Topic Modeling

Melanie Althage

https://orcid.org/0000-0001-5233-1061

Abstract Die Methode Topic Modeling wird häufig in den Digital Humanities eingesetzt, um die inhaltliche Struktur umfangreicher Textsammlungen, etwa mit Blick auf Diskurse oder Publikationstrends, zu analysieren. Der vorliegende Beitrag bietet eine Einführung in ihre methodologischen Grundlagen sowie einen Überblick über verschiedene Topic-Modeling-Algorithmen und ihre jeweiligen Anwendungsfelder. Zudem werden zentrale Workflow-Schritte wie die Aufbereitung der Textdaten (Preprocessing) und die Evaluation der Modellierungsergebnisse vorgestellt. Ziel ist es, eine solide Grundlage für die kritisch-reflektierte Nutzung dieser Methode in der theologischen Forschung bereitzustellen.

Keywords Topic Modeling, Text Mining, quantitative Textanalyse, Machine Learning, Natural Language Processing, Blended Reading, Distant Reading

1. Einleitung

In seinem Beitrag von 2006 fragte Gregory Crane „What do you do with a million books?“ (Crane 2006). Eine Frage, die mit dem tagtäglich wachsenden Angebot an digital verfügbaren Quellen immer mehr Relevanz gewinnt (siehe auch Stulpe & Lemke 2016, 18). Ein nicht unerheblicher Teil dieser Quellen liegt jedoch in nur schwach strukturierter Form vor, was den Zugriff auf die in ihnen enthaltenen Informationen erschwert. Wie können wir diese Fülle von Informationen und potenziellem Wissen effektiv erschließen und für die Forschung nutzbar machen? Eine Antwort hierauf bietet Topic Modeling, ein Clustering-Algorithmus, der große Textkorpora durch gemeinsame sprachliche Muster thematisch strukturiert und explorierbar macht. Wenn wir annehmen, dass sich Themen oder inhaltliche Konzepte durch ein spezifisches Set von Begriffen ausdrücken, die in den unterschiedlichen historischen Quellen häufig gemeinsam auftreten, können solche automatisierten Verfahren der Mustererkennung einen wertvollen Beitrag für die Forschung leisten.

In den digitalen Geistes- und Geschichtswissenschaften hat sich Topic Modeling dementsprechend als vielseitiges Instrument für eine breite Palette an Forschungsfragen etabliert. Es ermöglicht die Analyse von Forschungstrends in wissenschaftlichen Zeitschriften (exemplarisch: Mimno 2012; Wehrheim 2019; Wehrheim et al. 2022),

Melanie Althage: Topic Modeling, in: Christopher A. Nunn und Frederike van Oorschot (Hrsg.): Kompendium Computational Theology 1. Forschungspraktiken in den Digital Humanities, Heidelberg: heiBOOKS, 2024, S. 265–288. DOI: https://doi.org/10.11588/heibooks.1459.c21916

die Untersuchung von Diskursstrukturen in verschiedenen Publikationsorganen (Völkl et al. 2022; Bunout & von Lange 2019) oder auch die Einordnung der Digital Humanities als Disziplin im Vergleich zu anderen Fächern (Luhmann & Burghardt 2021). Auch im Kontext der Theologie findet die Methode zunehmend Anwendung. Christopher A. Nunn etwa präsentierte Topic Modeling in seiner Studie als Teil eines breiteren *Distant-Reading*-Ansatzes und nutzte den *DARIAH-DE TopicsExplorer* (Simmler et al. 2019), eine benutzerfreundliche Software, um ethische Themen in den Briefen des Augustinus von Hippo zu beleuchten (Nunn 2022). Mark Graves hingegen vertiefte sich für seine Studie zur Moraltheologie Thomas von Aquins in die modelltheoretischen und mathematisch-computationellen Aspekte von Topic Modeling. Er zeigte eindrücklich, wie die Methode genutzt werden kann, um komplexe moralische und theologische Konzepte in ihren verschiedenen Facetten zu analysieren und anschließend deren Einfluss auf päpstliche Enzykliken zu untersuchen (Graves 2022).

Um weitere Studien in der Theologie anzuregen, zielt der vorliegende Beitrag darauf, einen kritisch-reflektierten Einstieg in Methode und Workflow des Topic Modelings sowie dessen vielfältige Varianten und Konfigurationsmöglichkeiten zu bieten. Dabei werden nicht nur die Potenziale, sondern auch die Limitierungen und Herausforderungen skizziert, die bei der Anwendung dieses Verfahrens im Forschungsprozess zu berücksichtigen sind. Zunächst wird dazu das Grundkonzept von Topic Modeling erläutert, gefolgt von einem Überblick über verschiedene Algorithmen und deren Anwendungsfälle. Eine detaillierte Darstellung der mathematischen Prinzipien hinter den einzelnen Verfahren wird bewusst ausgespart; für vertiefende Informationen wird auf die jeweilige Fachliteratur verwiesen. Abschließend werden die zentralen Aspekte der Datenaufbereitung sowie Evaluation der Modellierungsergebnisse erörtert. Ziel ist es, eine solide Grundlage und erste Orientierung für die Anwendung von Topic Modeling in der theologischen Forschung zu bieten.[1]

