapitel 3.2.4 für die Sentiment Analyse verwendet.

Für die Implementierung des beschriebenen Konzepts zur inhaltlichen Selektion wurde die Python Library Gensim gewählt, die bereits über eine Implementierung des LDA Modells verfügt. Die Input-Daten für das LDA Modell ist der breites anhand des Datums selektierte und sprachlich standardisierte Korpus. Bevor die Artikeltexte für das Modell verwendet werden können, müssen diese zwei Vorbereitungsschritte durchlaufen. Die gewählten Schritte sind im Forschungsbereich NLP angesiedelt
Korpora, welche aus Nachrichtenartikeln bestehen, tendieren dazu ein sehr breites Vokabular zu verwenden (vgl. Martin & Johnson, 2015, S.111). Da Themen jedoch hauptsächlich durch Nomen ausgedrückt werden, kann wie in der Arbeit von Martin & Johnson (2015) der Korpus auf Nomen reduziert werden, um die größe des Gesamtvokabulars und somit die Komplexität des Modells zu reduzieren (vgl. Martin & Johnson, 2015, S.111f). Um spezifische Wortarten aus einem Text extrahieren zu können, müssen die darin vorkommenden Wörter annotiert werden. Die Zuordnung der Wortarten in Textdokumenten ist Teil der syntaktischen Analyse in und wird mit sogenannten Part-of-Speech (POS) Tagging Algorithmen umgesetzt (vgl. Rajman & Besançon, 1998, S.57). Nach der Zuweisung der Wortart werden im ersten Preprocessingschritt die Nomen der analysierten Sätzen behalten und alle anderen Wörter verworfen (vgl. Al Omran & Treude, 2017).

Der zweite Preprocessingschritt für das Topic Modelling ist die Lemmatisierung. Die Lemmatisierung beschreibt den Prozess, ein Wort auf sein Lemma zu reduzieren, also die Grundform eines Wortes (vgl. Kulkarni & Shivananda, 2019, S.54). So kann beispielsweise aus den verschiedenen Formen “sein”, “sind” und “war” die Grundform “ist” abgeleitet werden. Dieser Schritt führt zu einer weiteren Reduktion der Dimensionalität der Daten, da nur noch eine Wortform in der Bag-Of-Words Darstellung berücksichtig werden muss. Zudem beschreiben die Autoren Lau et al., durch die Lemmatisierung eine verbesserte Kohärenz der resultieren Topics zu erhalten (vgl. Lau et al., 2014, S.537).

Wichtig hierbei ist anzumerken, dass die für das Topic Modelling verarbeitete Version des Textes als Kopie gespeichert wird und nur für das Topic Modelling zur Verwendung kommt. Für die spätere Sentiment Bestimmung kommen andere Preprocessingschritte zum Einsatz.

3.2.5 Sentiment Analyse

In diesem Abschnitt wird die Quantifizierung der Daten vorgestellt, also die Transformation der Tonalität in messbare Zahlengrößen. Dafür werden die Artikel der Sprachräume mit Hilfe der Sentiment Analyse untersucht, wodurch die Stimmungslage messbar gemacht wird.

Die meisten Ansätze für die Sentimentbestimmung sind dem Supervised Learning zuzuordnen (vgl. Petz, 2019, S.65). Da der zuvor automatisch erstellte Korpus keine Labels mit der Stimmungsrichtung des Textes enthält muss in dieser Arbeit die Sentiment Analyse mittels Unsupervised Learning Verfahren durchgeführt werden. Eine eigenständige Entwicklung eines Tools zur Sentiment Bestimmung würde den Rahmen dieser Arbeit sprengen, sodass auf öffentlich zugängliche Tools gesetzt wird. Es wurden die von Loria et. al vorgestellte Library Textblob und die von Hutto & Gilbert vorgestellte Library vaderSentiment als die freizugänglichen Python Libraries identifiziert (vgl. Hutto & Gilbert, 2015; vgl. Loria et al., 2014). Des Weiteren sind der Öffentlichkeit zahlreiche Application Programmable Interfaceses (API) zugänglich, jedoch wurden diese aufgrund von hoher Nutzungskosten ausgeschlossen.

Die Qualität der Sentiment Analyse übt einen massiven Einfluss auf die Aussagekraft der Ergebnisse aus. Daher ist es von besonderer Relevanz diese Qualität hinsichtlich des Korpus zu maximieren. Da Wörter, Begriffe und sogar Sprachkonstrukte in anderen Domänen gänzlich andere Stimmungrichtungen ausdrücken können, ist die Wahl der Sentiment Library stark von dem spezifischen Koprus in dieser Arbeit abhängig (vgl. Petz, 2019, S.67f). Die identifizierten Libraries verfolgen ebenfalls den in der Literatur beschriebenen Wörtbuch-basierten Ansatz (siehe Kaptiel 2.3.3 Quantifizierung einer Meinung). Somit soll für diese Arbeit also die Library gewählt, die das Sentiment mit Hilfe ihren spezischen Regeln und ihrem spezischem Wörterbuch genauer abbilden kann. Da ein Unsupervised Learning Ansatz verwendet wird, kann die Qualität der Sentiment Libraries nicht ohne weiteres evaluiert werden, sodass vorab eine Stichprobe durchgeführt wird. Diese Stichprobe besteht 50 zufällig aus dem Korpus ausgewählten Artikeln, die manuell gelesen und mit einem Tonalitätswert zwischen -1 und 1 versehen werden. Anschließend wird die

automatische Sentiment Analyse derselben zufällig gewählten Korpus mit Hilfe des selektierten Python Libraries durchgeführt.

Dazu müssen beim Preprocessing die Daten zuerst Vorverarbeitet werden. Im Vergleich zu dem Preprocessing während der Datenselektion, fällt die Vorbereitung der Daten für die Sentiment Analyse deutlich schmaler aus. Zuerst werden mit Hilfe von Reguläre Ausdrücke alle Sonderzeichen, überflüssige Leerzeichen und Zahlen entfernt. Die einzige Ausnahme dabei ist das Ausrufezeichen, da dieses oft dafür eingesetzt wird die Stimmungslage in Texten zu transportieren. Danach wird die Tokenisierung durchgeführt, also dem aufteilen eines Textes in sinnvolle Elemente. Aus dem Forschungsziel in Zusammenhang mit dem quantitativen Ansatz wird klar, dass auf Dokuementenebene analysiert werden muss, um die Tonalität in einer zusammenfassenden Kennzahl auszudrücken und dadurch eine direkte Vergleichbarkeit zwischen den Sprachräumen zu schaffen.

In einer der bekanntesten Arbeiten zur Bestimmung der Stimmungsrichtung für ganze Dokumente, zeigt Turney eine Herangehensweise, die den Durchschnitt des Sentiments aller darin vorkommenden Sätze berechnet (vgl. Petz, 2019, S.65; vgl. Turney, 2002, S.1). Die Analyseebene in dieser Arbeit wird, angelehnt an Turney, auf Satzebene festgelegt, wonach sinnvolle Elemente bei der Tokenisierung ganze Sätze sind. Es werden zudem weitere Ansätze entwickelt, um das Sentiment ganzer Dokumente so realitätsnah wie möglich abzubilden. Diese sind von der Arbeit von Yang & Huang inspiriert und in Tabelle 2 näher beschrieben (vgl. Yang & Huang, 2019, S.2049).

	Beschreibung
Ansatz 1	Das Sentiment des Artikels ergibt sich aus dem Durchschnitt der Sentiments aller Tokens (Sätze) des Artikels (vgl. Turney, 2002).
Ansatz 2	Basierend auf dem Ergebnis vom Ansatz 1, wird zusätzlich auch das Sentiment der Überschrift berechnet welches mit doppeltem Gewicht in das Gesamtsentiment mit ein fließt. Die höhere Gewichtung wird dadurch begründet, dass die Überschrift des Artikels von besonderer

	Relevanz ist und dadurch bei dem Leser eine Grundstimmung für den Rest des Artikels erzeugt wird (vgl. Brinker et al., 2000, S.378).
Ansatz 3	In diesem Ansatz bekommt die Tonalität des ersten und letzten Fünftel des Artikelstextes eine höhere Gewichtung. Ähnlich zu Ansatz 2 besteht die Annahme, dass am Anfang und am Ende des Artikels relevantere Informationen aufzufinden sind. Der Gedanke besteht darin, dass zwar im Hauptteil mehrere Meinungen aufgeführt werden, allerdings zu Beginn und zum Ende die Kernaussagen des Artikels vermittels werden (vgl. Weischenberg, 2001, S.225).Falls durch eine zu kurze Artikellänge keine sinnvollen Abschnitte eingeteilt werden können, wird für diesen spezifischen Artikel das Ergebnis des Ansatz 1 verwendet.
Ansatz 4	Der komplette Artikeltext wird nicht tokenisiert und das Sentiment als Ganzes berechnet.

Tabelle 2: Beschreibung der verschiedenen Ansätze

Das Sentiment wird für jede der vorgestellten Ansätze mit beiden Libraries berechnet. Anschließend wird der Mittlere absolute Fehler (englisch Mean Absolut Error, abgekürzt MAE) mit der Gesamtanzahl der Stichproben T (T=50), dem manuell vergebenen Tonalitätwert x und dem automatisch berechneten Tonaliätswert x(dach) berechnet.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{x}_t - x_t\right|$$

Formel 1: Bestimmung des MAE

Quelle: Barrot, 2007, S.418

In der Tabelle 3 ist der MAE je Alternative und Python Library festgehalten:

	Textblob	**vaderSentiment**
Ansatz 1	*0.480*	*0.379*
Ansatz 2	*0.506*	*0.441*
Ansatz 3	*0.485*	*0.362*
Ansatz 4	*0.473*	*0.612*

Tabelle 3: MAE je Python Library und Ansatz

Mit diesen Ergebnissen kann nun die beste Kombination aus Ansatz und Sentiment Library für den Hauptdurchlauf gewählt werden. Für diese Arbeit wird der Ansatz 3 mit der Python Library vaderSentiment verwendet, da dieser mit 0.362 den geringsten MAE aufweist.

