ELISABETH GÜNTHER

TOPIC MODELING

Algorithmische Themenkonzepte in Gegenstand und Methodik der Kommunikationswissenschaft

Ausgezeichnet mit dem Herbert von Halem Nachwuchspreis

HERBERT VON HALEM VERLAG

HH

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Elisabeth Günther
Topic Modeling.
Algorithmische Themenkonzepte in Gegenstand und Methodik der Kommunikationswissenschaft
Köln: Halem 2022

D6

ELISABETH GÜNTHER studierte Kommunikationswissenschaft in Augsburg (B.A.) und Hohenheim (M.Sc.). Von 2012-17 war sie wissenschaftliche Mitarbeiterin an den Universitäten Hohenheim und Münster. Ihre Forschungsinteressen liegen im Bereich Journalismusforschung und Computational Methods. Seit 2017 arbeitet sie als Data Scientistin.

Alle Rechte, insbesondere das Recht der Vervielfältigung und Verbreitung sowie der Übersetzung, vorbehalten. Kein Teil des Werkes darf in irgendeiner Form (durch Fotokopie, Mikrofilm oder ein anderes Verfahren) ohne schriftliche Genehmigung des Verlages reproduziert oder unter Verwendung elektronischer Systeme (inkl. Online-Netzwerken) gespeichert, verarbeitet, vervielfältigt oder verbreitet werden.

© 2022 by Herbert von Halem Verlag, Köln

ISBN (Print): 978-3-86962-575-1
ISBN (PDF): 978-3-86962-569-0

Den Herbert von Halem Verlag erreichen Sie auch im Internet unter http://www.halem-verlag.de
E-Mail: info@halem-verlag.de

SATZ: Herbert von Halem Verlag
LEKTORAT: Rabea Wolf, Volker Manz
DRUCK: docupoint GmbH, Magdeburg
GESTALTUNG: Claudia Ott Grafischer Entwurf, Düsseldorf
Copyright Lexicon ©1992 by The Enschedé Font Foundry.
Lexicon® is a Registered Trademark of The Enschedé Font Foundry.

Elisabeth Günther

Topic Modeling

Algorithmische Themenkonzepte in Gegenstand und Methodik der Kommunikationswissenschaft

HERBERT VON HALEM VERLAG

Inhalt

Abkürzungsverzeichnis 15

TEIL I
RELEVANZ UND FRAGESTELLUNG

1. Die duale Bedeutung von Topic Modeling für die KW 20
   1.1 Forschungsleitendes Interesse 20
      1.1.1 *Algorithmische Themen im Gegenstand der KW* 22
      1.1.2 *Algorithmische Themen in der Methodik der KW* 23
      1.1.3 *Methodologische Problemstellung: Vergleich der manuell-deduktiven und automatisch-induktiven Themenanalyse* 24
   1.2 Zum Aufbau der Arbeit 26
   1.3 Danksagungen 29

TEIL II
ALGORITHMEN UND CODE

Einleitung: Zur Bedeutung und Begründung der algorithmischen Logik 32

2. Technologischer Wandel: Die Vierte Industrielle Revolution 33
   2.1 Big Data 35
      2.1.1 *Zwischen Gesellschaftswandel, Datenbankproblem und Pathos* 36
      2.1.2 *Digitalisierung und Datafizierung* 39
      2.1.3 *Algorithmische Logik als Treiber der Automatisierung* 43

2.2 Künstliche Intelligenz 47
2.2.1 *Wesen und Ziel der KI* 49
2.2.2 *Geschichte der KI* 56
2.2.3 *KI-Trends und Zukunftsprognosen* 60
2.3 Implikationen für Alltag und Gesellschaft 65
2.3.1 *Techno-Social Environment* 66
2.3.2 *Wie verändern Digitalisierung und Big Data die Choice Architecture?* 68
2.3.3 *Reverse-Turing-Test: Dehumanisierung im Techno-Social Environment* 72
2.3.4 *Ethik für Algorithmen* 78
2.4 Implikationen für die empirische Forschung 87
2.4.1 *Big Data und Algorithmen in der empirischen Forschung* 88
2.4.2 *Data Science und Computational Social Sciences* 94
2.4.3 *Kritik an Big Data und Co* 100
2.5 Zwischenfazit: Algorithmen und Code in der Makroperspektive 104
2.5.1 *Big Data und KI in Gegenstand und Methodik der KW* 104
2.5.2 *Zentrale Folgerungen für die Computational Social Sciences* 106

3. Algorithmische Verarbeitung natürlicher Sprache 111
3.1 Sprache als Zeichensystem (de Saussure) 114
3.1.1 *Zentrale Relevanz der Sprache* 114
3.1.2 *Dyadischer Zeichenbegriff und Weiterführung* 116
3.2 Zeichenlehre als Erkenntnistheorie (Peirce) 118
3.2.1 *Zeichenbegriff und Grundlagen* 118
3.2.2 *Syntaktik, Semantik und Pragmatik: Weiterführung nach Morris* 123
3.2.3 *Übergang in die Stochastik: Deduktion, Induktion und Abduktion* 124
3.3 Information und Kommunikation (Shannon/Weaver) 127
3.3.1 *Von der Zeichenlehre zur Informationstheorie* 129
3.3.2 *Definition von Information als stochastischer Prozess* 133
3.4 Algorithmische Verarbeitung von Sprache und Information 137
3.4.1 *Natural Language Processing* 138
3.4.2 *Machine Learning* 142

3.5 Kognition und sprachliche Repräsentation 146
3.5.1 *Einführung in die Kognitionswissenschaft* 148
3.5.2 *Computational Theory of Mind* 151
3.6 Zwischenfazit: Algorithmen und Code in der Mikroperspektive 155
3.6.1 *Computationale Verarbeitung natürlicher Sprache* 155
3.6.2 *Mensch vs. Maschine?* 157

4. Fazit: Die Omnipräsenz der algorithmischen Logik 160
4.1 Veränderte Bedingungen in Gegenstand und Methodik der Kommunikationswissenschaft 160
4.2 Leitmotive 162
4.2.1 *Motiv 1: Eine strikte Dichotomie aus ›Mensch vs. Maschine‹ greift fehl* 163
4.2.2 *Motiv 2: KI schafft keine ›Magic Buttons‹ – vom Ob zum Wie* 164
4.2.3 *Motiv 3: Die KW hat die Kompetenz und Verantwortung, im Diskurs um algorithmische Systeme eine zentrale Rolle einzunehmen* 166

TEIL III
DAS KONSTRUKT ›THEMA‹

Einleitung:
Zwei Perspektiven auf das Konstrukt ›Thema‹ 169

5. Das Thema als Einheit der öffentlichen Meinung 171
5.1 Was ist ›das Thema‹? 172
5.1.1 *Themenbegriff, -konstrukt, -konzept und -variable* 172
5.1.2 *Das Konstrukt ›Thema‹ in der KW-Forschungstradition* 173
5.2 Themen konstituieren Öffentlichkeit 175
5.2.1 *Drei Ebenen von Thematisierung* 175
5.2.2 *Integrationsfunktion von Themen* 177
5.3 Themenauswahl konstruiert Medienrealität 179
5.3.1 *Gatekeeper* 180
5.3.2 *News-Bias* 181
5.3.3 *Nachrichtenwert* 182

