

Elisabeth Günther

# Topic Modeling

Algorithmische Themenkonzepte  
in Gegenstand und Methodik der  
Kommunikationswissenschaft

HERBERT VON HALEM VERLAG

### **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](mailto: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.

# 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 <i>Algorithmische Themen im Gegenstand der KW</i>	22
1.1.2 <i>Algorithmische Themen in der Methodik der KW</i>	23
1.1.3 <i>Methodologische Problemstellung: Vergleich der     manuell-deduktiven und automatisch-induktiven     Themenanalyse</i>	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 <i>Zwischen Gesellschaftswandel, Datenbankproblem     und Pathos</i>	36
2.1.2 <i>Digitalisierung und Datafizierung</i>	39
2.1.3 <i>Algorithmische Logik als Treiber der Automatisierung</i>	43

2.2	Künstliche Intelligenz	47
2.2.1	<i>Wesen und Ziel der KI</i>	49
2.2.2	<i>Geschichte der KI</i>	56
2.2.3	<i>KI-Trends und Zukunftsprognosen</i>	60
2.3	Implikationen für Alltag und Gesellschaft	65
2.3.1	<i>Techno-Social Environment</i>	66
2.3.2	<i>Wie verändern Digitalisierung und Big Data die Choice Architecture?</i>	68
2.3.3	<i>Reverse-Turing-Test: Dehumanisierung im Techno-Social Environment</i>	72
2.3.4	<i>Ethik für Algorithmen</i>	78
2.4	Implikationen für die empirische Forschung	87
2.4.1	<i>Big Data und Algorithmen in der empirischen Forschung</i>	88
2.4.2	<i>Data Science und Computational Social Sciences</i>	94
2.4.3	<i>Kritik an Big Data und Co</i>	100
2.5	Zwischenfazit: Algorithmen und Code in der Makroperspektive	104
2.5.1	<i>Big Data und KI in Gegenstand und Methodik der KW</i>	104
2.5.2	<i>Zentrale Folgerungen für die Computational Social Sciences</i>	106
3.	Algorithmische Verarbeitung natürlicher Sprache	111
3.1	Sprache als Zeichensystem (de Saussure)	114
3.1.1	<i>Zentrale Relevanz der Sprache</i>	114
3.1.2	<i>Dyadischer Zeichenbegriff und Weiterführung</i>	116
3.2	Zeichenlehre als Erkenntnistheorie (Peirce)	118
3.2.1	<i>Zeichenbegriff und Grundlagen</i>	118
3.2.2	<i>Syntaktik, Semantik und Pragmatik: Weiterführung nach Morris</i>	123
3.2.3	<i>Übergang in die Stochastik: Deduktion, Induktion und Abduktion</i>	124
3.3	Information und Kommunikation (Shannon/Weaver)	127
3.3.1	<i>Von der Zeichenlehre zur Informationstheorie</i>	129
3.3.2	<i>Definition von Information als stochastischer Prozess</i>	133
3.4	Algorithmische Verarbeitung von Sprache und Information	137
3.4.1	<i>Natural Language Processing</i>	138
3.4.2	<i>Machine Learning</i>	142

3.5	Kognition und sprachliche Repräsentation	146
3.5.1	<i>Einführung in die Kognitionswissenschaft</i>	148
3.5.2	<i>Computational Theory of Mind</i>	151
3.6	Zwischenfazit: Algorithmen und Code in der Mikroperspektive	155
3.6.1	<i>Computationale Verarbeitung natürlicher Sprache</i>	155
3.6.2	<i>Mensch vs. Maschine?</i>	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	<i>Motiv 1: Eine strikte Dichotomie aus ›Mensch vs. Maschine‹ greift fehl</i>	163
4.2.2	<i>Motiv 2: KI schafft keine ›Magic Buttons‹ – vom Ob zum Wie</i>	164
4.2.3	<i>Motiv 3: Die KW hat die Kompetenz und Verantwortung, im Diskurs um algorithmische Systeme eine zentrale Rolle einzunehmen</i>	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	<i>Themenbegriff, -konstrukt, -konzept und -variable</i>	172
5.1.2	<i>Das Konstrukt ›Thema‹ in der KW-Forschungstradition</i>	173
5.2	Themen konstituieren Öffentlichkeit	175
5.2.1	<i>Drei Ebenen von Thematisierung</i>	175
5.2.2	<i>Integrationsfunktion von Themen</i>	177
5.3	Themenauswahl konstruiert Medienrealität	179
5.3.1	<i>Gatekeeper</i>	180
5.3.2	<i>News-Bias</i>	181
5.3.3	<i>Nachrichtenwert</i>	182