1 Die in den Abbildungen 1 bis 3 präsentierten Beispiele für Topics basieren auf den zwischen 1996 und Juni 2019 auf dem Fachkommunikationsportal H-Soz-Kult (https://www.hsozkult.de/, zuletzt aufgerufen am 19. 07. 2024) veröffentlichten deutschsprachigen Buchrezensionen (15 103 mit rund 18 Millionen Wörtern). Die ausgewählten Topics stammen aus einem Modell mit insgesamt 80 Topics, das mit dem in der Software *MALLET* (McCallum 2002) implementierten Algorithmus *Latent Dirichlet Allocation* (LDA; Blei et al. 2003) über den Python-Wrapper in *Gensim* (Řehůřek & Sojka 2010) im Rahmen des laufenden Dissertationsprojekts der Autorin generiert wurde; Arbeitstitel des Projekts: „Mining the Historian's Web – Methodenkritische Reflexion quantitativer Verfahren zur Analyse genuin digitaler Quellen am Beispiel der historischen Fachkommunikation". Sie stellen einen früheren Bearbeitungsstand dar. Die zur Illustration exemplarisch gegenübergestellten Topics in Tabelle 1 basieren wiederum auf ausgewählten deutschsprachigen Funeralschriften des 17. Jahrhunderts (299 mit rund 3 Millionen Wörtern). Diese wurden im Rahmen des DFG-Projekts „AEDit Frühe Neuzeit" in Zusammenarbeit mit dem Deutschen Textarchiv digitalisiert und gemäß den DTA-Transkriptionsrichtlinien maschinenlesbar aufbereitet. Zum Subkorpus „AEDit Frühe Neuzeit" in: Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften, Berlin 2024, URL: www.deutschestextarchiv.de/search/metadata?corpus=aedit,

2. Methodische Grundlagen

Topic Modeling ist eine Methode des Text Mining, die darauf abzielt, umfangreiche Textkorpora inhaltlich zu verstehen und zu erschließen (einführend: Blei 2012a, b; Brett 2012). Anders als bei Klassifikationsalgorithmen, bei denen Kategorien explizit vorgegeben werden *(supervised Machine Learning),* basiert Topic Modeling auf einem generativen probabilistischen Modellierungsprozess *(unsupervised Machine Learning).* Das heißt, in diesem Prozess werden die „Kategorien" bzw. „Topics" unmittelbar aus den Daten abgeleitet. Damit ähnelt es traditionellen Indexierungspraktiken, die seit dem 18. Jahrhundert genutzt werden, um effizient und gezielt auf bestimmte Texteinheiten zuzugreifen; es unterscheidet sich jedoch in seiner Methode: Statt fester Schlagworte werden in sich heterogene Wortcluster durch wahrscheinlichkeitstheoretische Berechnungen generiert (vgl. Piper 2018, 66–75; siehe exemplarisch Abb. 1).

Abb. 1 Exemplarische Auswahl von religionsgeschichtlichen Topics für die auf H-Soz-Kult veröffentlichten Buchrezensionen; Visualisierungsform: Wordclouds mit einer Gewichtung der Wörter gemäß ihrer Relevanz für das Topic.

Im klassischen Topic-Modeling-Prozess gemäß der *Latent Dirichlet Allocation* (LDA; Blei et al. 2003) wird von der Annahme ausgegangen, dass sich die Dokumente eines umfangreichen Korpus zu unterschiedlichen Anteilen aus einem festen Set von Themen zusammensetzen, und ferner, dass sich diese Themen als *latente,* i. e. verborgene, sprachliche Strukturen bzw. Muster aus den Textdaten über die Generierung von Topics rekonstruieren lassen (zu den Annahmen sowie zum Modellierungsprozess siehe Blei 2012b, 78–82). Um dies zu illustrieren, nehmen wir an, wir hätten Zugriff auf eine digital verfügbare Bibliothek mit theologischen Werken zum Christentum, deren inhaltliche Klassifikation über Schlagwörter verloren gegangen ist. Die Werke

zuletzt aufgerufen am 19. 07. 2024. Zur Vergleichbarkeit wurden auch diese Modelle mit dem MALLET-Wrapper von Gensim generiert.

könnten unter anderem Diskussionen über die Trinität, Erlösung, Ethik und Moral sowie Bibelexegesen enthalten. Topic Modeling ermöglicht nun die Rekonstruktion dieser latenten inhaltlichen Kategorien. Dabei werden allerdings nicht konkrete Schlagwörter wie „Trinität", sondern Gruppen von Wörtern (z. B. „Gott, Jesus, Geist, Vater, Sohn, heilig, Dreifaltigkeit, …") generiert, die statistisch auffällig oft gemeinsam in den einzelnen Dokumenten vorkommen. Ziel ist es also, Wortgruppen zu identifizieren, die durch die Interpretation ihrer Zusammensetzung einen Überblick über die inhaltliche Struktur unserer Bibliothek und ihrer einzelnen Werke ermöglichen.

Im ersten Schritt wird dazu jedes Wort der Werke zunächst zufällig einem Topic zugeordnet. Ebenso wird jedem Werk eine zufällige Zusammenstellung von Topics zugeschrieben. Im nächsten Schritt erfolgt die Überprüfung dieser initialen Zuweisungen. Dabei wird anhand der Häufigkeit und Kookkurrenz eines Wortes mit anderen Wörtern kontrolliert, ob die aktuelle Topic-Zuweisung angemessen ist oder es besser zu einem anderen Wortcluster passen würde. Ähnliches gilt für die einzelnen Dokumente: Ein Werk, in dem oft die Wörter „Jesus", „heilig", „Gnade", „Vergebung", „Sünde" und „befreien" vorkommen, könnte bspw. von dem Konzept der Erlösung handeln, wurde aber vielleicht initial dem Topic „Trinität" zugewiesen. Solche Zuweisungen werden dann aktualisiert.[2] Dieser Prozess wird viele, oft tausende Male wiederholt (sog. Iterationen), bis das Korpus „sinnvoll" strukturiert ist. Sinnvoll heißt in diesem Fall, dass kaum noch Zuweisungsänderungen nötig sind, da sich das Modell stabilisiert hat.[3]

Schlussendlich erhalten wir ein statistisches Modell unserer Bibliothek, das die Zuordnung der einzelnen Werke zu theologischen Themen ermöglicht und somit eine effiziente Orientierung innerhalb unseres Korpus gewährleistet. Repräsentiert wird dieses Modell durch zwei Formen von *Outputs.* Einerseits wird das Datenkorpus üblicherweise als *Document-Topic-Matrix* repräsentiert, also als eine Tabelle, in der für jedes Dokument die einzelnen Topic-Gewichtungen aufgeführt werden. Andererseits wird analog dazu eine *Topic-Word-Matrix* generiert, welche die prozentuale Gewichtung der einzelnen Wörter für die einzelnen Topics aufschlüsselt (siehe dazu auch Althage 2022, 260 f.). Damit abstrahieren wir von den konkreten Werken und greifen uns bestimmte, relevante Eigenschaften der einzelnen Texte (= die statistisch auffälligen Muster im Sprachgebrauch) als numerische Abbildungen heraus mit dem Ziel, das Korpus inhaltlich zu erschließen.