5.4 Einfluss und Beeinflussung massenmedialer Thematisierung 184
5.4.1 *Praeludium: Der Issue-Begriff in der politischen Kommunikation* 185
5.4.2 *Wirkungshypothese im Agenda-Setting-Ansatz* 187
5.4.3 *Issues als Gegenstand der strategischen Kommunikation* 189
5.4.4 *Neue Aspekte algorithmisch strukturierter Öffentlichkeit* 192
5.5 Thematische Ausdifferenzierung führt zu Fragmentierung? 195
5.5.1 *Ursprünge der Fragmentierungsthese* 196
5.5.2 *Stratifizierte und segmentierte Öffentlichkeit* 197
5.5.3 *Themen in der Filterblase* 199
5.5.4 *Kritik an der Fragmentierungsthese* 201
5.6 Zwischenfazit: Das Konstrukt ›Thema‹ in der Makroperspektive 205
5.6.1 *Die kommunikationswissenschaftliche Forschungstradition* 205
5.6.2 *Thematisierung in algorithmisch strukturierter Öffentlichkeit* 207

6. Das Thema als Wissensstruktur 210
6.1 Was ist ›das Thema‹? 213
6.2 Textverstehen bottom-up: Die Textoberfläche 216
6.2.1 *Propositionsmodelle* 216
6.2.2 *Makrostrukturmodelle: Einzug des Themenbegriffs* 218
6.3 Textverstehen top-down: Schematheoretische Ansätze 221
6.3.1 *Schema (Bartlett)* 221
6.3.2 *Skripte (Schank/Abelson)* 223
6.3.3 *Frames (Minsky)* 224
6.3.4 *Konstruktions-Integrations-Modell (Kintsch)* 225
6.3.5 *Das Themenkonstrukt in den schematheoretischen Ansätzen* 226
6.3.6 *Interdisziplinäre Anleihen* 229
6.4 Text im Kontext: Textverstehen anhand mentaler Modelle 232
6.4.1 *Mentale Modelle (Johnson-Laird)* 233
6.4.2 *Situationsmodelle (van Dijk/Kintsch)* 235
6.4.3 *Das Themenkonstrukt in mentalen Modellen* 237

6.5 Zwischenfazit: Das Konstrukt ›Thema‹ in der Mikroperspektive 238
6.5.1 *Die Entwicklung des Forschungsbereichs Textverstehen* 238
6.5.2 *Themendefinition im Bereich des Textverstehens* 240
6.5.3 *Verbindung zum Forschungsbereich KI* 241

7. Fazit: Es gibt kein universales Verständnis von ›Thema‹ 243
7.1 Unterschiedliche Ausgangspositionen für KW und KI 243
7.2 Zentrale Erkenntnisse zum Themenkonstrukt 244
7.2.1 *Themendefinitionen, Themenkonzepte* 244
7.2.2 *Ineinandergreifen der beiden Perspektiven* 246

TEIL IV
ALGORITHMISCHE THEMEN

Einleitung: Manuelle und algorithmische Themenanalysen 249

8. Das Thema als Variable in der kommunikationswissenschaftlichen Forschungstradition 254
8.1 Das klassische Vorgehen 256
8.2 Vom Themenkonzept zur Themenvariable 257
8.2.1 *Kategorienbildung* 257
8.2.2 *Ereignisbezug von Themen* 259
8.3 Codierphase als gelenkte Rezeption 261
8.4 Computationale Verfahren im klassischen Forschungsprozess 264
8.5 Zwischenfazit: Die klassische KW-Forschungspraxis 265

9. Algorithmische Themenanalyse 267
9.1 Computationale Verarbeitung natürlichsprachiger Texte 269
9.2 Latent Semantic Analysis (LSA) 270
9.2.1 *LSA als ›Theory of Meaning‹* 272
9.2.2 *LSA als Textanalyseverfahren* 273
9.2.3 *Mathematische Grundlagen* 274

9.3 Probabilistic Latent Semantic Analysis (PLSA) 281
9.3.1 *Generatives Modell* 282
9.3.2 *Schätzen der Modellparameter* 284
9.3.3 *Zentrale Errungenschaften im Topic Modeling* 287
9.4 Latent Dirichlet Allocation (LDA) 288
9.4.1 *LDA = PLSA + Bayes* 290
9.4.2 *Generatives Modell* 291
9.4.3 *Vorbereitung der Datenanalyse* 293
9.4.4 *Schätzen der Modellparameter* 294
9.5 Zwischenfazit: Die computationale Forschungspraxis 298

10. Fazit: Ganzheitliche Gegenüberstellung der Themenanalysen 301
10.1 Mensch vs. Maschine 302
10.2 Zentrale Unterschiede in der Untersuchungsanlage 305
10.3 Konsequenzen für den Forschungsprozess 308
10.4 Konsequenzen für die Qualität der Messung 310
10.4.1 *Reliabilität als triadische Beziehung* 311
10.4.2 *Apparat* 313
10.4.3 *Kontext* 315
10.4.4 *Material* 317
10.5 Komplementäre Instrumente zur Themenanalyse 319

TEIL V
FAZIT UND DISKUSSION

11. Topic Modeling ist fester Bestandteil der KW 321
11.1 Der Blick zurück 321
11.2 Integration der algorithmischen Themenanalyse in die KW 323
11.2.1 *These 1: Innovation und Allgegenwärtigkeit* 323
11.2.2 *These 2: Duale Relevanz* 324
11.2.3 *These 3: TM = manuelle Themenanalyse + Automatisierung* 324
11.2.4 *These 4: Umfassende Integration* 326
11.2.5 *These 5: Menschliche Intelligenz* 327

11.2.6 *These 6: Konzeptionelle Skills* 327
11.2.7 *These 7: Kompetenz und Verantwortung* 329
11.3 Kritische Reflexion 330
11.4 Der Blick nach vorn 331
11.4.1 *Sapir-Whorf-Hypothese: Zum Einfluss der (künstlichen) Sprache* 332
11.4.2 *Computational Turn, Computational Divides?* 334
11.4.3 *Das menschliche Themenverständnis im Techno-Social Environment* 336

Literaturverzeichnis 339

Abbildungsverzeichnis

Abb. 1	Deutungsraum der Arbeit zur Einordnung des Themenkonstrukts in die kommunikationswissenschaftliche Forschung	21
Abb. 2	Gliederung der Arbeit	27
Abb. 3	World's technological capacity to store information 1986 - 2014	40
Abb. 4	Ebenen der Automatisierung realweltlicher Probleme	45
Abb. 5	Definitionen Künstlicher Intelligenz als 4-Felder-Matrix	50
Abb. 6	Menschliche und Künstliche Intelligenz	51
Abb. 7	Steigendes Interesse an computationalen Analysen in Fachzeitschriftenartikeln der Ebsco-CMMC-Datenbank	90
Abb. 8	Steigendes Interesse an computationalen Analysen in Präsentationen und Postern der ICA-Jahrestagungen	90
Abb. 9	Data-Science-Venn-Diagramm	95
Abb. 10	Data-Science-Puzzle	97
Abb. 11	Dyadisches Zeichen nach de Saussure	117
Abb. 12	Triadische Zeichenkonzeption nach Peirce	120
Abb. 13	Schematische Darstellung des informationstheoretischen Kommunikationsmodells	132
Abb. 14	Sprach- und informationstheoretische Grundlagen für die algorithmische Themenanalyse via TM	156
Abb. 15	Themenbegriff, -konstrukt, -konzept und -variable	173
Abb. 16	Themen als Teil der Makrostruktur von Texten	220
Abb. 17	Themen als Bindeglied zwischen Text und Schemata	227
Abb. 18	Textverstehen anhand mentaler Modelle	234
Abb. 19	›Das Thema‹ im Forschungsbereich Textverstehen	239
Abb. 20	Konzeption und Operationalisierung des Themenbegriffs für die manuelle Inhaltsanalyse und Topic Modeling	251