3.3 Datenanalyse

Im letzten Schritt der Methodik werden die erhobenen Daten analysiert. In diesem Kapitel wird dargelegt, wie anhand der gesammelten und aufbereiteten Daten die aufgestellte Hypothese H1 verifiziert wird.

Nach der Datenerhebung liegen nun Daten vor, die miteinander verglichen werden können. Wie in der Arbeit von Niekler & Wiedemann, werden Unterschiede in Subkollektionen der Artikelgesamtheit durch das Vergleichen der Ergebnisse aus Text-Mining-Verfahren sichtbar gemacht (vgl. Niekler & Wiedemann, 2016, S.64). In dieser Arbeit werden die Subkollektionen anhand des Sprachraums der Artikelquelle gebildet. Anschließend werden diese Stichproben mittels T-Test analysiert, um zu prüfen, ob sich die Sentiment Werte der Stichproben stark genug (auch signifikant) voneinander unterscheiden, um die so die Hypothese akzeptieren zu können. Der T-Test ist dafür das bekannteste statistisches Verfahren, mit dem mit dem nachgewiesen werden kann, ob sich die empirisch gefundenen Mittelwerte von 2 Stichproben systematisch voneinander unterscheiden (vgl. Kim, 2015, S.540; vgl. Rasch, 2008, S.43)
In dieser Arbeit wird ein unabhängiger t-Test durchgeführt, da die Variable der Hypothese, also die Tonalität von Artikeln in unterschiedlichen Sprachräumen, unabhängig voneinander zu sehen sind. Spezifischer wird in dieser Arbeit ein einseitiger t-Test verwendet, da geprüft werden soll, ob der Mittelwert der deutschen Artikel negativer ist als der Mittelwert der angloamerikanischen Artikel und damit eine gerichtete Hypothese vorliegt.

Mit Hilfe des t-Test wird überprüft, wie hoch die Wahrscheinlichkeit ist, dass das vorliegende Ergebnis zufällig entstanden ist. In anderen Worten wird die Wahrscheinlichkeit berechnet, dass die Hypothese Irrtümlich angenommen wurde und wird auch p-Wert bezeichnet (vgl. Kuckartz et al., 2010, S.163). Liegt diese Wahrscheinlichkeit unter dem vorab definierten Schwellwert, dem Signifikanzniveau, kann die aufgestellte Hypothese angenommen werden. In dieser Arbeit wird ein Signifikanzniveau (alpha) von 5 Prozent festgelegt, welches Häufig in der Literatur zum Einsatz kommt (vgl. Wasserstein & Lazar, 2016).

4 Ergebnisse

In diesem Kapitel werden die Ergebnisse der Methodikdurchführung präsentiert. Es konnten im Zeitraum Januar 2019 bis Juni 2020 mittels Webscraping der in Kapitel 3.2.1.1 (Quellselektion) gewählten Webseiten ein Gesamtkorpus bestehend aus 21.646 Artikeln und 107.337.053 Wörtern für die Untersuchung generiert werden.

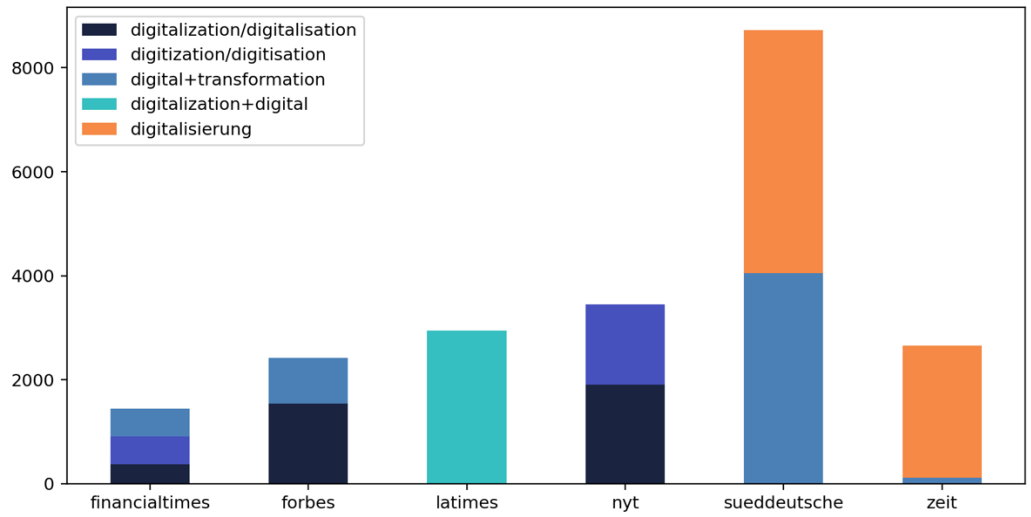


Abbildung 8: Anzahl der gesammelten Artikel je Stichwort und Quelle

In Abbildung 8 sieht man die Zusammensetzung des gesamten Korpus je Quellseite und Suchbegriff. Dabei wurden die Suchbegriffe nach amerikanischer Schreibweise mit "z" und nach britischer Schreibweise mit "s" zusammengefasst (*digitalisation* & *digitalization* sowie *digitisation* & *digitization*). Mit dem Suchbegriff *digital transformation* konnten Artikel bei Vier von Sechs Quellseiten gefunden werden. Von der deutschen Zeitschrift *Süddeutsche Zeitung* konnten mit 8.723 die meisten Artikel automatisch gesammelt werden. Zusammenfassend stehen als Rohmaterial 11.384 Artikel aus dem deutschen Sprachraum und 10.262 Artikel aus dem englischen Sprachraum zu weiterer Untersuchung zu Verfügung.

Darauf basierend wurde die Datenselektion durchgeführt. Das Ergebnis aller Selektionsschritt ist in Abbildung 9 festgehalten, um die Auswirkungen der verschiedenen Stufen zu visualisieren. Ein wenig reduziert waren nach der zeitlichen Selektion (Kapitel 3.4.1) noch 6.660 Artikel des englischen Sprachraums und 8.508 Artikel des deutschen Sprachraums übrig. Anschließend ergab die inhaltliche Selektion (Kapitel 3.4.2) eine weitere deutliche Verringerung der Artikelgesamtheit.

So bleiben 70 Prozent der Rohdaten nach der zeitlichen Selektion und noch 9,2 Prozent nach der inhaltlichen Selektion übrig.

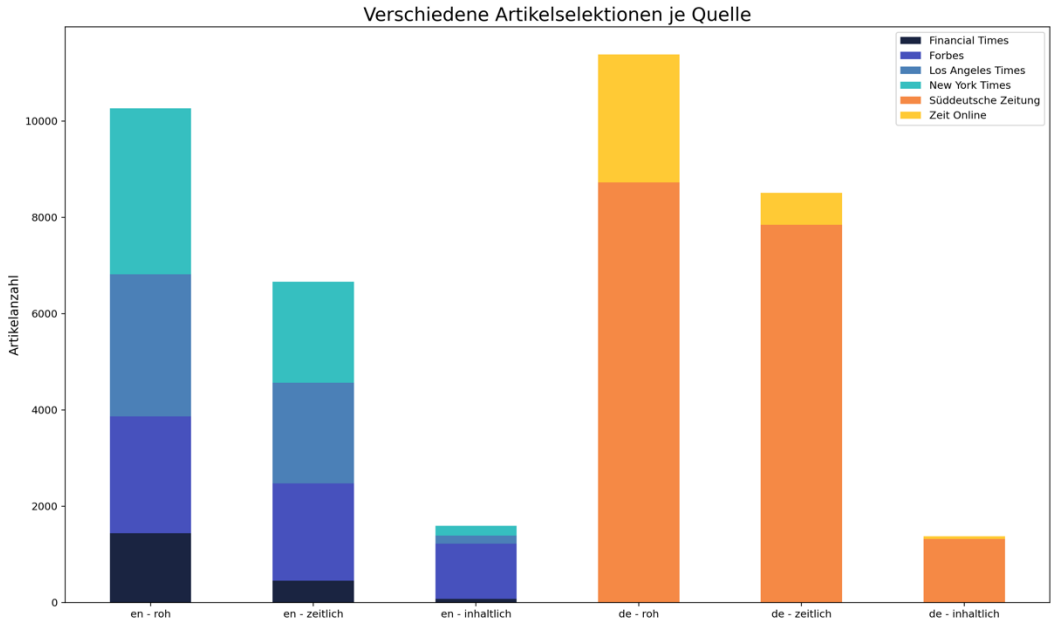


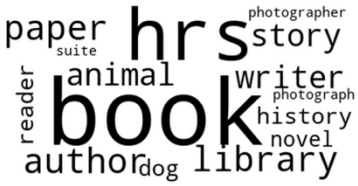
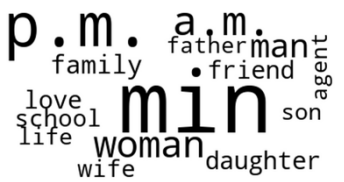
Abbildung 9: Anzahl der Artikel je Sprachraum, Quelle und Selektionsschritt

Zusammenfassend besteht der Korpus, der nach der Datenbereinigung als Untersuchungsmaterial für die Beantwortung der Forschungsfrage verwendet wird, aus **1.238** Artikel von englischen Quellseiten und **763** Artikel von deutschen Quellseiten. Die anglo-amerikanische Stichprobe setzt sich spezifischer hauptsächlich mit Artikeln Quelle Forbes (1.143 Artikel) zusammen. Am wenigsten bleiben nach der inhaltlichen Selektion bei der Financial Times bestehen (77 Artikel). Die deutsche Stichprobe weist eine ähnliche Struktur auf. Hier sind 1.319 Artikel von der Quelle Süddeutsche Zeitung und nur 58 Artikel von der Quelle Die Zeit übrig.