Für üblicherweise qualitativ-texthermeneutisch forschende Wissenschaften wie die Theologie mögen quantitative Textanalyseverfahren wie Topic Modeling, die

2 Diese „Zuweisung" eines Topics zu einem Dokument wird dabei als prozentualer Wahrscheinlichkeitswert ausgedrückt, der etwas über die Wahrscheinlichkeit des Auftretens dieses Wortclusters im Dokument oder Gesamtkorpus aussagt.

3 Bei den meisten Topic-Modeling-Algorithmen muss selbst festgelegt werden, wie viele Topics für ein Korpus in wie vielen Wiederholungen (Iterationen) zu generieren sind; hier empfiehlt es sich je nach Korpus und zu erwartender Themen-Diversität unterschiedliche Konfigurationen auszuprobieren (siehe dazu auch Kapitel 4).

Texte in Form von numerischen Repräsentationen verarbeiten, zunächst ungewohnt erscheinen, sie bieten allerdings mit ihrem makroanalytischen Ansatz (vgl. Jockers 2013; Graham et al. 2016) neuartige Perspektiven auf ihre Forschungsgegenstände. Zahlreiche Anwendungsmöglichkeiten für die Theologie sind denkbar. Vor allem für die Analyse dominanter Themen, Diskurse oder Konzepte – etwa in Predigten, Briefwechseln, Werken der Kirchenväter oder auch der Forschungsliteratur – kann diese Methode fruchtbar gemacht werden. Untersucht werden kann dabei nicht nur, wie sich die Schwerpunkte im Verlauf der Zeit verändern, sondern auch, wie unterschiedliche Themen oder Konzepte zueinander in Beziehung stehen. Durch die Analyse von Texten verschiedener religiöser Gruppen bzw. Autor*innen könnten zudem Unterschiede und Gemeinsamkeiten in den theologischen Ansichten herausgearbeitet werden; auch ließen sich diverse Textarten hinsichtlich ihrer sprachlichen und thematischen Charakteristika beleuchten.

Diese Anwendungsmöglichkeiten (siehe auch Althage 2022, 259 f.) ergeben sich insbesondere durch die Fähigkeit des Topic Modelings, Texte als Daten zu begreifen und so eine systematische und skalierbare Analyse durchzuführen. In traditionellen Forschungskontexten würden üblicherweise Stichproben oder Fallbeispiele für eine exemplarische Untersuchung herangezogen. Im Gegensatz dazu lassen sich computationelle Verfahren bei ausreichender Rechenkapazität auf beliebig große Quellenkorpora anwenden und somit auch Untersuchungszeiträume ausweiten. Solch umfangreiche Korpora sind für Menschen nur schwer mit gleichbleibenden Untersuchungs- und Relevanzkriterien zu überblicken, denn der menschliche Erkenntnisbildungsprozess und damit das, was aus den Quellen extrahiert wird, entwickelt sich dynamisch und ist von einer Vielzahl von Faktoren beeinflusst (Stichwort: Hermeneutischer Zirkel); für den Computer ist es dagegen eine Leichtigkeit, sehr große Datenmengen systematisch und konsistent zu verarbeiten. Die zu generierenden Topics speisen sich dabei allein aus den Daten und beruhen nicht auf zuvor mit bestimmten Annahmen definierten Kategorien.[4] Ein zusätzlicher Vorteil ist, dass Topic Modeling grundsätzlich auf beliebige Sprachen und daher auf beliebige Quellenbestände angewendet werden kann. Durch solch eine systematische Herangehensweise ermöglicht Topic Modeling eine tiefgreifende Analyse nicht nur eines, sondern tausender Dokumente, um verborgene thematische Strukturen zu identifizieren und zu interpretieren und dadurch ein tieferes Verständnis von den Eigenheiten des Forschungsobjektes zu gewinnen und vorgefertigte Annahmen herauszufordern.

4 Gleichzeitig bedeutet das allerdings auch, dass Art und Umfang der Datenvorverarbeitung das Modellierungsergebnis substanziell beeinflussen, s. Kapitel 4.2.

3. Topics: Definition und epistemologische Grenzen

Angesichts der zuvor skizzierten Anwendungsbreite von Topic Modeling ist es essentiell, den Begriff des „Topics" genauer zu definieren. Dies minimiert das Risiko von Fehlannahmen über die Erkenntnispotentiale dieser Methode. Wie bei vielen Text-Mining-Verfahren beruht die Modellierung von Topics hauptsächlich auf der Zählung von Worthäufigkeiten. Ein Topic ist in diesem Kontext eine Wahrscheinlichkeitsverteilung über das Vokabular der Textkollektion, die die Kookkurrenz bestimmter Wörter beschreibt (Blei 2012b, 78). Auch wenn Begrifflichkeiten wie „Topics" möglicherweise an die Topik oder „Topoi" erinnern (Piper 2018, 66–75; Horstmann 2018, 4–7), sind mit ihnen keine epistemologischen Aussagen über die Wahrscheinlichkeiten der Kookkurrenz, also über das gemeinsame Vorkommen, hinaus verbunden (vgl. Blei et al. 2003, 996, Anm. 1; Althage 2022, 267; siehe dazu auch Shadrova 2021). In geisteswissenschaftlichen Forschungskontexten ist die Anwendung von Topic Modeling allerdings für gewöhnlich mit zwei Annahmen verbunden: Erstens, die der Kohärenz der Topics, die besagt, dass die einem Topic zugeordneten Begriffe eine thematische oder konzeptuelle Verwandtschaft aufweisen sollten; und zweitens, die Annahme der Bedeutungsstabilität, nach der ein bestimmtes Topic, wenn es mehreren Dokumenten zugewiesen wurde, in all diesen Kontexten die gleiche Bedeutung oder Relevanz haben sollte (Schmidt 2012, 49).