Abb. 21 Schematische Darstellung der SVD-Reduktion 277
Abb. 22 VSM- und LSA-Repräsentationen 280
Abb. 23 PLSA Aspect Model in der Plate Notation 283
Abb. 24 Schematische Darstellung der PLSA-Dimensionsreduktion 286
Abb. 25 Generatives Modell für LDA in der Plate Notation 291
Abb. 26 Reliabilität der Themenanalyse als triadische Beziehung 312
Abb. 27 Menschliche und algorithmische Themenkonstruktion im Techno-Social Environment 337

Tabellenverzeichnis

Tab. 1	Arbeitsdefinitionen für ›Thema‹ und ›Issue‹ in der KW	174
Tab. 2	Arbeitsdefinitionen für ›Thema‹ und ›Theme‹ in der Psycholinguistik	241
Tab. 3	Arbeitsdefinitionen für die vier Themenbegriffe	245
Tab. 4	Zentrale Unterschiede zwischen LSA und PLSA	287
Tab. 5	Zentrale Unterschiede in der Untersuchungsanlage der manuellen Themenanalyse und des Topic Modelings	306
Tab. 6	Zentrale Unterschiede im Forschungsprozess der quantitativen Themenanalyse der KW und des Topic Modelings	309
Tab. 7	Reliabilität der manuellen Themenanalyse und des Topic Modelings: Wie robust sind die Verfahren gegen Variationen bzgl. Apparat, Kontext und Material?	314
Tab. 8	Themenkonzept und Umsetzung in der manuellen Themenanalyse und im Topic Modeling	325

Abkürzungsverzeichnis

AAAI	Association for the Advancement of Artificial Intelligence
AGI	Artificial General Intelligence
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
ASI	Artificial Superintelligence
BI	Business Intelligence
BoW	Bag of Words
CAPTCHA	Completely Automated Public Turing-Test to tell Computers and Humans Apart
CCS	Computational Communication Science
CM	Computational Methods
CMMC	Communication and Mass Media Complete
CSS	Computational Social Sciences
CTM	Correlated Topic Model
CToM	Computational Theory of Mind
DDoS	Distributed Denial of Service
DTM	Dokument-Term-Matrix
EM	Expectation-Maximization
FAT/ML	Fairness, Accountability, and Transparency in Machine Learning
GLM	General Linear Model
ICA	International Communication Association
IKT	Informations- und Kommunikationstechnologie
IoT	Internet of Things
IR	Information Retrieval
IT	Information Technology
KI	Künstliche Intelligenz
KL	Kullback-Leibler-Divergenz
KW	Kommunikationswissenschaft

LCA	Latente Klassenanalyse, engl. Latent Class Analysis
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MCMC	Markov-Chain-Monte-Carlo
MINT	Mathematik, Informatik, Naturwissenschaft und Technik
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MOOC	Massive Open Online Course
NASA	National Aeronautics and Space Administration
NLP	Natural Language Processing
PLSA	Probabilistic Latent Semantic Analysis
PR	Public Relations
SEO	Search Engine Optimization
STM	Structural Topic Model
SVD	Singulärwertzerlegung, engl. Singular Value Decomposition
TDM	Term-Dokument-Matrix
TM	Topic Modeling
VSM	Vector Space Model
WBS	wissensbasierte Systeme

TEIL I

RELEVANZ UND FRAGESTELLUNG

1. DIE DUALE BEDEUTUNG VON TOPIC MODELING FÜR DIE KW

1.1 Forschungsleitendes Interesse

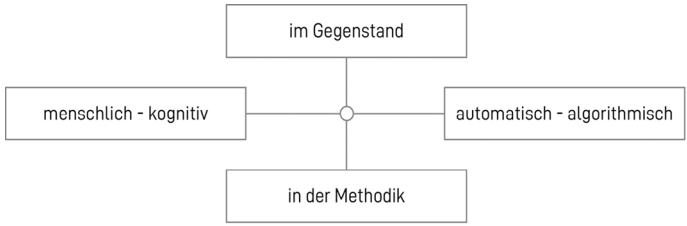
Als zentrales Konstrukt der Kommunikationswissenschaft (KW) ist *das Thema* heute relevanter denn je. Für unser Fach besitzt es durch seine grundlegende Funktion als Ordnungseinheit der öffentlichen Meinung eine besondere Bedeutung. Über Themen wird es möglich, sich aufeinander zu beziehen, zu wissen, *worum es geht*. Themen sind kollektive Bezugsgrößen, über die eine Diskussion ermöglicht wird, Öffentlichkeit strukturiert wird und auf denen Demokratie aufbaut. Themen konstituieren Öffentlichkeit. Sie sind so prägend, dass die Handlungskapazität einer Demokratie durch ihre Themenkapazität ausgewiesen werden kann (LUHMANN 1970).

Seinen aktuellen Bedeutungszuwachs über die sozialwissenschaftliche Forschung hinaus verdankt das Konstrukt dem Technologiewandel. Um exponentiell gewachsene Datenmengen, sog. *Big Data*, sinnhaft ordnen zu können, müssen algorithmische Systeme ›lernen‹, was Themen sind. Welche Webseiten passen am besten zu einer Suchanfrage, welche Nachrichtenartikel möchte eine Nutzerin wahrscheinlich als Nächstes lesen, und welche Wahlwerbung ist geeignet, um genau diesen Facebook-User von meiner Kandidatin zu überzeugen? Dreh- und Angelpunkt dieser und ähnlicher Aufgabenstellungen ist die Identifikation thematisch ähnlicher Inhalte, welche automatisch ausgewählt und ausgespielt werden. Methodisch umgesetzt wird die algorithmische Modellierung von Themen im State of the Art des Machine Learnings über Topic-Modeling-Algorithmen. Bereits heute sind Nutzer*innen dadurch in der täglichen digitalen Kom-

munikation (nicht immer bewusst) mit dem Ergebnis einer algorithmischen Modellierung von Themen in Kontakt.

ABBILDUNG 1

Deutungsraum der Arbeit zur Einordnung des Themenkonstrukts in die kommunikationswissenschaftliche Forschung



Quelle: eigene Darstellung

Topic Modeling (TM) ist in der KW bereits als innovatives Datenerhebungsverfahren etabliert (z. B. GÜNTHER/QUANDT 2016; MAIER et al. 2018). Die vorliegende Arbeit zielt in einer methodologischen Auseinandersetzung über den praktischen Einsatz hinaus und verfolgt gleichermaßen die theoretische Einordnung der algorithmischen Themenanalyse als neue Logik der Strukturierung öffentlicher Kommunikation. Um der zweifachen Relevanz für unsere Disziplin gerecht zu werden, wird *das Thema* erstens im Gegenstand der KW theoretisch begründet und zweitens im Sinne ihrer Fachmethodik verortet. Den Rahmen für die Auseinandersetzung mit diesen beiden Aspekten setzt der Technologiewandel: Dem traditionell eher intuitiv begriffenen Themenkonstrukt wurde durch Fortschritte im Bereich Künstliche Intelligenz (KI) ein algorithmisches Pendant zur Seite gestellt. Dabei handelt es sich jedoch nicht etwa um eine *Automatisierung* der klassischen kommunikationswissenschaftlichen Themenanalyse, sondern vielmehr um eine andere Messart einer anders operationalisierten Themenvariable. In ihrem Deutungsraum umspannt die vorliegende Arbeit diese beiden miteinander verbundenen Achsen (vgl. Abb. 1): *Das Thema* wird sowohl im Gegenstand als auch in der Methodik der KW verortet und als menschlich-kognitives wie auch als automatisch-algorithmisches Konstrukt theoretisch eingeordnet – *bzw. genau mit Fokus auf den Veränderungen, die sich durch diese Ergänzung ergeben*.