Nachfolgend wird noch spezifischer auf die Ergebnisse der inhaltlichen Selektion, also dem Topic Modelling, eingegangen, da dieser Schritt sehr komplex und von besonderer Relevanz ist. Zusätzlich wurden mit Hilfe dieser Ergebnisse Entscheidungen getroffen, die ausschlaggebend für die Reproduzierbarkeit dieser Arbeit sind.

Der Output des Modells sind die Wortgewichtungen je Topic und die Zuweisung der Artikel zu Topics. In Tabelle 4 sind die Wortgewichtungen je Topic dargestellt, die von dem Topic Modell erzeugt wurden. Wie in Kapitel 3.2.4.2 beschrieben werden die relevantesten 15 Begriffe des Topics für die Interpretation verwendet. Um die

Interpretation zu vereinfachen, werden diese als Wordcloud dargestellt. Dabei spiegelt die Relevanz des Begriffs für das jeweilige Topic die dargestellte Größe des Begriffs wider. So ist beispielsweise im Topic 0 das Wort “bank” wichtiger als das Wort “cash”. Die Interpretation der Versuchsteilnehmer der Zwischenstudie basieren auf diesen Wordclouds.

Topic 0	**Topic 1**
	
Topic 2	**Topic 3**
	
Topic 4	**Topic 5**
	
Topic 6	**Topic 7**
	
Topic 8	**Topic 9**
	

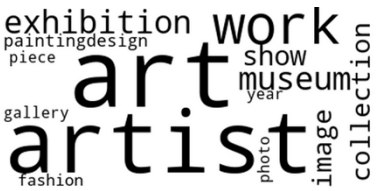
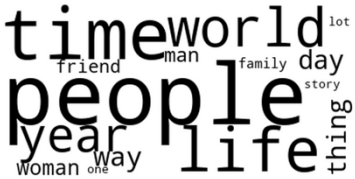
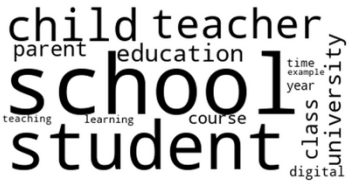
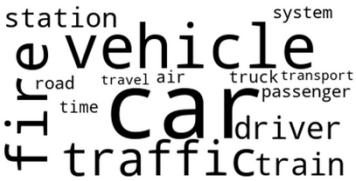
Topic 10	Topic 11
	
Topic 12	**Topic 13**
	
Topic 14	**Topic 15**
	
Topic 16	**Topic 17**
	
Topic 18	**Topic 19**
	
Topic 20	**Topic 21**
	

Tabelle 4: Wordclouds der Topics

Die Interpretationen der Versuchsteilnehmer sind dabei in Tabelle 5 festgehalten.

Topic	Versuchsperson 1	Versuchsperson 2	Versuchsperson 3	Interpretation
0	Finanzen	Bankenwelt	Geldgeschäfte	**Finanzwelt**
1	World Health	Coronakrise	Covid 19	**Coronakrise**
2	Kreativität	Buchauthor	Schreiben	**Bücher**
3	Digitalisierung	Digitale Welt	Digitalisierung	**Digitalisierung**
4	Darstellende Kunst	Kultur	Unterhaltung	**Kultur**
5	Familie	Familienzeit	Zeit	**Familienzeit**
6	Stadtplanung	Bürozentrum	Städteplanung	**Städteplanung**
7	Marketing	Kunde ist König	Kundenbeziehung	**Kunden**
8	Staat	Wahlen	Staat	**Staat**
9	Arbeitswelt	Mitarbeiterzufriedenheit	Arbeitsplatz	**Artbeitswelt**
10	Kunst	Kunstausstellung	Kunstszene	**Kunst**
11	Leben	Beziehungen	?	**Alltag**
12	Sport	Sportveranstaltung	Sport	**Sport**
13	Wirtschaft	Aktienwert	Börsennotiertes Unternehmen	**Unternehmen**
14	Bildung	Schulausbildung	Ausbildung	**Bildung**
15	Gesundheit	Medizinische Versorgung	Gesundheitssystem	**Gesundheitswesen**
16	Fortbewegung	Transportmittel	Mobilität	**Mobilität**
17	Politik/Wahlen	Wahlkampf	Wahl	**Wahlen**
18	Digitaler Kontent	Internetseite	Medien	**Digitale Medien**
19	Digitaler Staat	?	Digitales Geschäft	**Digitaler Staat**
20	Wirtschaftsinformatik	Datensicherheit	Informtionssysteme	**Wirtschaftsinformatik**
21	Fertigung	Stromherstellung	Industrie	**Fertigungsindustrie**

Tabelle 5: Interpretationen der Wordclouds je Versuchsperson

Letztlich wurde eine höhere Abstraktion gesucht, mit dem alle der Drei Interpretationen abgedeckt werden kann. Diese Abstraktion ist in der Spalte Interpretation festgehalten und dient zur Identifikation der Topics, die dem Bereich Digitalisierung zuzuordnen sind. Auch wenn diese Abstraktion nicht vollständig akkurat alle Einzelinterpreationen abdeckt, ist jedoch sehr gut Abbildbar, ob das jeweilige Topic in Richtung der Digitalisierung zeigt.

So konnten mit Hilfe der Zwischenstudie die Topics Digitalisierung, Digitale Medien, Digitaler Staat und Wirtschaftsinformatik als die Topics identifiziert werden, die vermutlich Themen der Digitalisierung beinhalten. Ausgeschlossen für die weitere Untersuchung sind alle jene Topics, die untersuchungsfremde Inhalte thematisieren. Beispiele hierfür sind Topic 12 bei dem einstimmig das Thema Sport oder Topic 15 bei dem das Thema Gesundheitswesen interpretiert wurde.

In Abbildung 10 wird zudem auch die Verteilung der Artikel je Topic vor der Datenselektion dargestellt. Dabei werden alle die Artikel gezählt bei der Wahrscheinlichkeit, dass das jeweilige Topic in dem Artikel vorkommt, größer als 30 Prozent ist. Die selektierten Topics orange markiert. Aus der Abbildung lässt sich entnehmen, dass während der initialen Datenerhebung viele Artikel selektiert wurden, bei denen das Thema Digitalisierung nicht vermutet werden kann. Darunter fällt auch das Topic Alltag, das den meisten Artikeln zugewiesen ist.

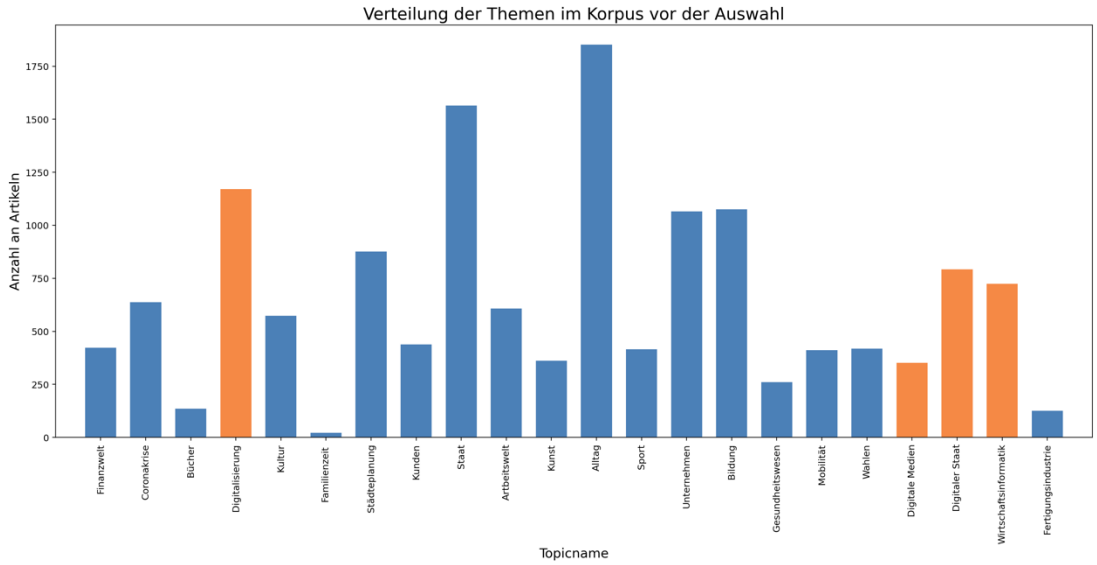


Abbildung 10: Anzahl der Dokumente je Topic vor der Auswahl

Aus dem finalen, bereinigten Korpus, der für die Untersuchung verwendet wird, sind in Abbildung 11 beispielhaft einige standardisierte (teilweise aus dem deutschen ins Englische übersetzt) Überschriften zufällig gewählter Artikel mit dazugehöriger Quelle und den höchstgewichtesten Topics gelistet:

„How To Choose KPIs For Your Digital Marketing Strategies“
– Forbes (Topic Digitale Medien & Kunden)

„Workers can learn to love artificial intelligence“
– Financial Times (Topic Digitalisierung & Artbeitswelt)

„Laptops and school WiFi can come“ **– Die Sueddeutsche (Topic Digitaler Staat & Bildung)**

„How Healthcare Organizations Can Protect Data Into The Next Decade And Beyond“ **– Forbes (Topics Digitalisierung & Gesundheitswesen)**

„What is a German foreign intelligence agency allowed to do?“
– Die Zeit (Topics Staat & Wirtschaftsinformatik)

Abbildung 11: Überschriften zufällig ausgewählter Artikel mit Topic und Quelle

Hier wird nochmals deutlich, dass Artikel eine Mischung aus unterschiedlichen Topics sind. Beispielsweise ist der Artikel mit der Überschrift „*How Healthcare Organizations Can Protect Data Into The Next Decade And Beyond?*“ eine Mischung der Topics Digitalisierung und Gesundheitswesen, was ein sehr treffendes Ergebnis ist.