Topic-Modelle *verstehen* die Bedeutung und Konzepte indes nicht, die Menschen mit den Wörtern eines Textes verbinden, denn der Computer ist in dieser Hinsicht „semantisch blind" (Schwandt 2018, 108. 133). Entsprechend kritisch machte Benjamin Schmidt deutlich, dass Topics nicht an sich bedeutsam sind, sondern es erst durch unsere Interpretation werden (Schmidt 2012; siehe auch Horstmann 2018, 10). David Blei wiederum wies darauf hin, dass Topics in diesem Sinne wie Themen aussehen können, da Wörter, die häufig miteinander kookkurrieren, dazu tendieren, zum selben Themenfeld zu gehören (Blei 2012a, 9). Dies basiert auf dem Prinzip der distributionellen Semantik (Piper 2018, 13; Schöch 2017, 14), nach dem sich die Bedeutung von Wörtern aus der gemeinsamen Vorkommenshäufigkeit mit anderen Wörtern in einem bestimmten Kontext ergibt. Ein solcher Kontext kann ein Dokument, ein Absatz, aber auch ein einzelner Satz sein. Um nun die Textdaten auf einer *Bedeutungsebene* zu erschließen, werden diese Häufigkeitsbeziehungen zwischen den Wörtern beispielsweise durch Koordinaten in einem Vektorraum *(vector space)* numerisch repräsentiert und dadurch computationell verarbeitbar (Turney & Pantel 2010; Blei 2012a, 9; Piper 2018, 13–18; siehe dazu auch Althage 2022, 266 f.).

Auch wenn Topics also häufig der Einfachheit halber mit „Themen" oder anderen semantischen Kategorien gleichgesetzt werden, darf nicht vergessen werden, dass sie nicht gleichbedeutend sind (Uglanova & Gius 2020, 72). Das zeigt sich zusätzlich daran, dass in einem Topic-Modell immer auch Wortcluster vorhanden sind – und das zum Teil sehr stark gewichtet –, die eher allgemeinere stilistische Eigenschaften einer bestimmten Textart beschreiben (*Metatopics,* siehe Abb. 2) oder auf eine Hetero-

Abb. 2 Beispiele für *Metatopics* aus den Rezensionen von H-Soz-Kult

Abb. 3 Beispiele für *Noisy Topics* aus den Rezensionen von H-Soz-Kult

genität in den Textdaten hindeuten, die sich u. a. in sprachraumspezifischen Topics niederschlagen (siehe Abb. 3; vgl. zu verschiedenen Topic-Formen Boyd-Graber et al. 2014, 234–237; Schöch 2017, 23–26; Althage 2022, 267–269). Gerade die letzteren, hier als *Noisy Topics* bezeichneten, könnten ein Ansatzpunkt für eine weitergehende Vorverarbeitung der Textdaten sein (siehe Kap. 4.2). Im Kontext der zuvor genannten Annahmen sollte zudem beachtet werden, dass ein Topic mit der gleichen Gewichtung in zwei verschiedenen Dokumenten auf der Wortebene vollkommen unterschiedliche Schwerpunkte setzen kann.[5] Topics stellen also keine eigenständigen epistemologischen Einheiten mit festem semantischen Kern dar, sondern vielmehr ein hermeneutisches Instrument (Rockwell & Sinclair 2016). Sie ermöglichen eine strukturierte

5 Siehe etwa die Rezensionen unter https://www.hsozkult.de/publicationreview/id/reb-25382 und https://www.hsozkult.de/publicationreview/id/reb-26856, die mit einer Distribution von jeweils 32 % für das Topic 42 (siehe Abb. 1) unterschiedliche inhaltliche Dimensionen des Topics widerspiegeln. Beide wurden zuletzt am 18.06.2024 aufgerufen.

Annäherung an umfangreiche Textsammlungen, sollten jedoch immer im Kontext der zugrundeliegenden Quellen interpretiert werden, wobei insbesondere die Annahmen der Kohärenz und Bedeutungsstabilität zu prüfen sind.

4. Topic Modeling Workflow

Die konkrete Anwendung von Topic Modeling im Forschungskontext erfordert einen sorgfältig durchdachten und kritisch reflektierten sowie dokumentierten Workflow (siehe Abb. 4, nächste Seite). Dieser beinhaltet die Auswahl eines geeigneten Verfahrens, die Vorbereitung der Textdaten *(Preprocessing)* für die Generierung der Topic-Modelle sowie die Evaluation der Ergebnisse unter Berücksichtigung verschiedener Konfigurationen des Preprocessings und der Topic-Modellierung. In der Regel handelt es sich hierbei um einen iterativen Prozess, in dem immer wieder zwischen den einzelnen Bearbeitungsschritten hin und her gesprungen werden kann, um die Modellierungsergebnisse mit Blick auf die Fragestellung zu optimieren. Ist dann ein adäquates Topic-Modell gefunden, bieten sich vielfältige Visualisierungsmöglichkeiten für die Ergebnisse an, von einfachen Wortlisten und Wordclouds über Balken-, Liniendiagramme oder Scatterplots zur Veranschaulichung von Ausprägungen und Entwicklungen, Heatmaps für Korrelationen bis hin zu Netzwerken für Relationen zwischen den Clustern. Dieses Kapitel konzentriert sich nun nachfolgend auf die Auswahl eines passenden Algorithmus, das Preprocessing der Daten sowie die Evaluation der Ergebnisse.