1.1.1 *Algorithmische Themen im Gegenstand der* KW

Forschungsinteresse 1: *In der digitalen Kommunikation ergänzen automatisch-algorithmisch generierte Themen die bisher sozial-kollektiv ausgehandelte Struktur von Öffentlichkeit. Was bedeutet dies für den Gegenstand der kommunikationswissenschaftlichen Forschung?*

Die transformativen Prozesse des Technologiewandels bringen mit den enormen Fortschritten im Bereich der KI eine veränderte Medienlandschaft mit sich. Dies impliziert potenzielle Auswirkungen auf die relevanten Themen unserer Gesellschaft, welche in Anbetracht eines zunehmend komplexen und diversen Medienerlebens längst nicht mehr reduziert werden können auf *das, was in der Tagesschau kommt*. Über die algorithmische Modellierung thematischer Kontexte können digitale Plattformen ihren Nutzer*innen im Idealfall genau diejenigen Inhalte anzeigen, die sie individuell interessieren. Die KW diskutiert unter dem Stichwort der *Filterblasen* (PARISER 2011; LIAO/FU 2013; NGUYEN et al. 2014) und *Echoräume* (VICARIO et al. 2016; GARRETT 2009; ZOLLO et al. 2017) mögliche Risiken einer solchen Personalisierung: Scheitern die klassischen Massenmedien (durch Spezialisierung und mangelnde Reichweite) an der Vermittlung eines gemeinsamen Sets allgemein relevanter Themen, so leidet ihre Integrationsfunktion für die Gesellschaft. Dadurch, so die These, drohe im Extremfall eine Fragmentierung von Öffentlichkeit. Aktuelle empirische Belege geben zwar vorerst Entwarnung,[1] die Diskussion um die Rolle digitaler Technologien für die Themenbreite im gesellschaftlichen Diskurs muss dennoch geführt werden und betont die Relevanz der Themenvariable in der aktuellen kommunikationswissenschaftlichen Forschung.

Bei der Auseinandersetzung mit diesen Entwicklungen drängt sich an dieser Stelle die Beobachtung auf, dass digitale Technologie längst nicht mehr *auf der anderen Seite* steht: In den 1950er-Jahren platzierte der berühmte Turing-Test, ein Gedankenexperiment zur Feststellung des Erfolgs Künstlicher Intelligenz, Mensch und Maschine noch in voneinander getrennte Räume. Die Künstliche Intelligenz habe ihr Ziel erreicht, so die Vorgabe, wenn Fragesteller*innen die beiden anhand ihrer Antworten nicht

1 Etwa Bruns (2019) oder der Digital News Report des Reuters Instituts von Newman, Fletcher, Kalogeropoulos, Levy und Kleis Nielsen (2017).

mehr auseinanderhalten könnten: wenn das technische System also, hier bezogen auf das sprachliche Verhalten, eine imaginäre Grenze zwischen Mensch und Maschine überschritten habe (TURING 1950; vgl. Kap. 3.5). Diese Grenze, und mit ihr die simple Dichotomie von Mensch und Maschine – natürlich vs. künstlich, intuitiv vs. regelgeleitet, kreativ vs. deterministisch, denkend vs. rechnend –, ist in ihrer ursprünglich gedachten Form nicht mehr aufrechtzuerhalten. Durch den Technologiewandel sind digitale Technologien mittlerweile so eng mit unserem Alltag und unserem sozialen Leben verwoben, dass eine Aussage darüber, was tatsächlich »echt« ist, zunehmend schwerfällt. Im Turing-Test führte ein Computer sprachliches Verhalten über vorgegebene Regelsysteme aus, durch unvorhergesehene Fragen konnten gute Spielleiter*innen ihn also recht einfach austricksen. Heute sind digitale Anwendungen komplexer und nicht immer als solche zu erkennen. Für die vorliegende Arbeit bedeutet dies, dass wir uns im Forschungsobjekt und im (digitalen) Alltag auch abgesehen von der Diskussion um die Filterblasen bereits regelmäßig und nicht immer bewusst mit algorithmischen Themen konfrontiert sehen.

1.1.2 *Algorithmische Themen in der Methodik der KW*

Forschungsinteresse 2: *In der Informatik bzw. KI wurden im Zuge des Technologiewandels neue Methoden zur algorithmischen Modellierung der Themenstruktur von großen Textarchiven entwickelt. Welche Möglichkeiten bieten diese für die kommunikationswissenschaftliche Forschung?*

Mit dem TM steht der KW ein innovatives Verfahren zur algorithmischen Themenanalyse zur Verfügung. Sie beruht auf einer probabilistischen Modellierung der syntaktischen Struktur in großen Textsammlungen und hat ihre Wurzeln im Forschungsbereich KI. Als vollautomatisches Verfahren verspricht TM nicht nur, große Textmengen ohne menschlichen Codieraufwand zu klassifizieren, sondern die enthaltenen Themen zudem selbstständig zu inferieren. Forscher*innen müssen also weder ein Codebuch erstellen noch Coder*innen schulen und anleiten.

In Anbetracht der durch Digitalisierung und Datafizierung stetig zunehmenden Datenmenge, auf welche die KW zur Beantwortung ihrer Forschungsfragen zurückgreifen kann, sind Verfahren dieser Art vielversprechend: Manuelle Untersuchungen sind teuer und zeitaufwendig, sodass

sie mit Blick auf die Größe des Untersuchungsmaterials forschungspraktischen Grenzen unterliegen. In genau diesem Aspekt punkten algorithmische Verfahren mit der Möglichkeit einer effizienten Verarbeitung digitaler Textarchive einer Größenordnung, die die Kapazitäten manueller Codierleistung sprengen. Im Nachwort des *Big Data*-Special Issues des *Journal of Communication* betonte Parks (2014) die Relevanz der Auseinandersetzung mit den (voll-)automatischen Verfahren. Zwar seien die in der KW erzielten Erfolge in diesem Bereich noch klein, für die Entwicklung des Fachs sei die Auseinandersetzung aber von entscheidender Bedeutung: »Relatively little of the work done in this early stage will stand the test of time, but all of it will likely be critical in the ongoing process of conceptual and methodological advance« (S. 355). Entsprechend hoch war das Interesse an der 2017 erstmals auf der ICA-Jahrestagung vertretenen Interest Group *Computational Methods* (CM), welche der Verbreitung neuer Methoden gewidmet ist (VAN ATTEVELDT 2017): Gleich das erste Panel, mit dem sich die neu gegründete Interessengruppe innerhalb der Fachgesellschaft präsentierte, war als *Applications of Topic Modeling* der algorithmischen Themenanalyse gewidmet. Diese Arbeit beschreibt die theoretische Grundlage dieses aktuell hochrelevanten Verfahrens und ordnet die Methode konzeptionell in das empirische Repertoire der KW ein.

1.1.3 *Methodologische Problemstellung: Vergleich der manuell-deduktiven und automatisch-induktiven Themenanalyse*

Aus den beiden ersten Forschungsinteressen dieser Arbeit ergibt sich eine methodologische Problemstellung: Wie können manuelle und algorithmische Verfahren, die sich auf mehreren Ebenen grundlegend unterscheiden, sinnvoll einander gegenübergestellt werden? Das Grundproblem bei der Auseinandersetzung mit Methoden der algorithmischen Themenanalyse lässt sich recht einfach so zusammenfassen, dass einerseits die Menge der zu analysierenden Textdokumente stetig wächst, natürliche Sprache andererseits aber komplex ist. Wenngleich sich Verfahren der automatischen Inhaltsanalyse in der KW also mittlerweile fest etabliert haben (z. B. VAN ATTEVELDT 2008; SCHARKOW 2012; WETTSTEIN 2015; GÜNTHER/QUANDT 2016), scheint insbesondere gegenüber algorithmischen Analyseverfahren eine gewisse Skepsis bestehen zu bleiben. Verständlicherweise: Coder*innen können

Texte nicht nur inhaltlich verstehen, sondern über ihr Weltwissen auf einer höheren Abstraktionsebene einordnen. Dies versetzt sie in der Lage, den tatsächlichen Rezeptionsprozess in strukturierter Form nachzuvollziehen. Die KI kann zwar riesige Datenmengen verarbeiten, ihre komplexen mathematischen Modelle basieren jedoch im Grunde auf simplen Worthäufigkeiten. Inwiefern können Computer also inhaltliche Kategorien analysieren, wenn sie Textinhalte nur auszählen und nicht *verstehen* können?