Zuletzt werden die Ergebnisse der Datenanalyse vorgestellt. Hinzu ist in Abbildung 12 ein Histogramm der Tonalitätswerte je Sprachraum abgebildet. Deutlich erkennbar ist, dass die Verteilung der Artikel aus dem deutschen Sprachraum weiter links auf der X-Achse liegt als die Verteilung des anglo-amerikanischen Sprachraums. Die durchschnittliche Tonalität im deutschen Sprachraum (0.08) ist negativer als die durchschnittliche Tonalität im anglo-amerikanischen Sprachraum (0.195). Dabei ist plus Eins die maximal zu erreichende Tonalität im positiven Bereich und minus Eins die maximal zu erreichende Tonalität im negativen Bereich.
Der durchgeführte Hypothesentest ergab einen p-Wert von 2.22e-82. Dieser Wert liegt deutlich unter dem aufgestellten Signifikanzniveu von alpha=0.05. Mit dieser geringen Irrtumswahrscheinlichkeit, kann davon ausgegangen werden, dass die Ergebnisse nicht zufällig entstanden sind.

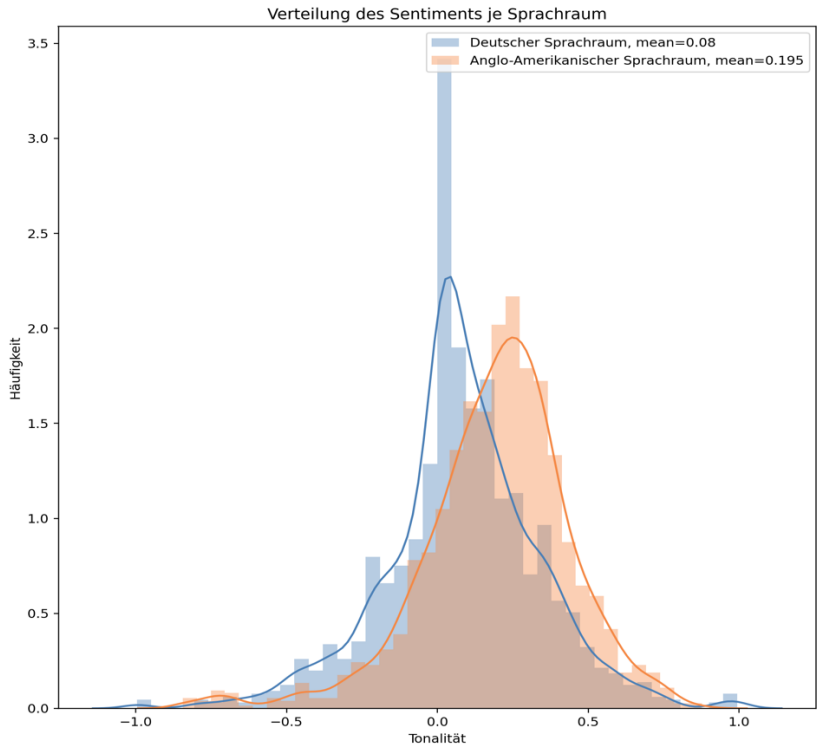


Abbildung 12: Verteilung der Tonalität je Sprachraum (Histogramm mit Gaussian Kernel Density Estimate)

Somit muss die Nullhypothese H0 “Im deutschen Sprachraum wird mit der gleichen Tonalität über die Digitalisierung gesprochen wie im anglo-amerikanischen Sprachraum” abgelehnt werden und die Hypothese H1 “Im deutschen Sprachraum wird mit einer negativeren Tonalität über die Digitalisierung gesprochen als im anglo-amerikanischen Sprachraum” angenommen werden.

Weiter differenziert sind in Abbildung 13 die durchschnittliche Tonalität je Quelle dargestellt. Hierbei wird nur der bereinigte Korpus betrachtet bei dem davon ausgegangen werden kann, dass die Digitalisierung thematisiert wird. Zusätzlich vermerkt sind hier nochmals die Anzahl der Artikel je Quelle.

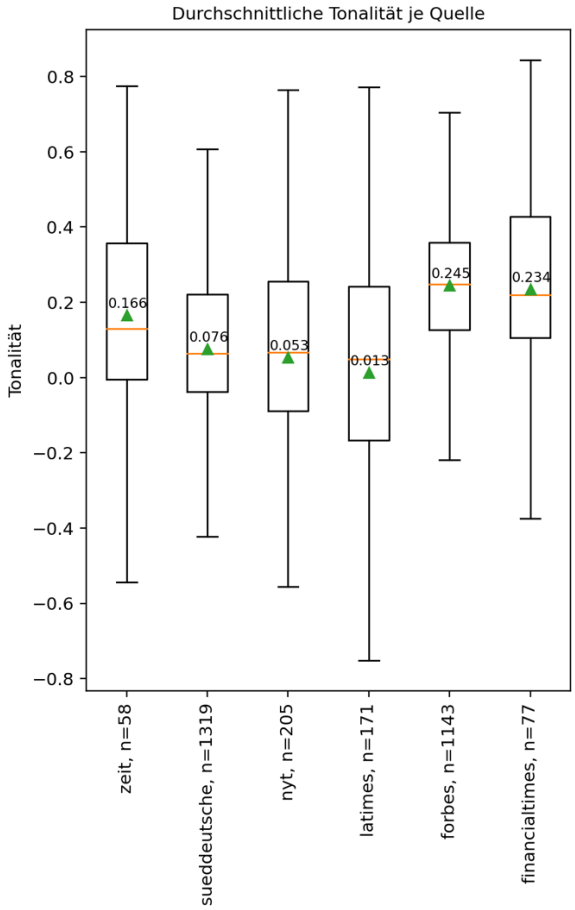


Abbildung 13: Durchschnittliche Tonalität je Quelle

So spricht die Quelle Forbes im anglo-amerikanischen Sprachraum durchschnittlich am positivsten über die Digitalisierung und im deutschen Sprachraum die Quelle Süddeutsche am negativsten.

Zuletzt wird der Sachverhalt nochmals in Abbildung 14 verdeutlicht, in der die Tonalität je Sprachraum im Zeitverlauf dargestellt ist. Dabei wurde für die Darstellung

Exponential Smoothing angewendet und eine Trendline je Sprachraum definiert. Auch in dieser Darstellung wird erneut verdeutlicht, dass der deutsche Sprachraum im Vergleich eine negativere Tonalität im fast kompletten Zeitraum aufweist. Es kann weiterhin anhand des negativen Vorzeichens der Trendlinie eine Verschlechterung des deutschen Sprachraums und anhand des positiven Vorzeichens eine leichte Verbesserung im anglo-amerikanischen Sprachraum hinsichtlich der Stimmungsrichtung gezeigt werden.

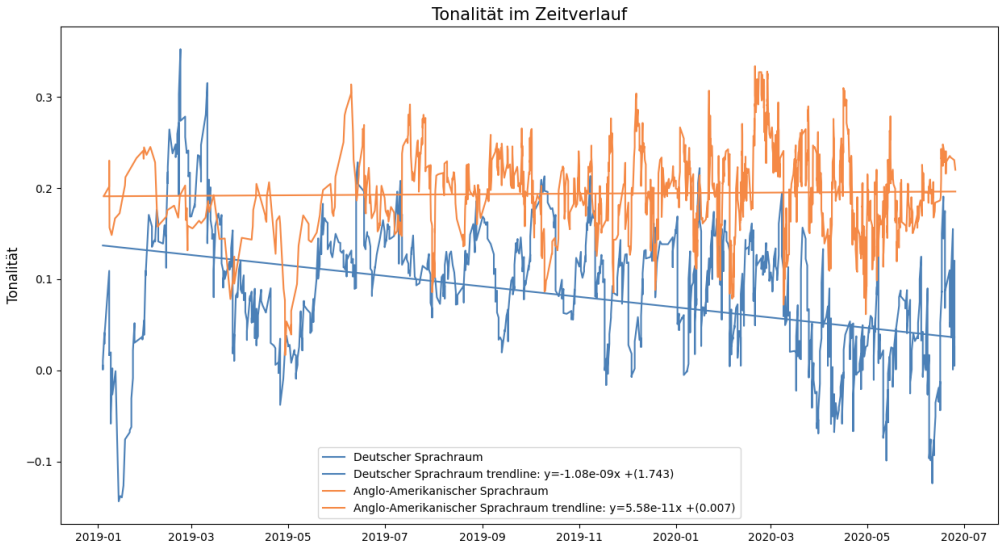


Abbildung 14: Tonalität je Sprachraum im Zeitverlauf mit Trendlinien

An dieser Stelle wird nochmals benotet, dass die vorgestellten Ergebnisse sich auf den Zeitraum Januar 2019 bis Juli 2020 beziehe, wie auf der X-Achse der Abbildung 14 entnehmbar.

5 Diskussion

In diesem Kapitel werden die vorgestellten Ergebnisse diskutiert, indem mögliche Erklärungen sucht werden und die resultierende Aussagekraft abgewägt wird. Ferner werden methodische Limitationen aufgezeigt und schließlich auf die Beantwortung der Forschungsfrage eingegangen.