4.1 Wahl des Topic-Modeling-Verfahrens

Zu Beginn des Forschungsvorhabens ist zu eruieren, welcher Algorithmus in welcher Implementierung für eine gegebene Fragestellung in Frage kommt (siehe überblickshaft Jelodar et al. 2019; Vayansky & Kumar 2020; Churchill & Singh 2022). Diese Entscheidung sollte auf einem Vergleich verschiedener Ansätze und ihrer jeweiligen Ergebnisse basieren (siehe als exemplarische Entscheidungshilfe Abb. 5, übernächste Seite). Dabei sind diverse Faktoren zu berücksichtigen, etwa die Übereinstimmung der theoretisch-methodologischen Annahmen des Verfahrens mit den eigenen Erkenntnisinteressen sowie die verfügbaren Konfigurationsmöglichkeiten (von der Topic-Anzahl bis zur Hyperparameteroptimierung) und ihre Auswirkungen auf den Output. Das Ziel ist eine kritische Auseinandersetzung mit den Potenzialen und Grenzen der in Betracht kommenden Verfahren.

Unter den zahlreichen Optionen hat sich LDA (Blei et al. 2003) als besonders populär in den digitalen Geisteswissenschaften erwiesen und ist auch in etlichen ge-

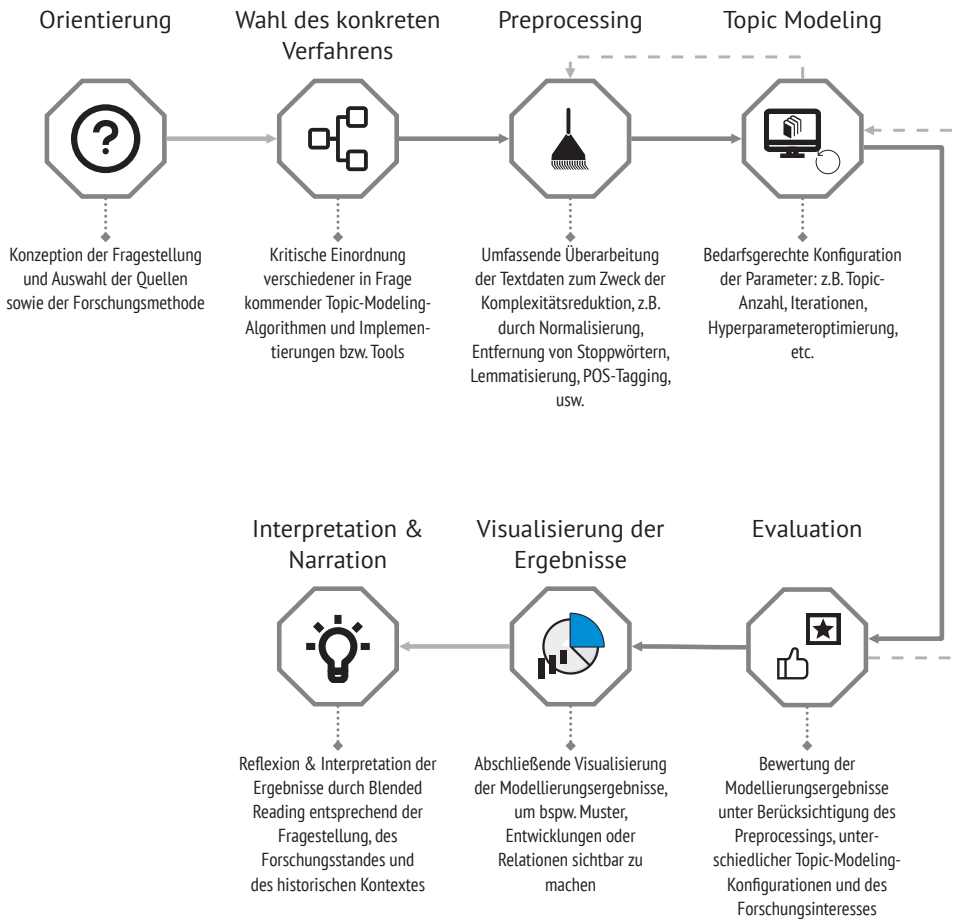

Abb. 4 Schematischer Topic-Modeling-Workflow

brauchsfertigen Werkzeugen und Programmbibliotheken implementiert.[6] Es wurde in den unterschiedlichsten Disziplinen bereits erfolgreich eingesetzt, um umfangreiche Textkollektionen im Sinne eines heuristischen Werkzeugs explorativ zu erschlie-

6 Zu berücksichtigen ist allerdings, dass es verschiedene Ausprägungen von LDA gibt, die entsprechend unterschiedliche Modellierungsergebnisse generieren können. So unterscheiden sich bspw. die Implementierungen in MALLET und Gensim hinsichtlich ihrer Inferenzalgorithmen für die Ableitung der Topics. Die Gensim-Implementierung ist für die Handhabung sehr großer Datenmengen konzipiert und legt einen Schwerpunkt auf die Performanz; die Ergebnisse können daher weniger kohärent sein. Im Gegensatz dazu beansprucht MALLET für die Modellierung der Topics zwar mehr Rechenzeit, produziert dafür aber i. d. R. kohärentere und robustere Modellierungsergebnisse – auch bei kleineren Textkorpora (siehe dazu Althage 2022, 261–263; Hodel et al. 2022; Boyd-Graber et al. 2014, 231–233).