Um die Relevanz und Anwendbarkeit von TM beurteilen zu können, ist ein grundlegendes Verständnis über dessen Funktionsweise und damit auch die Feststellung von Gemeinsamkeiten und Unterschieden zur manuellen Inhaltsanalyse zentral. In dieser Arbeit soll es jedoch nicht um den direkten Abgleich von TM-Ergebnissen mit denen einer klassischen KW-Themenanalyse gehen, wie er an anderer Stelle bereits vorgestellt wurde (VAN ATTEVELDT et al. 2014). Ein solcher hilft bei der Beantwortung einer wichtigen Ausgangsfrage – *Entsprechen die Ergebnisse denen einer klassischen Inhaltsanalyse?* –, markiert damit aber allenfalls den Beginn unseres Verständnisses der vielversprechenden Modelle.

Forschungsinteresse 3: *Wie können manuelle und algorithmische Themenanalyse als Ganzes gegenübergestellt werden, um der KW eine Grundlage zu bieten, Unterschiede in den Ergebnissen sachkundig unter Berücksichtigung der jeweiligen Stärken und Schwächen einordnen zu können?*

Werden in der vorliegenden Arbeit die Unterschiede zwischen algorithmischer und klassischer Themenanalyse fokussiert, soll das jedoch nicht von ihren starken gemeinsamen Wurzeln und den zahlreichen gemeinsamen Bezügen in ihrer Entwicklung ablenken. Ein Vergleich entlang der Genese des zentralen Grundanliegens, nämlich der Rolle von Themen im menschlichen (Text-)Verstehen, ist daher nicht nur für das Verständnis von TM äußerst gewinnbringend, sondern beleuchtet in der Gegenüberstellung auch die klassische Themenanalyse neu. Das zunehmende Interesse der KW an den neuen Verfahren eröffnet also auch einen neuen Zugang zu zentralen Ausgangsfragen im Fach, wie hier zur Rolle von Themen. Damit motivieren die algorithmischen Konzepte sowohl im Gegenstand wie auch im methodischen Zugang eine Rückbesinnung auf die Grundlagen, welche in der KW wie auch in der KI im Bereich der Semiotik, angewandt auf die Inhaltsanalyse im Bereich des Textverstehens, verortet sind. Verfolgt man diese Gemeinsamkeiten, so wird sichtbar, dass die algorithmische Inhalts-

analyse nicht vollständig von extern in die KW eingeführt werden muss. Zentrale Konzepte sind bereits in der klassischen Themenanalyse angelegt. KW und KI haben diese Grundlagen im Sinne ihrer jeweiligen Ausrichtung unterschiedlich und unabhängig voneinander umgesetzt – im Rahmen von Data Science und der Computational Social Sciences (CSS) ragen sie nun in höchst spannender Weise wieder ineinander.

1.2 Zum Aufbau der Arbeit

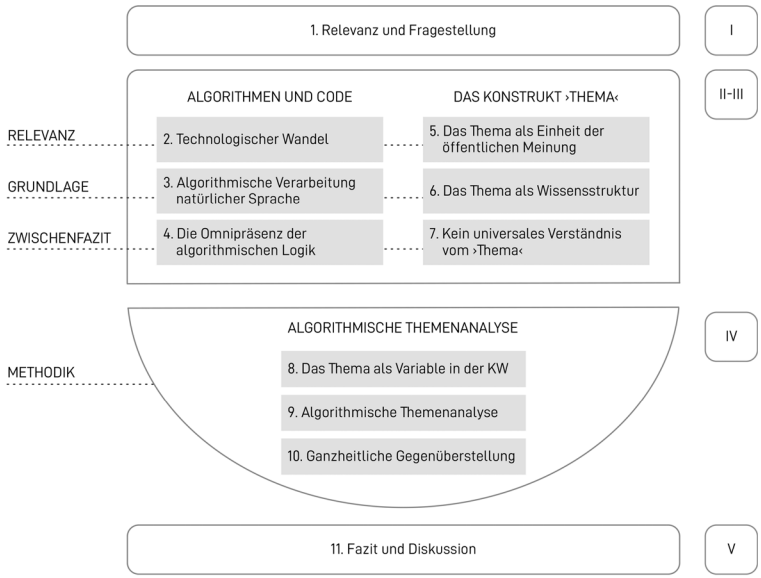
Da es sich bei der algorithmischen Themenanalyse via TM um eine Methodeninnovation für die KW handelt, muss nicht nur das Verfahren selbst, sondern auch der Kontext algorithmischer Verfahren insgesamt eingeführt werden. Die Arbeit legt also zum einen eine Grundlage für die CSS als Teilbereich der Data Science und beschreibt zum anderen darauf aufbauend algorithmische Themenkonstrukte als spezielles Anwendungsgebiet. Aus der grundlegend interdisziplinären Natur der CSS ergeben sich nun nicht nur ungewöhnlich viele Bezüge, sondern aus kommunikationswissenschaftlicher Sicht auch Bezüge zu ungewöhnlichen Fachrichtungen. Folglich spannt diese Arbeit einen sehr weiten Bogen und verbindet Grundlagen aus verschiedensten Disziplinen – unter ihnen die KI, Semiotik, Informationstheorie und Psycholinguistik.

Leser*innen wird dadurch einige Geduld abverlangt: Bevor algorithmische Themenkonstrukte beschrieben und zum in der KW etablierten Themenverständnis in Bezug gesetzt werden können, müssen zunächst einige terminologische und konzeptuelle Grundlagen gelegt werden. Diese sind für eine kompetente Auseinandersetzung mit *dem Thema* an sich und mit der algorithmischen Themenanalyse via TM notwendig. Die Bezüge sind in Einleitung sowie Zwischenfazit zu jedem Kapitel skizziert, die volle Bedeutung ihres Zusammenspiels kann aber erst im späteren Verlauf, basierend auf dem so gelegten theoretischen Fundament, herausgearbeitet werden. Im Folgenden soll ein grober Überblick die Struktur der vorliegenden Arbeit veranschaulichen (vgl. Abb. 2). Auf diese Einleitung (Teil I) folgen mit den Theorieblöcken zur *algorithmischen Logik* und zum *Konstrukt ›Thema‹* (Teile II und III) zwei zentrale Fundamente. Sie bereiten den Boden für die daran anschließende Beschreibung der *algorithmischen Themenanalyse* via TM und deren Gegenüberstellung mit der manuellen quantitativen Themenanalyse der KW (Teil IV). Fazit und Diskussion schließen die Arbeit ab (Teil V).

Kapitel 1: Das aktuelle Kapitel leitet in die Arbeit ein, begründet ihre Relevanz und stellt die Forschungsinteressen vor.