Im ersten Schritt werden jedoch die Ergebnisse dieser Studie mit den Ergebnissen verwandter Studien eingeordnet. Wie im Forschungsstand vorgestellt, wurden zwei vergleichbare Studien im vergleichbaren Zeitraum zur Fragestellung identifiziert (siehe Kapitel 2.1 Forschungsstand). Obwohl diese Studien zu ähnlichen Fragestellungen verfasst wurden, gibt es einige Unterschiede zu dieser Arbeit. Während in den Studien von Störk-Biber et al. und dem Vodafone Institut für Gesellschaft und Kommunikation Versuchspersonen gebeten wurden konkrete Fragestellungen im Hinblick auf ihr jeweiliges Heimatland zu beantworten, wurde in dieser Arbeit lediglich die Stimmungsrichtung je Sprachraum mittels einer Kennzahl ausgedrückt, ohne dabei weiter inhaltlich differenzieren zu können (Störk-Biber et al., 2020; Vodafone Institut für Gesellschaft und Kommunikation, 2019). Ferner sind Länder nicht mit Sprachräumen vergleichbar, wodurch die Ergebnisse der Studien nicht direkt mit den Ergebnissen dieser Arbeit komparabel sind. Jedoch kann ein gemeinsamer Konsens in den Ergebnissen gefunden werden, wenn Länder aus dem jeweiligen Sprachraum als Vergleichsgröße herangezogen werden.
Störk-Biber et al. zeigen zum einen, dass die Meinung zum Thema Digitalisierung in Deutschland sehr zwiegespalten sei (vgl. Störk-Biber et al., 2020, S.30). In dieser Arbeit fällt der Median der Tonalität im deutschen Sprachraum mit 0.06 nahe Null. Das bedeutet, dass genau so viele Artikel mit negativer Tonalität wie Artikel mit positiver Tonalität im untersuchten Korpus vorliegen. Somit wird die von Störk-Biber et al. dargestellte Janusköpfigkeit auch in dieser Arbeit widergespiegelt. Zudem zeichnet sich auch in dieser Arbeit ab, dass es Analog zu Störk-Biber et al. in Deutschland keine allgemeine Ablehnung zur Technik gibt, da die durschnittliche Tonalität mit 0.09 des deutschsprachigen Raum im leicht positiven Bereich liegt (vgl. Störk-Biber et al., 2020, S.24). Allerdings ist die Entwicklung weiter zu betrachten, da wie in Abbildung 14 ersichtlich ein negativer Trend der Stimmungsrichtung vorliegt.

Weiterhin werden in den Studien von TechnikRadar und Vodafone Ergebnisse geliefert, die mögliche Gründe für die Verifizierung der Hypothese aufzeigen. Dazu TechnikRader zeigt in seiner Arbeit, dass es in Großbritannien (vgl. anglo-amerikanischer Sprachraum) ein größerer Anteil der Bevölkerung hinsichtlich Gesellschaft und Lebensqualität positive Auswirkungen von der Digitalisierung erwartet im Vergleich zu dem Anteil in Deutschland (vgl. deutscher Sprachraum) (vgl. TechnikRadar, 2019, S.16). Darüber hinaus werden in der Studie des Vodafone Institute for Society and Communications ähnliche Sachverhalte vorgestellt. Zum einen wird aufgeführt, dass die USA und Großbritannien (vgl. anglo-amerikanischer Sprachraum) einen höheren wahrgenommen Digitalisierungsgrad im Vergleich zu Deutschland aufweisen. Zum anderen glauben fast 60 Prozent der deutschen hinter dem Digitalisierungsgrad anderer Länder zurückbleibt (vgl. Vodafone Institute for Society and Communications, 2018, S.3). Aus diesen Gründen liegt die Vermutung nahe, dass im deutschen Sprachraum mit einer negativeren Tonalität im Vergleich zum anglo-amerikanischen Sprachraum gesprochen wird.

Als nächstes wird genauer auf die Datengrundlage eingegangen. Während der Datensammlung konnte ein sehr großer Korpus von mehr als 20.000 Artikeln gesammelt werden. Aus der Methodik ersichtlich wurde die Datenbereinigung stark thematisiert und sehr gründlich durchgeführt. Diese Bereinigung hat zu einer massiven Verkleinerung des Korpus geführt. Wie in Abbildung 9 zu sehen, führte die inhaltliche Selektion zu der größten Verkleinerung der Datengrundlage. Dabei ist die Qualität des verwendeten Topic Modelling Models ausschlaggebend für die Rechtfertigung dieser enormen Einschränkung. Da das verwendete Modell aus dem Unsupervised Learning Bereich stammt, ist eine Evaluierung der Datenqualität schwierig, da die tatsächlichen Themen der Artikel nicht als Metadaten zur Verfügung stehen. Es wird allerdings eine kleine Stichprobe quantitativ untersucht, um zu prüfen, ob das Topic Modell annehmbare Ergebnisse erzeugt. Hierfür werden in Tabelle 6 acht Überschriften von Artikeln gelistet, bei den sich das Modell die höchste Zuversicht (engl. Confidence) aufweist, dass der jeweilige Artikel zu dem Thema Digitalisierung oder Digitaler Staat gehört.

Rang	Topic: Digialisierung	Topic: Digitaler Staat
1	*"How To Leverage Digital Transformation To Make Workplaces More Diverse And Inclusive"*	*"Municipalities apply for funding for digitization"*
2	*"Digital Transformation Roadmap For Laggards"*	*"State government Up to nationwide fast internet"*
3	*"Add Dexterity Without Disruption: Let Low-Code Empower IT And Drive Digital Transformation"*	*"Altmarkt Zweckverband wants to start broadband expansion soon"*
4	*"How to Reduce the Complexities of Change In Digital Transformation"*	*"Concept for digital nodes decided"*
5	*"Technology Decisions to Avoid Digital Transformation Exhaustion"*	*"Application phase for digitization price starts"*
6	*"How To Keep Pace With Digital Transformation And Avoid Becoming Obsolete"*	*"Ministry of Defense million euros for consultants"*
7	*"Culture, Not Tech, Is Key To Driving Digital Transformation"*	*"Brandenburg Nationwide fast internet up to"*
8	*"Must Have Exec Ed: Matching Business Models With Business Algorithms"*	*"Federal states hardly call up funds from the digital pact"*

Tabelle 6: Evaluation des Topic Modells

Durch Analyse der Überschriften der Zugeordneten Artikel wird deutlich, dass fast alle Artikel inhaltlich das zugewiesene Thema thematisieren. So sind dem Topic Digitalisierung also tatsächlich Artikel zugeordnet, bei den vermutet werden kann, dass sie die Digitalisierung thematisieren. Dasselbe gilt für das Topic Digitaler Staat. Aus diesem Grund kann davon ausgegangen werden, dass das Topic Modell gut funktioniert und die Artikel innerhalb eines Topics inhaltlich sehr ähnlich sind. Ferner kann auch die hohe vorgeschlagene Themenkohärenz in Abbildung 7 bestätigt werden, da die Interpretationen der Versuchsteilnehmer in vielen Fällen sehr ähnlich sind (siehe Tabelle 5). Aus der semantischen Nähe der Interpretationen kann schließlich auch abgeleitet werden, dass die richtigen Themen für die spätere Datenselektion gewählt wurden.

Aus diesen Erkenntnissen kann geschlussfolgert werden, dass der Korpus nach der inhaltlichen Selektion in hohem Maße nur noch Artikel beinhaltet, die zum Thema Digitalisierung geschrieben wurde. Somit ist auch die Datenqualität des Untersuchungsmaterial hoch. Zudem wird auch die Wirksamkeit des implementieren Topic Modells klar. Obwohl für diesen Ansatz auch menschliche Ressourcen benötigt werden, ist er drastisch effizienter im Vergleich zu einer manuellen

Themenfeststellung, da nur die gefundenen Topics anhand von wenigen Begriffen interpretiert werden müssen und nicht jeder Artikel im gesamten Korpus gelesen werden muss. So kann mit geringem Aufwand Wissen über den Inhalt aller Artikel des Korpus generiert werden.

Darüber hinaus wird nun auch die Qualität der Quantifizierung der Tonalität, also die Qualität Sentiment Analyse, näher betrachtet. Das Repräsentieren der Stimmungrichtung eines gesamten Artikels in einer Kennzahl ist bereits konzeptionell eine schwere Aufgabe. Vor allem, da der Korpus aus Zeitungsartikeln besteht, in dem häufig mehrere Sichtweisen zu einem Thema geschildert werden. Damit können vor allem auch mehrere Stimmungsrichtungen in einem Artikel vorkommen. Der Modalwert von Null des deutschen Sprachraums lässt sich womöglich neben der Zwiegespaltenheit auch durch diesen Effekt erklären.
Für die Sentiment Analyse konnte bereits mit Hilfe der vorab durchgeführten Stichprobe eine Evaluierung mit der Kennzahl MAE durchgeführt werden. Obwohl bereits die alternative mit dem geringsten MAE gewählt wurde, ist die Abweichung des automatisch zugewiesenen Sentiments zum manuell vergebenen Sentiment mit 36,2 Prozent immer noch sehr hoch. Eine mögliche Erklärung für die hohe Abweichung und den hohen MAE ist, dass das Wörterbuch der verwendeten Python Library nicht für den Wortschatz des Korpus ausgelegt wurde. VaderSentiment wurde ursprünglich für die Sentimentbestimmung von Sozial Medien Text Daten kreiert (vgl. Hutto & Gilbert, 2015). Der Korpus enthält allerdings vorwiegend Nachrichtendaten aus dem Technologie Bereich, sodass Sentiment Wörter gegebenenfalls nicht im Wörterbuch vorkommen oder mit falscher Bedeutung. In Hinsicht auf Tabelle 3 weißt vaderSentiment allerdings bei drei von vier Ansätzen den geringsten MAE auf. Dadurch kann davon ausgegangen werden, dass vaderSentiment im Zusammenhang mit dem Korpus dieser Arbeit besser geeignet ist als Textblob. Allerdings muss hierbei auch die Realitätsnähe des manuell vergegeben Sentimentwerts in Frage gestellt werden. Wie zuvor beschrieben besteht eine große Herausforderung dabei die Stimmungsrichtung eines ganzen Dokuments mittels einer Kennzahl auszudrücken. Somit fiel auch das manuelle Labeling solcher Artikel schwer, die mehrere verschiedene Sichtweisen und damit verschiedene Tonalitäten umfassen. Zudem ist die Wahrnehmung von Stimmungsrichtungen sehr subjektiv. Limitiert auf den

Umfang dieser Arbeit wurde das manuelle Labeling nur von einer Testperson durchgeführt, wodurch die Ausdruckskraft der Grundwahrheit geschwächt ist. Je mehr Versuchspersonen in einer weiterführenden Studie befragt würden, desto näher würde das tatsächliche Sentiment des Artikels erhoben werden.