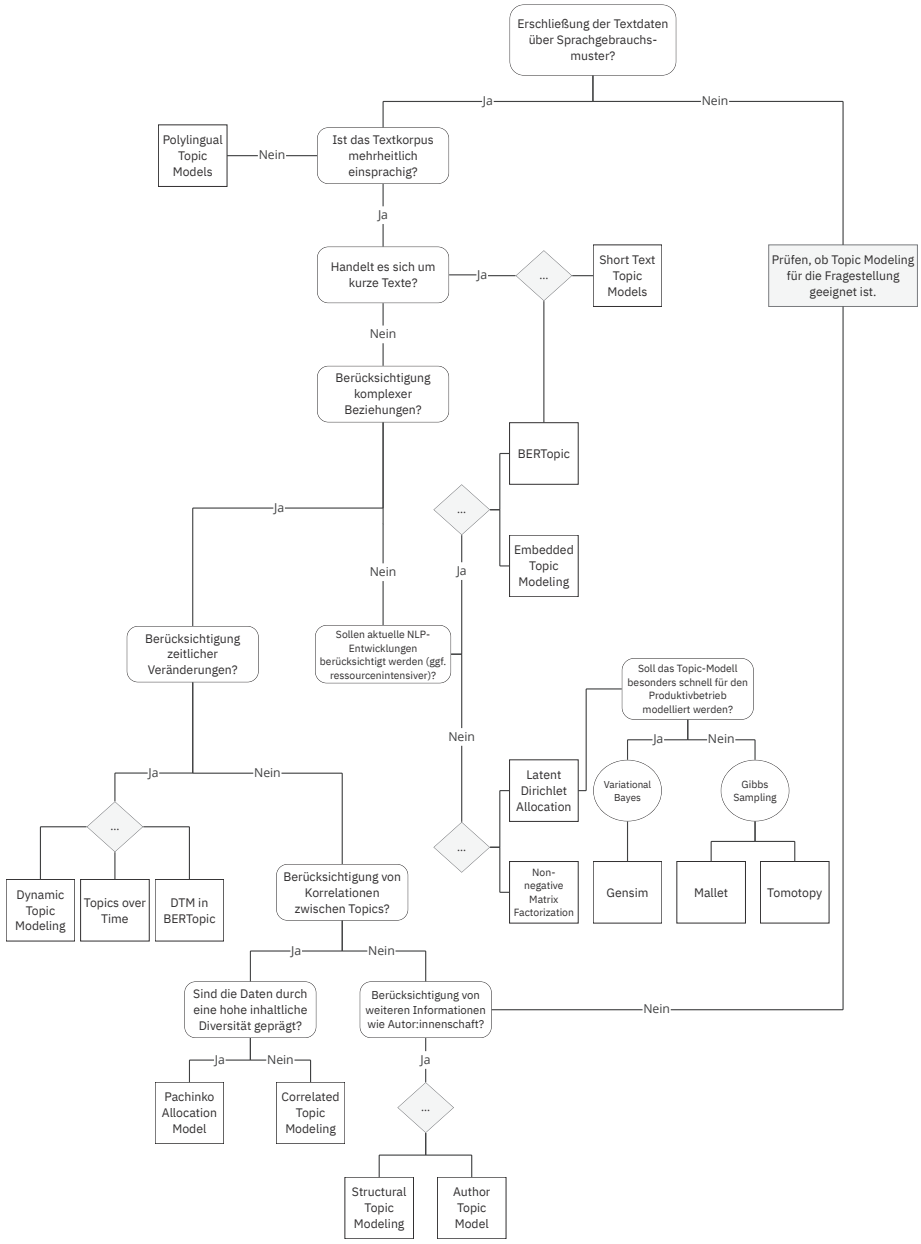

Abb. 5 Exemplarischer Entscheidungsbaum (aufbauend auf Vayansky & Kumar 2020, insb. 14, Abb. 8; Churchill & Singh 2022; Jelodar et al. 2019); einige zentrale Fragen zum Ziel oder den Eigenschaften des Forschungsgegenstandes können dabei helfen, das passende Verfahren oder Tool auszuwählen.*

* Zu *Non-negative Matrix Factorization* siehe: Lee & Seung 1999; *Topics over Time:* Wang & McCallum 2006; *Pachinko Allocation Model:* Li & McCallum 2006; zu *Embedded Topic Modeling* exemplarisch: Dieng et al. 2020.

ßen. In den digitalen Geisteswissenschaften wird es dabei auch häufig für historische Untersuchungen eingesetzt, die sich über längere Zeiträume erstrecken (z. B. Wehrheim et al. 2022; Snickars 2022; Grant et al. 2021). Da es allerdings weder die zeitliche und relationale Dimension der Daten noch ihre Kontextualität im Modellierungsprozess berücksichtigt, sind solche Informationen nachträglich auf das Modell anzuwenden (Althage 2022). Verfahren wie *Dynamic Topic Modeling* (DTM; Blei & Lafferty 2006; Grootendorst 2022; für ein Anwendungsbeispiel: Guldi 2019) berücksichtigen demgegenüber die Temporalität der Topics bereits im Modellierungsprozess. Damit können sie zeigen, wie z. B. Diskurse im Laufe der Zeit entstehen, sich entwickeln und wieder verschwinden.

Stehen wiederum statt der zeitlichen Dimensionen eher die Relationen zwischen verschiedenen Clustern im Vordergrund, bieten sich Verfahren wie *Correlated Topic Modeling* (CTM; Lafferty & Blei 2005; Blei & Lafferty 2007) an, mit dem sich Korrelationen zwischen Topics ermitteln lassen. *Structural Topic Modeling* (STM; Roberts et al. 2014; Küsters & Garrido 2020) ermöglicht es dagegen, die Wortcluster in Bezug auf spezifische Kontextinformationen zu modellieren. Dies ist besonders nützlich, wenn man untersuchen möchte, wie sich z. B. Faktoren wie Geschlecht und soziale Gruppenzugehörigkeit oder auch das Genre auf die Generierung der Topics auswirken. Auch der Einfluss von Autor*innenschaft auf das Topic-Modell kann mittels STM untersucht werden, daneben wurden aber auch spezielle *Author Topic Models* entwickelt (Rosen-Zvi et al. 2004).

Die Vielseitigkeit von Topic Modeling zeigt sich auch in seiner Anwendbarkeit auf verschiedene Textarten, von wissenschaftlichen Artikeln über historische Dokumente bis hin zu Tweets. Insbesondere für kürzere Texte wie besagte Tweets oder Titel können jedoch spezialisierte Modelle wie *Short Text Topic Models* (Cheng et al. 2014; Zuo et al. 2016; Zhao et al. 2021) besser geeignet sein. Bei mehrsprachigen Textsammlungen wiederum können *Polylingual Topic Models* (Mimno et al. 2009) oder *BERTopic* (Grootendorst 2022) verwendet werden, um thematische Konsistenzen über verschiedene Sprachen hinweg zu identifizieren.