Kapitel 2 leitet das erste theoretische Standbein der Arbeit ein, welches sich auf verschiedenen Ebenen mit der Bedeutung von Algorithmen und Code für Gesellschaft und Forschung auseinandersetzt (Teil II). Kapitel 2 begründet die allgemeine Relevanz dieser Arbeit auf makroperspektivischer Ebene und beschreibt die weitreichenden Auswirkungen des technologischen Wandels, welcher Konzepte aus dem Forschungsbereich der KI in den (digitalen) Alltag rückt. Welche Implikationen ergeben sich daraus für Gesellschaft und empirische Sozialforschung?

ABBILDUNG 2
Gliederung der Arbeit



Quelle: eigene Darstellung

Kapitel 3 begründet die quantitative Verarbeitung natürlichsprachiger Texte, wie sie den zuvor beschriebenen Umwälzungen zugrunde liegt: Als kleinster gemeinsamer Nenner von KW und KI steht das Zeichen als Grundeinheit von Sprache, Wissen und Kognition. Das Kapitel führt in Semiotik,

Informationstheorie und Kognitionswissenschaft ein und schließt mit einem Überblick über Machine Learning (ML) und Natural Language Processing (NLP).

Kapitel 4 bietet ein Zwischenfazit zu Teil II. Dazu werden der makro- (Kapitel 2) und der mikroperspektivische Blick (Kapitel 3) auf Algorithmen und Automatisierung miteinander verbunden. Die wichtigsten Kernaussagen werden in Form von drei Leitmotiven für die folgende Auseinandersetzung mit Konzepten und Anwendungen aus dem Bereich der KI wiederholt.

Kapitel 5 lenkt den Fokus der Arbeit auf das zweite theoretische Standbein, das Konstrukt ›Thema‹ (Teil III), welches hier auf makroperspektivischer Ebene als Ordnungseinheit der öffentlichen Meinung verortet wird. Da die KW über keine einheitliche Definition für dieses zentrale Konstrukt verfügt, wird seine Bedeutung über verschiedene Forschungsbereiche des Fachs hin verfolgt.

Kapitel 6 beschreibt Themen als Wissensstrukturen, die eine zentrale Rolle bei der menschlichen Informationsverarbeitung einnehmen. Gemäß dem Forschungsinteresse dieser Arbeit fußt die definitorische Grundlage auf Erkenntnissen der Psycholinguistik zum Textverstehen. Hier entwickelte Konzepte und Theoriemodelle bilden die Grundlage für die algorithmische Themenanalyse via TM.

Kapitel 7 beinhaltet ein Zwischenfazit zu Teil III. Das Kapitel greift die vorgestellten Facetten des Themenbegriffs als Einheit der öffentlichen Meinung (Kapitel 5) und als Wissensstruktur (Kapitel 6) auf und diskutiert Konsequenzen der digitalen Transformation für die kommunikationswissenschaftliche Forschungspraxis.

Kapitel 8 leitet Teil IV ein, welcher die beiden theoretischen Standbeine der Arbeit (Teile II und III) zu einer methodologischen Auseinandersetzung mit der algorithmischen Themenanalyse via TM verbindet. Als Grundlage für eine Gegenüberstellung stellt Kapitel 8 die klassische KW-Themenanalyse vor und diskutiert Vorteile und Problemstellen. In Bezug auf die Themenkonzeption nimmt diese üblicherweise eine makroperspektivische Annäherung vor (Kapitel 5). Für das Verständnis der manuellen Codierung sind

die mikroperspektivischen Grundlagen zum Textverstehen aber ebenso relevant (Kapitel 6).

Kapitel 9 führt in die algorithmische Themenanalyse via TM ein, welche ganz konkret an die psycholinguistischen Theoriemodelle anschließt (Kapitel 6), methodologisch begründet über die beschriebenen sprach- und informationstheoretischen Grundlagen (Kapitel 3), umgesetzt in computationaler Verarbeitung großer Textmengen (Kapitel 2). Das Kapitel zeichnet die Entwicklung anhand der drei kanonischen Modell-Gruppen LSA, PLSA und LDA nach.

Kapitel 10 stellt die TM-Verfahren der klassischen quantitativen Themenanalyse in einem ganzheitlichen Vergleich gegenüber: Dazu wird der Mensch-Maschine-Vergleich neu aufgegriffen, zentrale Unterschiede werden in der Untersuchungsanlage aufgeschlüsselt und Konsequenzen für den Forschungsprozess und die Qualität der Messung diskutiert.

Kapitel 11 rundet die Arbeit ab, indem es die zentralen Schlüsse in Bezug auf die theoretische Einordnung algorithmischer Themenkonstrukte in Gegenstand und Methodik der KW in Form von sieben Thesen formuliert. Abschließend folgen eine kritische Reflexion der Arbeit und eine Diskussion von Zukunftsszenarien, die mögliche Entwicklungen der KW und ihres Gegenstandsbereichs in Zeiten des technologischen Wandels betreffen.

1.3 Danksagungen

Ich danke an erster Stelle Thorsten Quandt, der mich überhaupt erst dazu motiviert hat, eine Promotionsstelle anzutreten, und mich fachlich wie menschlich immer kompetent und großzügig unterstützt hat. Emese Domahidi, die von Anfang an mein Verständnis von Wissenschaft und Kooperation beeindruckt hat. Meinem Bruder Christian, der mir mit »Warum nicht?« die Scheu vor Code und Programmierung nahm. Michael Scharkow, der mich bei meinen ersten Schritten in R und der automatischen Inhaltsanalyse begleitete. Florian Buhl für all unsere produktiven Dispute und für den Impuls, mich theoretisch mit dem Themenkonstrukt auseinanderzusetzen. Anna Volpers für mein Praktikum und die vielen Gespräche über Semiotik, Zeichen und Bedeutung. Jens Vogelgesang, Helle Sjøvaag,

Rodrigo Zamith und Christian Baden für den Support. Ben Fretwurst für die zahllosen Diskussionen zu Reliabilität und öffentlicher Thematisierung, und für alles andere. Raik Roth für die wundervolle Unterstützung und Freundschaft. Grazy, Julia, Monett und Nancy: Ich bin dankbar, dass ich euch an meiner Seite hatte. Raphael: für alles.

TEIL II

ALGORITHMEN UND CODE

EINLEITUNG: ZUR BEDEUTUNG UND BEGRÜNDUNG DER ALGORITHMISCHEN LOGIK

Die Auseinandersetzung mit algorithmischen Themenkonzepten in der KW bedarf der Festlegung zentraler Ausgangspunkte. Teil II fokussiert folglich zunächst auf die transformativen Prozesse des technologischen Wandels, der als gesellschaftlicher Rahmen sowohl das Fundament als auch die Folgen computationaler Verfahren und algorithmischer Konstrukte abbildet (Kapitel 2). Anschließend folgt eine Einführung in die sprach- und informationstheoretischen Grundlagen, die dem beschriebenen Wandel als kleinste Einheit zugrunde liegen und für die Begründung und das Verständnis computationaler Verfahren und algorithmischer Konstrukte zentral sind (Kapitel 3). Weshalb ist es notwendig, einen so großen Bogen zu beschreiben, wo diese Arbeit doch auf ein sehr spezielles Verfahren der algorithmischen Inhaltsanalyse fokussiert? Gerade weil es sich um ein sehr spezielles Verfahren handelt, ist es notwendig, dieses nicht isoliert zu betrachten – es handelt sich beim Topic Modeling nicht um einen *Magic Button*, der sich ohne Weiteres in eine fachfremde Disziplin übertragen lässt. Um die Verfahren adäquat einordnen, zu einer sinnvollen Anwendung bringen und nicht zuletzt ihre Relevanz in der alltäglichen digitalen Kommunikation begreifen zu können, ist daher ein Grundverständnis ihres Entstehungskontextes essenziell, mitsamt den Grundannahmen und Zielsetzungen, welche diesen Prozess über Jahrzehnte hinweg geprägt haben. Teil II bildet das erste Standbein für diese Zielsetzung und ist allgemeiner auf die Bedingungen und Grundlagen computationaler Verfahren und algorithmischer Konstrukte gerichtet. Teil III wird anschließend speziell auf das Themenkonstrukt in KW und KI fokussieren.