5.4	Einfluss und Beeinflussung massenmedialer Thematisierung	184
5.4.1	<i>Praeludium: Der Issue-Begriff in der politischen Kommunikation</i>	185
5.4.2	<i>Wirkungshypothese im Agenda-Setting-Ansatz</i>	187
5.4.3	<i>Issues als Gegenstand der strategischen Kommunikation</i>	189
5.4.4	<i>Neue Aspekte algorithmisch strukturierter Öffentlichkeit</i>	192
5.5	Thematische Ausdifferenzierung führt zu Fragmentierung?	195
5.5.1	<i>Ursprünge der Fragmentierungsthese</i>	196
5.5.2	<i>Stratifizierte und segmentierte Öffentlichkeit</i>	197
5.5.3	<i>Themen in der Filterblase</i>	199
5.5.4	<i>Kritik an der Fragmentierungsthese</i>	201
5.6	Zwischenfazit: Das Konstrukt ›Thema‹ in der Makroperspektive	205
5.6.1	<i>Die kommunikationswissenschaftliche Forschungstradition</i>	205
5.6.2	<i>Thematisierung in algorithmisch strukturierter Öffentlichkeit</i>	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	<i>Propositionsmodelle</i>	216
6.2.2	<i>Makrostrukturmodelle: Einzug des Themenbegriffs</i>	218
6.3	Textverstehen top-down: Schematheoretische Ansätze	221
6.3.1	<i>Schema (Bartlett)</i>	221
6.3.2	<i>Skripte (Schank/Abelson)</i>	223
6.3.3	<i>Frames (Minsky)</i>	224
6.3.4	<i>Konstruktions-Integrations-Modell (Kintsch)</i>	225
6.3.5	<i>Das Themenkonstrukt in den schematheoretischen Ansätzen</i>	226
6.3.6	<i>Interdisziplinäre Anleihen</i>	229
6.4	Text im Kontext: Textverstehen anhand mentaler Modelle	232
6.4.1	<i>Mentale Modelle (Johnson-Laird)</i>	233
6.4.2	<i>Situationsmodelle (van Dijk/Kintsch)</i>	235
6.4.3	<i>Das Themenkonstrukt in mentalen Modellen</i>	237

6.5	Zwischenfazit: Das Konstrukt ›Thema‹ in der Mikroperspektive	238
6.5.1	<i>Die Entwicklung des Forschungsbereichs Textverstehen</i>	238
6.5.2	<i>Themendefinition im Bereich des Textverstehens</i>	240
6.5.3	<i>Verbindung zum Forschungsbereich KI</i>	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	<i>Themendefinitionen, Themenkonzepte</i>	244
7.2.2	<i>Ineinandergreifen der beiden Perspektiven</i>	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	<i>Kategorienbildung</i>	257
8.2.2	<i>Ereignisbezug von Themen</i>	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	<i>LSA als ›Theory of Meaning‹</i>	272
9.2.2	<i>LSA als Textanalyseverfahren</i>	273
9.2.3	<i>Mathematische Grundlagen</i>	274