Nach dieser Übersicht über verschiedene Topic-Modeling-Verfahren und deren Anwendungsmöglichkeiten stellt sich nun die Frage, wie diese Modelle praktisch in den Forschungsprozess integriert werden können. Dafür gibt es eine Vielzahl von Möglichkeiten; so können bspw. gebrauchsfertige Tools wie der *TopicsExplorer*[7] oder *Topics* in *Voyant Tools*[8] verwendet werden. Da die Konfigurationsmöglichkeiten gerade bei dieser Methode sehr große Auswirkungen auf die Ergebnisse haben, sollten diese Softwarelösungen vor allem als Einstiegshilfe verstanden werden, um sich mit dem Modellierungsprozess vertraut zu machen. Während die Anzahl der zu gene-

7 S. https://dariah-de.github.io/TopicsExplorer, zuletzt aufgerufen am 19. 07. 2024.

8 Voyant Tools. URL: http://voyant-tools.org, zuletzt aufgerufen am 18.06.2024 (Rockwell & Sinclair 2016).

rierenden Topics und der Iterationen bei diesen Tools oft noch frei gewählt werden kann, werden komplexere Komponenten wie die sog. Hyperparameter, die das Verteilungsprofil der Topics beeinflussen (Wallach et al. 2009; Boyd-Graber et al. 2014, 233), in der *Black Box* versteckt. Auch die Möglichkeiten zur Evaluation der Modellierungsergebnisse oder ihr Export als nachnutzbare Daten ist hier eingeschränkt. Zu berücksichtigen ist zudem, dass Topic Modeling nicht nur einen einzelnen Algorithmus repräsentiert, sondern, wie wir gesehen haben, eine ganze Familie von Algorithmen, die alle auf die ein oder andere Art und Weise das Ziel verfolgen, Texte anhand ihrer Sprachgebrauchsmuster zu gruppieren, um ihre thematische Struktur zu explorieren.

Es empfiehlt sich daher, auf umfangreichere Anwendungen wie das besonders häufig genutzte Framework *MALLET*[9], den *interactive Leipzig Corpus Miner* (iLCM)[10] oder Implementierungen verschiedener Algorithmen in Programmiersprachen wie *Python* oder *R* zu setzen, die es ermöglichen, die Verfahren entsprechend der eigenen Bedarfe zu konfigurieren. Bibliotheken wie *Gensim*[11], *Scikit-Learn*[12], *Tomotopy*[13], *BERTopic*[14] oder *OCTIS*[15] in Python bieten bspw. gleich mehrere Lösungen in einer Umgebung an.[16] Die Wahl des geeigneten Algorithmus und einer entsprechenden Implementierung stellt dabei nur den ersten Schritt in einem komplexen Prozess dar; wie wir im nächsten Kapitel diskutieren werden, ist etwa die sorgfältige Vorbereitung des Textkorpus für historische und stilistisch variantenreiche Textdaten unerlässlich.

4.2 Preprocessing

Topic Modeling kann grundsätzlich auf beliebige Texte aus beliebigen Sprachräumen angewendet werden, allerdings ist insbesondere für historisch forschende Diszipli-

9 MALLET. A Machine Learning for Language Toolkit. URL: http://mallet.cs.umass.edu, zuletzt aufgerufen am 18.06.2024 (McCallum 2002).

10 iLCM. DFG Funded Project. URL: https://ilcm.informatik.uni-leipzig.de, zuletzt aufgerufen am 18.06.2024 (Niekler et al. 2023).

11 Gensim. Topic modelling for humans. URL: https://radimrehurek.com/gensim, zuletzt aufgerufen am 18.06.2024 (Řehůřek & Sojka 2010).

12 Scikit-learn. Machine Learning in Python. URL: https://scikit-learn.org/stable/index.html, zuletzt aufgerufen am 18.06.2024.

13 Tomotopy. URL: https://bab2min.github.io/tomotopy, zuletzt aufgerufen am 18.06.2024.

14 S. https://maartengr.github.io/BERTopic/index.html, zuletzt aufgerufen am 19. 07. 2024, vgl. Grootendorst 2022.

15 OCTIS. Comparing Topic Models is simple! URL: https://github.com/MIND-Lab/OCTIS, zuletzt aufgerufen am 18.06.2024 (Terragni et al. 2021).

16 Ein Blick in die Dokumentationen der Bibliotheken gibt Aufschluss über die Konfigurationsmöglichkeiten und häufig ersten Beispielcode. In der Regel lassen sich über das Internet zahlreiche nützliche Tutorials recherchieren.

nen zu berücksichtigen, dass die zuvor vorgestellten Verfahren in der Regel anhand von Textdaten entwickelt und getestet wurden, die einer modernen Sprachstufe entsprechen und damit normierter sind als Texte des Mittelalters oder der Frühen Neuzeit; auch literarische Texte mit ihren zahlreichen stilistischen Besonderheiten können hier eine Herausforderung darstellen (siehe etwa Uglanova & Gius 2020). LDA bspw. wurde u. a. anhand von englischsprachigen Nachrichten- sowie wissenschaftlichen Fachartikeln getestet (Blei et al. 2003). Da das Verfahren Regelmäßigkeiten sichtbar macht, ist ein sprachlich und orthographisch möglichst homogenes Korpus zuverlässiger zu modellieren als z. B. Funeralschriften des 17. Jahrhunderts, die noch keinen vergleichbaren schriftsprachlichen Regularien unterlagen und entsprechend verschiedene Schreibweisen für dieselben Konzepte aufweisen können sowie zahlreiche lateinische Bemerkungen und Zitate. Je komplexer und variantenreicher die Datenbasis, desto weniger konsistent und vorhersehbar können die Ergebnisse der Modellierung potenziell sein.