2. Technologischer Wandel: Die vierte industrielle Revolution

Den Ausgangspunkt für dieses Kapitel, und damit für die gesamte Arbeit, bildet der technologische Wandel, welcher Alltag, Arbeit und Kommunikation über die vergangenen Jahrzehnte so tiefgreifend verändert hat, dass er vom Weltwirtschaftsforum als *Vierte Industrielle Revolution* bezeichnet wird (Schwab 2016). Entsprechend bezeichnet Andrew Ng[2] die Künstliche Intelligenz (KI) als die »neue Elektrizität«:

> »AI is the new electricity. Just as electricity transformed many industries roughly 100 years ago, AI will also now change nearly every major industry [. . .]. The industrial revolution freed humanity from much repetitive physical drudgery; I now want AI to free humanity from repetitive mental drudgery, such as driving in traffic« (NG 2017, Par. 6).

Damit steht der gegenwärtig zu beobachtende technologische Wandel in einer eindrucksvollen Reihe revolutionärer Neuerungen, welche das Arbeits- und Zusammenleben über die vergangenen Jahrhunderte immer wieder grundlegend veränderten: Die erste industrielle Revolution ermöglichte die maschinelle Produktion durch Wasserkraft und Dampfmaschine, die zweite beschrieb die dank Elektrizität heute selbstverständliche Massenproduktion, die dritte letztlich eine Automatisierung der Produktion durch Elektronik und Informationstechnologie. Aktuelle Entwicklungen wie Internet of Things (IoT) oder autonomes Fahren gehen jedoch erneut weit über das Konzept der digitalen Revolution hinaus: Alltag und Arbeit werden durch Informations- und Kommunikationstechnologie (IKT) nicht

2 Andrew Ng ist Mitgründer des Google-Brain-Projekts sowie der MOOC-Plattform Coursera und eine der aktuell wohl einflussreichsten Persönlichkeiten der KI.

nur ergänzt und unterstützt, sondern sind grundlegend von digitalen Technologien durchdrungen (FRISCHMANN 2016; FRISCHMANN/SELINGER 2018) und *basieren* in weiten Teilen auf Code (LESSIG 2001). Der Gründer des Weltwirtschaftsforums Klaus Schwab fasst dies wie folgt zusammen:

> »There are three reasons why today's transformations represent not merely a prolongation of the Third Industrial Revolution but rather the arrival of a Fourth and distinct one: velocity, scope, and system impact. The speed of current breakthroughs has no historical precedent. When compared with previous industrial revolutions, the Fourth is evolving at an exponential rather than linear pace. Moreover, it is disrupting almost every industry in every country. And the breadth and depth of these changes herald the transformation of entire systems of production, management, and governance« (SCHWAB 2016, Par. 1).

Die aktuellen Umwälzungen unterscheiden sich außerdem insofern von den vorangegangenen, als Technologie und technologischer Fortschritt immer weniger als solche erkennbar sind, sondern nahezu unsichtbar mit dem Alltag verwoben scheinen – galt die Einführung der Personal Computer ganz offensichtlich als Revolution und löste eine klar sichtbare Umgestaltung von Büroplätzen und privaten Arbeitszimmern aus, ließ das Smartphone dieselbe Rechenleistung schon wenige Jahre später in die Hosentasche wandern. Schon kurz darauf verschwanden große Festplatten in der Cloud, während die moderne KI, versteckt in Kühlschränken und Kreditkarten, längst selbstverständlicher Teil des Alltags ist. Noch näher – buchstäblich unter die Haut – gehen uns die in Kombination mit der Bioinformatik entwickelten bionischen Prothesen, implantierte Mikrochips oder eSkin-Applikationen (vgl. Kap. 2.3.1).

Im Zentrum der digitalen Revolution steht die nun computervermittelte Kommunikation – das zentrale Forschungsobjekt unserer Disziplin wurde digitalisiert, und mit ihm unsere Forschungsmethoden: Basierend auf probabilistischen Modellen und der Quantifikation menschlichen Verhaltens als digitale Spuren greifen Algorithmen (je nach Perspektive der Betrachter*innen) unterstützend, lenkend oder gar manipulierend in die digitale Kommunikation ein. Die Kommunikationswissenschaft sucht folglich eine Realität zu beschreiben, in der die Struktur von Öffentlichkeit durch Algorithmen mitgestaltet wird. Gleichzeitig bringt dieser Wandel neue Techniken der Datenverarbeitung und innovative statistische Verfahren hervor, welche für die empirische Sozialforschung unter den sich rasant ändernden Bedingungen hochrelevant sind und zunehmend in das Methodenrepertoire aufgenommen werden. Sichtbar wird diese Entwick-

lung letztlich in einem wachsenden Begriffswald an neuen Konzepten: Big Data, Digitalisierung, Datafizierung, Algorithmen, Künstliche Intelligenz, Machine Learning, Hadoop, Data Mining, Data Science und Computational Social Science sind insbesondere im kommunikationswissenschaftlichen Kontext zwar nicht mehr ganz neu, sorgen aber nach wie vor für Verwirrung. Wie groß ist beispielsweise Big Data, und was unterscheidet Machine Learning von der klassischen Statistik?

Dieses Kapitel hat nun zwei Ziele: Zum einen soll es terminologische wie konzeptuelle Unklarheiten beseitigen (Kapitel 2.1 und 2.2) und so die Auseinandersetzung mit einem Teilbereich der beschriebenen Transformation, dem algorithmischen Thema, vorbereiten. Das zweite Ziel dieses Kapitels sieht daran anknüpfend vor, ein grundlegendes Verständnis dafür zu erarbeiten, weshalb die darauf folgende Auseinandersetzung mit dem (algorithmischen) Thema heute relevanter ist denn je. Dazu muss die Arbeit zunächst einen Schritt zurückgehen, um für Gesellschaft und Kommunikationsforschung relevante Aspekte der digitalen Revolution als großes Ganzes zu skizzieren – mit Fokus auf den Implikationen für das Forschungsobjekt (Kapitel 2.3) und den empirischen Zugang (Kapitel 2.4) der Kommunikationswissenschaft.

2.1 Big Data[3]

Der technologische Wandel katapultierte über die vergangenen Jahre zahlreiche neue Begriffe und Konzepte ins Zentrum unseres Fachs: Die Kommunikationswissenschaft – ehemals als Buchwissenschaft bezeichnet – beschäftigt sich heute unter anderem mit Big Data (Kapitel 2.1.1), Digitalisierung und Datafizierung (Kapitel 2.1.2) und mit Algorithmen (Kapitel 2.1.3). Diese Bezüge zu für die KW ungewöhnlichen Fachrichtungen ergeben sich aus der grundlegend interdisziplinären Natur der CSS und sollen im Folgenden sprachlich und konzeptionell ausgearbeitet werden.