Eine hohe Aussagekraft der Sentiment Analyse kann jedoch dadurch erahnt werden, da fünf von sechs der vorgestellten Ansätze ähnliche Ergebnisse liefern. Die Ergebnisse aller Ansätze sind in Abbildung 15 festgehalten. Bis auf Ansatz 2 mit der Library Textblob, zeigen sämtliche Ansätze ein hoch signifikantes Ergebnis, bei dem der Durchschnitt der Tonalität im detuschen Sprachraum negativer als der Durchschnitt des anlgo-amerikanischen Sprachraum ist.

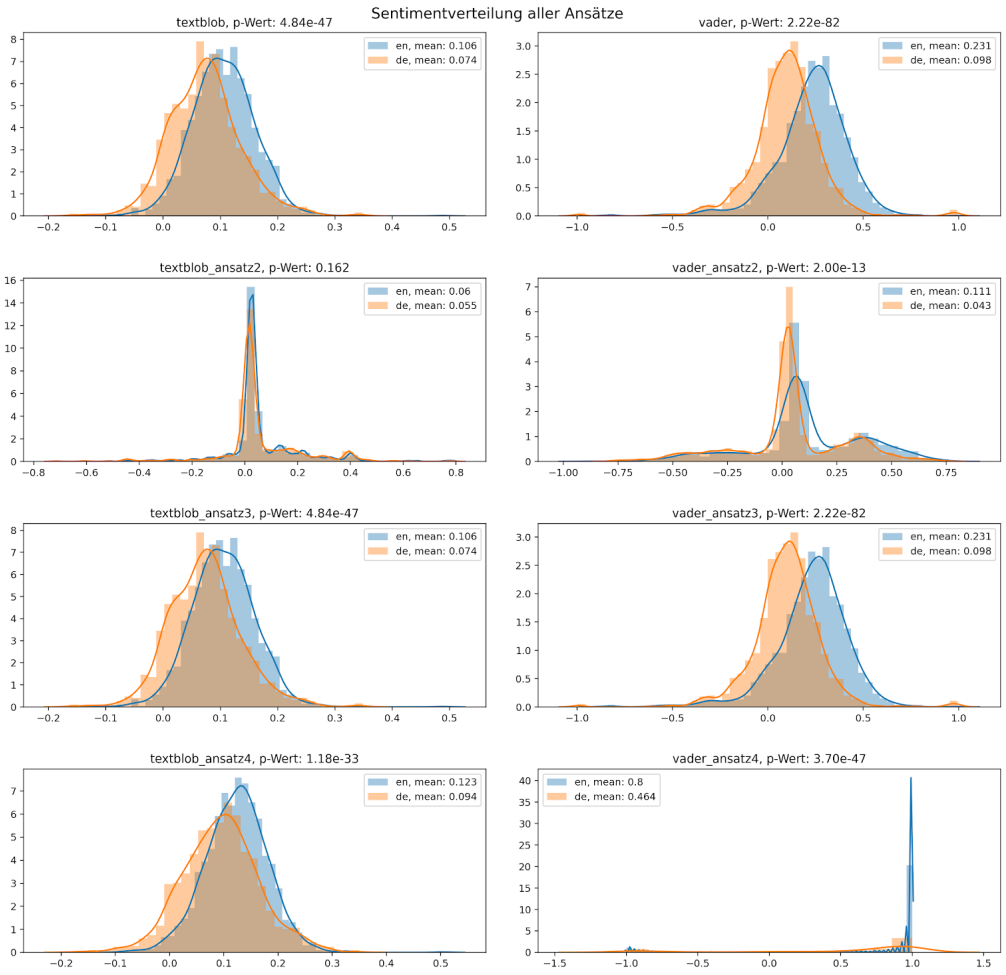


Abbildung 15: Ergebnisse aller Ansätze

Die Ergebnisse in dieser Arbeit unterliegen aber zusätzlich weiteren Einflüssen, die Auswirkungen auf die Aussagekraft haben. Eine international vergleichende Studie, wie diese Arbeit, wirft eine methodische Schwierigkeit auf, da ein Umgang der

verschiedenen Sprachen geschaffen werden muss. Aus Gründen der Vergleichbarkeit wurde wie in Kapitel 0 eine Standardisierung der Texte, also eine Übersetzung durchgeführt. Fraglich ist hierbei, ob die Übersetzungen so genau sind, dass sie die Inhalte und vor allem Stimmungsrichtungen der Texte nicht verzerren. Um den Effekt der Verzerrung zu eliminieren, könnte man in einer weiterführenden Studie die englischen Texte ins Deutsche übersetzen. Zwei Herausforderungen wurden bereits identifiziert: Einerseits ist die automatische Übersetzung mit der Google Translate API bei einem Korpus dieser Größe sehr kostspielig. Andererseits stehen für die deutsche Sprache nur eine öffentliche Python Library für die Sentiment Analyse zur Verfügung, sodass diese zwangsläufig ohne Aternativen verwendet werden muss.
Werden die Auswirkungen der Datenselektion näher betrachtet, fällt auf, dass der resultierende Korpus in Bezug auf die Quellendiversität sehr unausgeglichen ist (siehe Abbildung 9). Der untersuchte Korpus setzt sich hauptsächlich aus einer Quelle je Sprachraum zusammen. 96 Prozent der Artikel des deutschen Sprachraums stammen von der Quelle Süddeutsche Zeitung. Gleiches zeichnet sich im anglo-amerikanischen Sprachraum ab, in dem 81 Prozent der Artikel von der Quelle Forbes verfasst wurden. Die anderen Quellen haben daher kaum Einfluss auf das Untersuchungsergebnis. Aufgrund des ressourcenaufwändigem Webscraping wurden nachtraglich keine neuen Daten gesammelt. Jedoch sollte für eine höhere Aussagekraft in aufbauenden Arbeiten auf eine höhere Quellendiversität des bereinigten Korpus geachtet werden. Weitere sinnvolle Quellen wäre beispielweise das Soziale Netzwerk Twitter.

Schließlich kann in Abbildung 16 unter Berücksichtigung der Metadaten gezeigt werden, dass im deutschen Sprachraum nicht nur zu den Themen der Digitalisierung eine durchschnittliche negativere Tonalität im Vergleich zum anglo-amerikanischen Sprachraum vorliegt. In den meisten anderen Topic ist dieser Effekt auch zu sehen. Daher kommt das Ergebnis potentiell nicht nur durch eine schlechtere Einstellung gegenüber der Digitalisierung zustande, sondern womöglich auch, da die deutschen Medien generell negativer Berichterstatten. Diese Vermutung wirft damit weiteres Forschungspotential für zukünftige Studien auf.

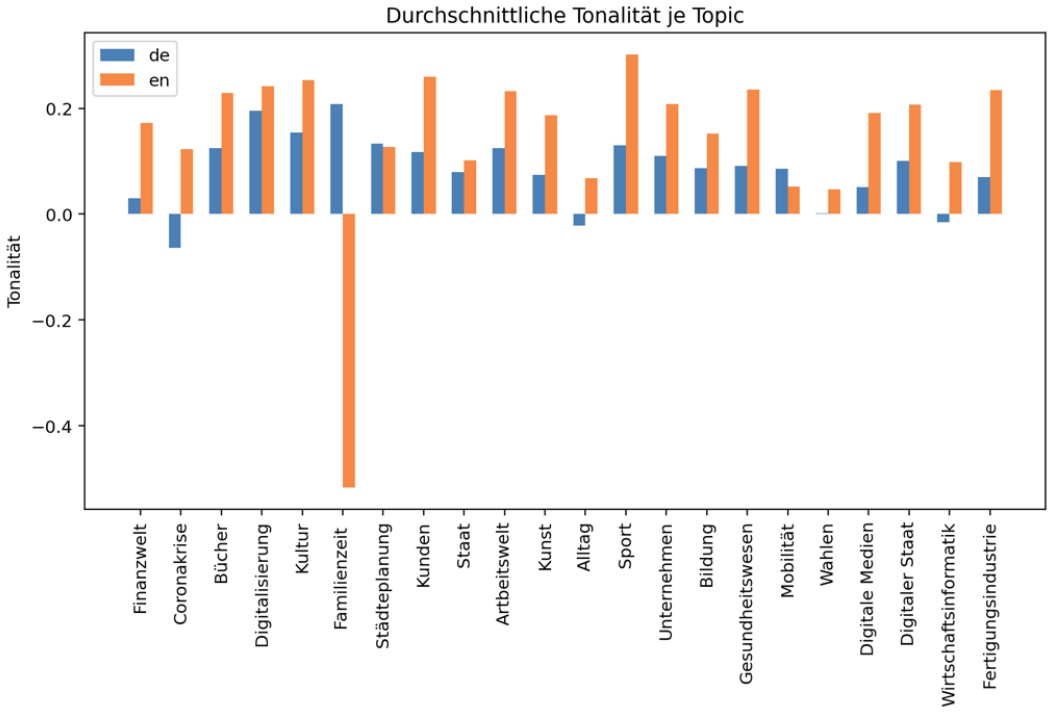


Abbildung 16: Durchschnittliches Sentiment je Topic

Nachfolgend wird diskutiert wie und mit welcher Aussagekraft die Forschungsfrage beantwortet werden kann. Aufgrund Übereinstimmung der verschiedenen Sentiment Analyse Ansätze, der hohen Datenqualität des Untersuchungsmaterials und den ähnlichen Befunden anderer Wissenschaftler, besteht zusammenfassend eine hohe Plausibilität der gefundenen Ergebnisse. Insbesondere, da die ähnlichen Ergebnisse der anderen Studien mittels verschiedener Methoden gefunden wurden. Daher besteht in Zusammenhang mit der Validität der Abbildung der öffentlichen Meinung durch Nachrichten- und Zeitschriftenartikel mit dieser Arbeit ein Hinweis darauf, dass im deutschen Sprachraum tatsächlich negativer über die Digitalisierung gesprochen wird als das im anglo-amerikanischen Sprachraum.