9.3	Probabilistic Latent Semantic Analysis (PLSA)	281
9.3.1	<i>Generatives Modell</i>	282
9.3.2	<i>Schätzen der Modellparameter</i>	284
9.3.3	<i>Zentrale Errungenschaften im Topic Modeling</i>	287
9.4	Latent Dirichlet Allocation (LDA)	288
9.4.1	<i>LDA = PLSA + Bayes</i>	290
9.4.2	<i>Generatives Modell</i>	291
9.4.3	<i>Vorbereitung der Datenanalyse</i>	293
9.4.4	<i>Schätzen der Modellparameter</i>	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	<i>Reliabilität als triadische Beziehung</i>	311
10.4.2	<i>Apparat</i>	313
10.4.3	<i>Kontext</i>	315
10.4.4	<i>Material</i>	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	<i>These 1: Innovation und Allgegenwärtigkeit</i>	323
11.2.2	<i>These 2: Duale Relevanz</i>	324
11.2.3	<i>These 3: TM ≠ manuelle Themenanalyse + Automatisierung</i>	324
11.2.4	<i>These 4: Umfassende Integration</i>	326
11.2.5	<i>These 5: Menschliche Intelligenz</i>	327



11.2.6	<i>These 6: Konzeptionelle Skills</i>	327
11.2.7	<i>These 7: Kompetenz und Verantwortung</i>	329
11.3	Kritische Reflexion	330
11.4	Der Blick nach vorn	331
11.4.1	<i>Sapir-Whorf-Hypothese: Zum Einfluss der (künstlichen) Sprache</i>	332
11.4.2	<i>Computational Turn, Computational Divides?</i>	334
11.4.3	<i>Das menschliche Themenverständnis im Techno-Social Environment</i>	336
	Literaturverzeichnis	339

## TEIL I

### RELEVANZ UND FRAGESTELLUNG

## 1. DIE DUALE BEDEUTUNG VON TOPIC MODELING FÜR DIE KW

### 1.1 Forschungsleitendes Interesse

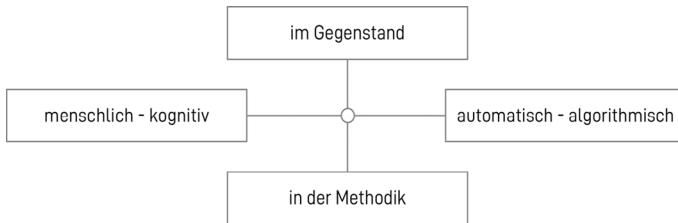
Als zentrales Konstrukt der Kommunikationswissenschaft (KW) ist *das Thema* heute relevanter denn je. Für unser Fach besitzt es durch seine grundlegende Funktion als Ordnungseinheit der öffentlichen Meinung eine besondere Bedeutung. Über Themen wird es möglich, sich aufeinander zu beziehen, zu wissen, *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,<sup>1</sup> 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 ATTEVELD 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 ATTEVELD 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

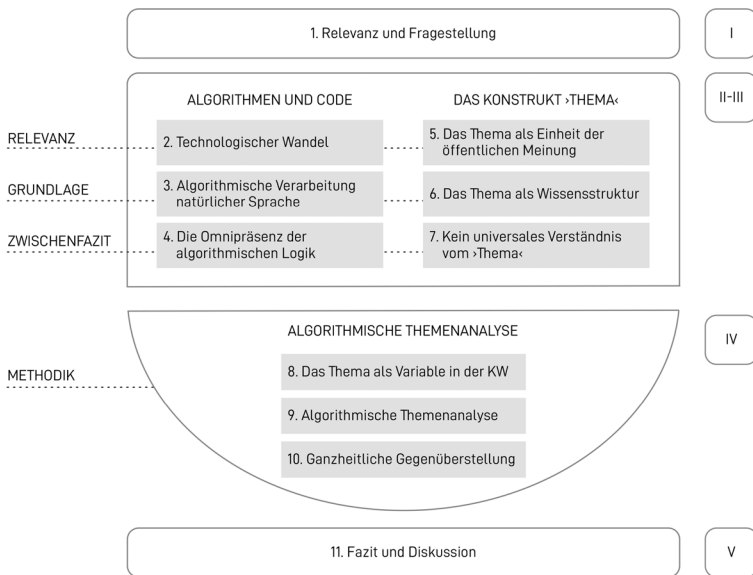
Da es sich bei der algorithmischen Themenanalyse via TM um eine Methodeninnovation für die kw handelt, muss nicht nur das Verfahren selbst, sondern auch der Kontext algorithmischer Verfahren insgesamt eingeführt werden. Die Arbeit legt also zum einen eine Grundlage für die css als Teilbereich der Data 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.