3 Eine ausführliche Recherche zum etymologischen Hintergrund des Big-Data-Begriffs findet sich bei Diebold (2012).

2.1.1 *Zwischen Gesellschaftswandel, Datenbankproblem und Pathos*

Im Diskurs um den technologischen Wandel war *Big Data* in den vergangenen Jahren ein nahezu unumgängliches Schlagwort. Unter den genannten Konzepten spannt es den weitesten, allerdings auch den am wenigsten belastbaren Bezugsrahmen auf: Während die Bedeutung nämlich zunächst klar erscheint – es geht um große Datenmengen –, gibt es keine einheitliche Vorstellung darüber, was genau damit gemeint ist. Eine Ursache dafür liegt darin, dass der Begriff durch eine Kombination aus dem hohen Tempo des Technologiewandels und dem gleichzeitig starken öffentlichen Interesse sehr früh in den populärwissenschaftlichen Diskurs eingeführt wurde, maßgeblich bevor eine gründliche wissenschaftliche Auseinandersetzung erfolgen konnte (GANDOMI/HAIDER 2015).

Grundsätzlich stellt Big Data ganz allgemein das positive Pendant zum Begriff der *Informationsflut* dar (bzw. *information overload*), welcher den Diskurs noch bis vor wenigen Jahren dominierte: Große Datenmengen sind dort mit Chaos und Verwirrung assoziiert, gelten hier dagegen als Schatz, wertvolle Ressource und Währung moderner Unternehmensmodelle. Die positive Konnotation beruht in großen Teilen auf enormen Fortschritten im Bereich der Informatik und insbesondere der KI, welche in Bezug auf Infrastruktur und neue Möglichkeiten der Analyse von großen Datenmengen erzielt wurden (vgl. Kap. 2.2). Diese skalieren nicht nur die bisherigen Anwendungen in dem Sinne, dass sie anstatt Mega- und Gigabytes nun auch Terra- und Petabytes verarbeiten können – sie beschreiben vielmehr ein analytisches Mindset, welches basierend auf ganz neuen Tools und empirischen Zugängen eine alternative Perspektive auf Wissen bietet. Damit verweist der Begriff Big Data auf verschiedene, miteinander verwobene Aspekte einer Entwicklung, die ihre Wurzeln in der Statistik hat und die ihre Relevanz aus der digitalen Revolution bezieht: Bei ausreichend großer Datengrundlage, so die Prämisse, treten irrelevante Informationen oder durch unsaubere Daten generiertes Rauschen in den Hintergrund, sodass dahinter liegende Muster erkennbar werden. Diese alles andere als neue Logik ist heute prominent im ML verkörpert, welches im Rahmen der Data Science im sozialwissenschaftlichen Bereich Einzug hält (vgl. Kap. 2.4). Big Data liegt nun als eine Art Sammelbegriff über dieser Entwicklung:

> »[B]ig data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights, or create new forms of value, in ways that change mar-

kets, organizations, the relationship between citizens and government, and more« (MAYER-SCHÖNBERGER/CUKIER 2013: 6).

Der Pathos, mit dem Autor*innen wie Mayer-Schönberger und Cukier (2013) Big Data als revolutionäre Errungenschaft preisen, ist insbesondere für die ökonomisch geprägten Definitionen typisch und damit auf Meta-Ebene ebenfalls Teil des Phänomens. Boyd und Crawford (2012) schließen neben Technologiewandel und Analyse-Innovation auch den (mit Begriffen wie ›belief‹ und ›aura‹ religiös anmutenden) Hype um genau diese Entwicklungen ein: Big Data wird von ihnen definiert als

»a cultural, technological, and scholarly phenomenon that rests on the interplay of:

1. *Technology*: maximising computational power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
2. *Analysis*: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
3. *Methodology*: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy« (BOYD/CRAWFORD 2012: 663).

Mayer-Schönberger und Cukier (2013) sowie Boyd und Crawford (2012) bieten zwei anschauliche Beispiele dafür, dass sich die zahlreich vorhandenen Definitionen nicht nur inhaltlich, sondern auch in ihrer (impliziten) Bewertung stark voneinander unterscheiden. Dies liegt unter anderem daran, dass die Autor*innen in ihrer Beschreibung davon, *was Big Data eigentlich ist*, aus grundlegend unterschiedlichen Perspektiven auf das Phänomen Bezug nehmen: Je nach fachlicher Ausrichtung wird die Entwicklung mit Betonung neuer Möglichkeiten betriebswirtschaftlicher Wertschöpfung, der besonderen Charakteristika der Datengrundlage, der aus der Datenmenge entstehenden Herausforderungen für die IT-Infrastruktur oder sogar als eine Art Synonym zur Data Science verstanden. Grob können an dieser Stelle vier verschiedene Perspektiven unterschieden werden. Während alle ihre Berechtigung besitzen, ist für diese Arbeit in erster Linie die letztgenannte relevant:

Big Data als Geschäftsmodell: Beiträge aus der Perspektive der BWL oder Wirtschaftsinformatik betonen Möglichkeiten der Wertschöpfung, die durch die neue Nutzung bereits vorhandener Informationen entstehen. Das betrifft nicht nur Innovationen im Sinne neuer Produkte (etwa Online-Tools zum Vergleich von Hotel- oder Flugpreisen), sondern hat Auswirkungen auf das Unternehmensmodell als Ganzes (z. B. B. MARR 2015): Auf Intuition

oder Erfahrung basierende Entscheidungen des Managements werden in diesem Sinne durch ein *data-driven decision making* abgelöst – Business Analytics bzw. Business Intelligence wandelt sich zu bzw. integriert Data Science als zentralen Bestandteil.

Die vier Vs von Big Data: Eng verbunden mit der vorigen Perspektive ist die Definition anhand allgemeiner Charakteristika der Datengrundlage, welche als die vier Vs von Big Data bekannt sind – *Volume* (Daten haben hohen Speicherbedarf), *Velocity* (werden in nahezu Echtzeit generiert, sind volatil), *Variety* (sind strukturell vielfältig) und *Veracity* (sind ungeordnet und fehlerbehaftet) (KITCHIN 2014, 2015). Die Liste wurde von verschiedenen Autor*innen um weitere Merkmale von Big Data erweitert (KITCHIN/MCARDLE 2016), u. a. um *Exhaustivity* (statt einer Stichprobenziehung wird ein System durch die Daten vollständig abgebildet; MAYER-SCHÖNBERGER/CUKIER 2013) und *Relationality* (gemeinsame Aspekte verbinden verschiedene Datenquellen zu einem großen Ganzen; BOYD/CRAWFORD 2012).[4]

Big Data als Backend-Problem von IT-Systemen: Die beschriebenen Eigenschaften von Big Data stellen besondere Herausforderungen an die IT. In der Informatik steht der Begriff folglich oftmals synonym zu bestimmten Techniken im Umgang mit großen Datenmengen, allen voran Map/Reduce und Hadoop (z. B. FASEL/MEIER 2016; vgl. Kap. 2.4.2) Diese stellen sicher, dass Datenmengen, welche die Kapazität eines einzelnen Speichermediums überschreiten, sowie Analyseschritte, welche nicht in den Arbeitsspeicher passen, dennoch zuverlässig und effizient bearbeitet werden können. Ein Big-Data-Profi ist aus dieser Perspektive also in erster Linie im Umgang mit Datenbanken versiert. Dies trifft zwar zuvorderst auf gelernte Informatiker*innen zu, allerdings werden technische Skills dieser Art in einem gewissen Rahmen auch für Sozialwissenschaftler*innen relevant (GÜNTHER/TRILLING/VAN DE VELDE 2017).

Big Data is not about the Data: Wie dieser Titel eines programmatischen Beitrags von King (2016) beschreibt, spielen aus Sicht der Sozialwissenschaften bei Big Data weder die Daten noch ihre Größe die Hauptrolle. Vielmehr

4 Ein ausführlicher Überblick über die in der Diskussion vorgeschlagenen Charakteristika von Big Data findet sich bei Kitchin (2015) oder Kitchin und McArdle (2016).