Dieser Befund deutet darauf hin, dass die Bevölkerung des deutschen Sprachraums nur erschwert die Veränderungen der Digitalisierung adaptiert, da sich möglicherweise kritischer mit den Risiken der Digitalisierung auseinandergesetzt wird.

6 Fazit

In dieser Arbeit wurde ein quantitativer Forschungsansatz entwickelt, der die Stimmungsrichtung der öffentlichen Meinung zum Thema Digitalisierung im deutschen mit dem anglo-amerikanischen Sprachraum vergleicht. Dafür wurde zuerst der bisherige Forschungsstand vorgestellt, wobei die Umfrage als konventionelle Methodik der Datenerhebung in der Meinungsforschung identifiziert wurde. Daraus abgeleitet besteht die Neuheit dieser Arbeit darin, die Daten mit Hilfe eines Public Opinion Mining Systems automatisiert und computergestützt zu erheben. Für die Realisierung des Ansatzes wurde zu Beginn die Datensammlung durch die Implementierung von Webscraper umgesetzt, anschließend die gesammelten Daten unter Einsatz eines LDA Topic Models bereinigt und schließlich die Stimmungsrichtung mit Hilfe der Sentiment Library vaderSentiment quantifiziert.

Mit diesem System konnte gezeigt werden, dass im deutschen Sprachraum negativer über die Digitalisierung gesprochen wird als im anglo-amerikanischen Sprachraum. Dabei wird die weitverbreitete Behauptung, im deutschen Sprachrum würde eine allgemeinen Technikskepsis vorliegen, nicht bestätigt. Allerdings besteht ein Hinweis darauf, dass im deutschen Sprachraum im Vergleich zum anglo-amerikanischen Sprachraum tendenziell negativer gesprochen wird, da zu vielen anderen Themen wie zum Beispiel Sport, Bildung oder Staat auch eine negativere Tonalität erhoben wurde. Dieses Indiz sollte in einer weiterführenden Studie näher untersucht werden, unter anderem auch, um die Aussagekraft dieser Arbeit zu bestätigen.

Nachdem die Digitalisierung Einfluss auf die gesamte globale Gesellschaft nimmt, besteht auch weiterhin die Dringlichkeit, die Stimmungsrichtung der öffentlichen Meinung zum Thema zu verfolgen. Das in dieser Arbeit entwickelte Public Opinion Mining System ist gut für diese Aufgabe geeignet, da, im Gegensatz zur Umfrage, nach der initialen Entwicklungsphase auf Knopfdruck neue, aktuelle Stimmungsbilder erzeugt werden können.
Ferner wird das sprachübergreifendes Public Opinion Mining zukünftig in der praktischen Anwendung sehr relevant werden, da für Staat und Industrie ein hohes Erkenntnisinteresse, in der Meinung der globalisierten Bevölkerung vorliegt.

Quellenverzeichnis

Aggarwal, Charu C. 2015. *Data Mining*. Cham: Springer International Publishing.

Al Omran, Fouad Nasser A., und Christoph Treude. 2017. „Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments“. S. 187–97 in *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*. Buenos Aires, Argentina: IEEE.

Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, und Krys Kochut. 2017. „A brief survey of text mining: Classification, clustering and extraction techniques“. *arXiv preprint arXiv:1707.02919*.

Araujo, Matheus, Julio Reis, Adriano Pereira, und Fabricio Benevenuto. 2016. „An evaluation of machine translation for multilingual sentence-level sentiment analysis“. S. 1140–1145 in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*.

Bendel, Oliver. 2018. „Definition: Digitalisierung“. *Digitalisierung*. Abgerufen 3. August 2020 (https://wirtschaftslexikon.gabler.de/definition/digitalisierung-54195/version-277247).

Blei, David. 2011. *Introduction to Probabilistic Topic Models*.

Blei, David M., Andrew Y. Ng, und Michael I. Jordan. 2003. „Latent dirichlet allocation“. *Journal of machine Learning research* 3(Jan):993–1022.

Bleich, Erik, und A. Maurits van der Veen. 2018. „Media Portrayals of Muslims: A Comparative Sentiment Analysis of American Newspapers, 1996–2015“. *Politics, Groups, and Identities* 1–20.

Bundesministerium für Wirtschaft und Energie. 2020. „Den digitalen Wandel gestalten“. Abgerufen 16. Juni 2020 (https://www.bmwi.de/Redaktion/DE/Dossier/digitalisierung.html).

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde, und Hagen Langer, Hrsg. 2010. *Computerlinguistik und Sprachtechnologie: eine Einführung*. 3., überarb. und erw. Aufl. Heidelberg: Spektrum, Akad. Verl.

Chaovalit, Pimwadee, und Lina Zhou. 2005. „Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches“. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*.

Denecke, Kerstin. 2008. „Using SentiWordNet for multilingual sentiment analysis“. S. 507–12 in.

Ding, Xiaowen, Bing Liu, und Philip S. Yu. 2008. „A holistic lexicon-based approach to opinion mining“. S. 231–240 in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, *WSDM ’08*. Palo Alto, California, USA: Association for Computing Machinery.

Fayyad, Usama, Gregory Piatetsky-Shapiro, und Padhraic Smyth. 1996. „From Data Mining to Knowledge Discovery in Databases“. *AI Magazine* 17(3):37–37.

Früh, Werner. 2017. *Inhaltsanalyse: Theorie und Praxis*. 9., überarbeitete Auflage. Konstanz: UVK Verlagsgesellschaft mbH.

Gadatsch, Andreas. 2017. „Einfluss der Digitalisierung auf die Zukunft der Arbeit“. S. 193–213 in *Controlling und Leadership: Konzepte – Erfahrungen – Entwicklungen*, herausgegeben von A. Gadatsch, A. Krupp, und A. Wiesehahn. Wiesbaden: Springer Fachmedien.

Gillen, Philippe, und Achim Wambach. 2018. „Ableitungen zum Einfluss der Digitalisierung auf die Volkswirtschaft“. S. 159–68 in *Digitalisierung im Einkauf*, herausgegeben von F. Schupp und H. Wöhner. Wiesbaden: Springer Fachmedien.

Gyorodi, Cornelia, Robert Gyorodi, George Pecherle, und Andrada Olah. 2015. „A Comparative Study: MongoDB vs. MySQL“. S. 1–6 in *2015 13th International Conference on Engineering of Modern Electric Systems (EMES)*. Oradea, Romania: IEEE.

Hamidian, Kiumars, und Christian Kraijo. 2013. „DigITalisierung – Status quo“. S. 1–23 in *Digitalisierung und Innovation*, herausgegeben von F. Keuper, K. Hamidian, E. Verwaayen, T. Kalinowski, und C. Kraijo. Wiesbaden: Springer Fachmedien Wiesbaden.

Hanni, Chill, und Meyn Hermann. 1996. „Funktionen der Massenmedien in der Demokratie“.

Hippner, Hajo, und René Rentzmann. 2006. „Text Mining“. *Informatik-Spektrum* 29(4):287–90.

Hu, Minqing, und Bing Liu. 2004. „Mining and summarizing customer reviews“. S. 168–177 in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, *KDD ’04*. Seattle, WA, USA: Association for Computing Machinery.

Hutto, C. J., und Eric Gilbert. 2015. „VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text“.

Khalifa, Osama, David Wolfe Corne, Mike Chantler, und Fraser Halley. 2013. „Multi-Objective Topic Modeling“. S. 51–65 in *Evolutionary Multi-Criterion Optimization*, *Lecture Notes in Computer Science*, herausgegeben von R. C. Purshouse, P. J. Fleming, C. M. Fonseca, S. Greco, und J. Shaw. Berlin, Heidelberg: Springer.

Kim, Dong Sung, und Jong Woo Kim. 2014. „Public Opinion Mining on Social Media: A Case Study of Twitter Opinion on Nuclear Power“. S. 224–28 in.

Kim, Soo-Min, und Eduard Hovy. 2004. „Determining the Sentiment of Opinions“. S. 1367–1373 in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: COLING.

Kim, Tae Kyun. 2015. „T test as a parametric statistic“. *Korean Journal of Anesthesiology* 68(6):540–46.

Kim, Yoosin, Seung Ryul Jeong, und Imran Ghani. 2014. „Text Opinion Mining to Analyze News for Stock Market Prediction“. 13.

Kröhling, Andreas. 2017. „Digitalisierung – Technik für eine nachhaltige Gesellschaft?“ S. 23–49 in *CSR und Digitalisierung*, *Management-Reihe Corporate Social Responsibility*, herausgegeben von A. Hildebrandt und W. Landhäußer. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kuckartz, Udo, Stefan Rädiker, Thomas Ebert, und Julia Schehl. 2010. „t-Test: zwei Mittelwerte vergleichen“. S. 147–66 in *Statistik*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kulkarni, Akshay, und Adarsha Shivananda. 2019. *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning Using Python*. Berkeley, CA: Apress.

Lau, Jey Han, David Newman, und Timothy Baldwin. 2014. „Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality“. S. 530–539 in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics.

Lau, Raymond Y. K., Chapmann C. L. Lai, Jian Ma, und Yuefeng Li. 2009. „Automatic Domain Ontology Extraction for Context-Sensitive Opinion Mining“. 19.

Lemke, Matthias, und Gregor Wiedemann. 2016. „Einleitung Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse“. S. 1–13 in *Text Mining in den Sozialwissenschaften*, herausgegeben von M. Lemke und G. Wiedemann. Wiesbaden: Springer Fachmedien Wiesbaden.

Lenz, Justus. 2019. „Digitalisierung – Hype oder Revolution?: Die Auswirkungen der digitalen Transformation auf Wirtschaft und Arbeitsmarkt“. S. 149–58 in *Arbeit einspunktnull*. Nomos Verlagsgesellschaft mbH & Co. KG.

Li, Xiu, und Liping Gao. 2013. „The Design and Implementation of an Internet Public Opinion Monitoring and Analyzing System“. S. 176–80 in *2013 International Conference on Service Sciences (ICSS)*.

Liu, Bing. 2011. *Web Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Liu, Bing. 2012. „Sentiment analysis and opinion mining“. *Synthesis lectures on human language technologies* 5(1):1–167.

Liu, Bing, und Lei Zhang. 2012. „A survey of opinion mining and sentiment analysis“. S. 415–463 in *Mining text data*. Springer.

Liu, Qiong, und Ying Wu. 2012. „Supervised Learning“. S. 3243–45 in *Encyclopedia of the Sciences of Learning*, herausgegeben von N. M. Seel. Boston, MA: Springer US.

Loria, Steven, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, und E. Dempsey. 2014. „Textblob: simplified text processing“. *Secondary TextBlob: simplified text processing* 3.

Manning, Christopher D., Prabhakar Raghavan, und Hinrich Schütze. 2008. *Introduction to information retrieval*. New York: Cambridge University Press.

Martin, Fiona, und Mark Johnson. 2015. „More Efficient Topic Modelling Through a Noun Only Approach“. S. 111–115 in *Proceedings of the Australasian Language Technology Association Workshop 2015*. Parramatta, Australia.

McGregor, Shannon C. 2019. „Social Media as Public Opinion: How Journalists Use Social Media to Represent Public Opinion“. *Journalism* 20(8):1070–86.

Meier, Andreas. 2018. *Werkzeuge der digitalen Wirtschaft: Big Data, NoSQL & Co.* Wiesbaden: Springer Fachmedien Wiesbaden.

Miner, Gary, Robert Nisbet, Andrew Fast, Thomas Hill, und Dursun Delen, Hrsg. 2012. *Practical text mining and statistical analysis for non-structured text data applications*. 1st ed. Waltham, MA: Academic Press.

Mohammad, Saif M., Mohammad Salameh, und Svetlana Kiritchenko. 2016. „How Translation Alters Sentiment". *Journal of Artificial Intelligence Research* 55:95–130.

Neuberger, Dr Christoph. 2004. „Wandel der aktuellen Öffentlichkeit im Internet". 26.

Pang, Bo, und Lillian Lee. 2008. „Opinion Mining and Sentiment Analysis". 94.

Petereit, Dieter. 2020. „Deepl mit Update: Google-Translate-Konkurrenz wird schlauer". *t3n Magazin*. Abgerufen 24. Juni 2020 (https://t3n.de/news/deepl-update-schlauer-1250207/).

Petz, Gerald. 2019. *Opinion Mining im Web 2.0: Ansätze, Methoden, Vorgehensmodell*. Wiesbaden: Springer Fachmedien Wiesbaden.

Raithel, Jürgen. 2006. *Quantitative Forschung: ein Praxiskurs*. 1. Aufl. Wiesbaden: VS, Verl. für Sozialwiss.

Rajman, M., und R. Besançon. 1998. „Text Mining: Natural Language Techniques and Text Mining Applications". S. 50–64 in *Data Mining and Reverse Engineering: Searching for semantics. IFIP TC2 WG2.6 IFIP Seventh Conference on Database Semantics (DS-7) 7–10 October 1997, Leysin, Switzerland*, *IFIP — The International Federation for Information Processing*, herausgegeben von S. Spaccapietra und F. Maryanski. Boston, MA: Springer US.

Rajman, Martin, und Martin Vesely. 2004. „From Text to Knowledge: Document Processing and Visualization: A Text Mining Approach". S. 7–24 in *Text Mining and its Applications*, *Studies in Fuzziness and Soft Computing*, herausgegeben von S. Sirmakessis. Berlin, Heidelberg: Springer.

Rasch, Björn, Hrsg. 2008. *Quantitative Methoden: Einführung in die Statistik. Bd. 1: [...]: mit 25 Tabellen*. 2., erw. Aufl., korrigierter Nachdr. Heidelberg: Springer.

Runkler, Thomas A. 2015. *Data Mining*. Wiesbaden: Springer Fachmedien Wiesbaden.

Sbalchiero, Stefano, und Maciej Eder. 2020. „Topic Modeling, Long Texts and the Best Number of Topics. Some Problems and Solutions“. *Quality & Quantity* 54(4):1095–1108.

Scholz, Thomas. 2011. „Ein Ansatz zu Opinion Mining und Themenverfolgung für eine Medienresonanzanalyse.“ S. 7–12 in.

Scholz, Thomas, Stefan Conrad, und Isabel Wolters. 2012. „Comparing Different Methods for Opinion Mining in Newspaper Articles“. S. 259–64 in *Natural Language Processing and Information Systems*, *Lecture Notes in Computer Science*, herausgegeben von G. Bouma, A. Ittoo, E. Métais, und H. Wortmann. Berlin, Heidelberg: Springer.

Schweiger, Wolfgang. 2017. „Öffentliche Meinung und Meinungsbildung online“. S. 113–53 in *Der (des)informierte Bürger im Netz*. Wiesbaden: Springer Fachmedien Wiesbaden.

Shang, Songtao, Minyong Shi, Wenqian Shang, und Zhiguo Hong. 2015. „Research on public opinion based on Big Data“. S. 559–62 in *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*.

Shapiro, Adam Hale, Moritz Sudhof, und Daniel Wilson. 2020. „Measuring news sentiment“. Federal Reserve Bank of San Francisco.

Shirsat, Vishal S., Rajkumar S. Jagdale, und S. N. Deshmukh. 2017. „Document Level Sentiment Analysis from News Articles“. S. 1–4 in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*.

Stoffel, Kilian. 2009. „Web + Data Mining = Web Mining“. *HMD Praxis der Wirtschaftsinformatik* 46(4):6–20.

Störk-Biber, Constanze, Jürgen Hampel, Cordula Kropp, und Michael Zwick. 2020. „Wahrnehmung von Technik und Digitalisierung in Deutschland und Europa: Befunde aus dem TechnikRadar“. *HMD Praxis der Wirtschaftsinformatik* 57(1):21–32.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, und Manfred Stede. 2011. „Lexicon-Based Methods for Sentiment Analysis“. *Computational Linguistics* 37:267–307.

Tausendpfund, Markus. 2018. „Operationalisierung“. S. 107–37 in *Quantitative Methoden in der Politikwissenschaft: Eine Einführung*, *Grundwissen Politik*, herausgegeben von M. Tausendpfund. Wiesbaden: Springer Fachmedien.

Turney, Peter. 2002. „Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews“. S. 417–424 in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Vodafone Institut für Gesellschaft und Kommunikation. 2019. „The Tech Divide Die Unterschiedliche Wahrnehmung Der Digitalisierung In Europa, Asien Und Den Usa (Politik)“. Abgerufen 27. Februar 2020 (https://www.vodafone-institut.de/wp-content/uploads/2019/02/Politik_Tech_Divide_VFI.pdf).

Vodafone Institute for Society and Communications. 2018. „The Tech Divide Die Unterschiedliche Wahrnehmung Der Digitalisierung In Europa, Asien Und Den Usa (Industrie und Arbeit)“. Abgerufen 25. Juli 2020 (https://www.vodafone-institut.de/wp-content/uploads/2018/11/VFI_Industrie_und_Arbeit_Tech_Divide.pdf).

Walther, Ralf. 2001. „Web Mining“. *Informatik-Spektrum* 24(1):16–18.

Wasserstein, Ronald L., und Nicole A. Lazar. 2016. „The ASA Statement on *p* -Values: Context, Process, and Purpose“. *The American Statistician* 70(2):129–33.

Wei-ping, Zhu, Li Ming-xin, und Chen Huan. 2011. „Using MongoDB to implement textbook management system instead of MySQL“. S. 303–5 in *2011 IEEE 3rd International Conference on Communication Software and Networks*.

Weischenberg, Siegfried. 2001. „Feature-Aufbau“. S. 225–49 in *Nachrichten-Journalismus*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Wiedemann, Gregor, und Andreas Niekler. 2016. „Analyse qualitativer Daten mit dem ‚Leipzig Corpus Miner‘“. S. 63–88 in *Text Mining in den Sozialwissenschaften*, herausgegeben von M. Lemke und G. Wiedemann. Wiesbaden: Springer Fachmedien Wiesbaden.

Yang, Heng-Li, und Hung-Chang Huang. 2019. „Sentiment Classification for Web Search Results“. *Journal of Internet Technology* 20(7):2043–2053.

Zeller, Frauke, Jens Wolling, und Pablo Porten-Cheé. 2010. „Framing 0/1. Wie die Medien über die ‚Digitalisierung der Gesellschaft ‘berichten“. *M&K Medien & Kommunikationswissenschaft* 58(4):503–524.