Um die inhaltlichen Charakteristika der Texte herauszufiltern (siehe Tab. 1, nächste Seite), empfiehlt es sich daher, die Komplexität der Textdaten dadurch zu reduzieren, dass das Vokabular vereinheitlicht und normalisiert wird; dieser Arbeitsschritt wird *Preprocessing* genannt (vgl. z. B. Maier et al. 2018, 97 f. 100–102. 110). Art, Umfang und Reihenfolge der einzelnen Verarbeitungsschritte sind dabei nicht trivial und hängen zum einen vom gewünschten Verfahren und zum anderen wesentlich vom zu verarbeitenden Quellentyp und der jeweiligen Fragestellung ab. Zu berücksichtigen ist bei der Auswahl und Anordnung der einzelnen Verarbeitungsschritte die Sprache sowie der Grad der Standardisierung und Normierung. Während es für moderne Sprachen immer mehr Ressourcen gibt, ist das Angebot für historische Sprachstufen noch spärlich, was auf ein aufwändigeres Preprocessing hinauslaufen kann. In jedem Fall empfiehlt sich – ob moderne oder historische Texte –, sehr sorgfältig bei der Vorbereitung der Daten vorzugehen und die einzelnen Entscheidungen in diesem oft iterativen Prozess zu dokumentieren, um die Nachvollziehbarkeit des Vorgehens zu gewährleisten.

Zu den zwingenden Vorbereitungsschritten zahlreicher Textanalysemethoden zählt die *Tokenisierung*. Dabei wird der Text in kleinere Einheiten zerlegt, auch *Token* genannt, die bearbeitet, gezählt, verglichen und neu kombiniert werden können. Üblicherweise wird die Tokenisierung auf Wortebene durchgeführt. Während Menschen intuitiv solche lexikalischen Einheiten erkennen können, muss dies für den Computer explizit formalisiert werden. Je nach Sprache geht dies mit ganz eigenen Herausforderungen einher. Für das Topic Modeling relevant ist bspw. der Umgang mit Mehrworteinheiten, die mal durch Bindestriche kenntlich gemacht werden, mal aber auch nicht, wie z. B. „Heiliger Geist“. Typischerweise würde bei der Tokenisierung die Verbindung zwischen den beiden Wörtern aufgelöst werden („Heiliger“, „Geist“), so dass die einzelnen Wortbestandteile unabhängig voneinander prozessiert werden. Was als Untersuchungseinheit im jeweiligen Forschungskontext gelten soll, ist demnach zu reflektieren. Für solche Aufgaben der natürlichen Sprachverarbei-

Tab. 1 Exemplarische Gegenüberstellung einer Auswahl von Topics vor und nach einem ersten Preprocessing (Funeralschriften des 17. Jahrhunderts, AEDit Frühe Neuzeit)

Vor dem Preprocessing	**Nach dem Preprocessing (Tokenisierung, Entfernung von Satzzeichen und Zahlen, Lemmatisierung, POS-Tagging, Lowercasing)**
und der die das zu mit auch er nicht den dem ist sie von ein wie des sich Gott daß	kind eltern job söhnlein kinderlein kindlein lieb gerecht taufe bräutigam töchterlein braut gerechtigkeit item de christus matt arm justitia himmlisch
Frau Kinder Mutter Adeligen Eltern Kind Adelige Gn. Edlen/E. J. Kinderlein liebes Weib Eltern/Rahel geborene Söhnlein Kindlein Job	frau lieb mutter adelig junker weib adelige witwe edl gn geboren rahel herz schmerz trost schwester gestreng kind kreuz augenlust
Prediger Lehrer & Kirchen Amt M. Zuhörer Stadt ad Anno D. Prediger/c. treuen Gemeine Schulen Fürstlichen Fürstl. treue	prediger kirche lehrer amt jahr zuhörer wort treu prophet groß lehre schule stadt apostel anno ehrwürdig predigt knecht fürstl mann
Dann dann wann wider sonder „ Vers deren lang dieselbige Leibs Tods Kapitel gern Edlen Arzt Sara Sohns seliger dieweil	christus arzt jesus arznei kreuz kapitel doktor luc matth apotheke christi wunde joh apotheker volk hiob medikus leiden heiland jude
Sie Er daß die der Ihr als eine Ihm von zu Frau Die Ich dem den GOTT sich Der Ihre	frau seele hoch himmel welt mutter tod haus freude träne auge herrlichkeit ps leben liebe braut ehre vater land tugend

tung *(Natural Language Processing)* gibt es bereits zahlreiche etablierte Werkzeuge.[17] Auch die Modellierung von *Bi-* (Wortpaaren) und *Trigrammen* (Worttripeln) kann dabei helfen, Phrasen, die sich aus kookkurenten Termen zusammensetzen, wieder zu einem Token zusammenzufügen, wodurch der lokale Bezug teilweise erhalten bleibt.

Für das weitere Preprocessing gibt es keinen vorgeschriebenen Weg; es haben sich allerdings einige Schritte etabliert, die je nach Einsatzzweck modular und auf das Datenkorpus zugeschnitten Verwendung finden können (siehe Abb. 6). Dazu zählt etwa die *Segmentierung*. Insbesondere wenn es sich um Korpora handelt, die sich aus sehr umfangreichen Einzeldokumenten zusammensetzen (z. B. ein Korpus von Büchern), kann es sinnvoll sein, die Dokumente in kleinere Einheiten zu zerlegen (etwa auf Kapitel- oder Absatzebene). Auch die Entfernung von Satzzeichen oder der sog. Stoppwörter hat sich etabliert (siehe kritisch zum Umgang mit Stoppwörtern etwa Schofield et al. 2017). Hierbei handelt es sich um Funktionswörter wie „der“, „und“, „zu“ etc., die, da sie sehr häufig in Texten vorkommen, als statistisch sehr markante Eigenschaften von Texten die Topics dominieren würden (siehe Tab. 1). Auch wenn sie für den menschlichen Umgang mit Sprache von zentraler grammatikalischer Be-

17 In Python bspw. NLTK. Natural Language Toolkit. URL: https://www.nltk.org; spaCy. Industrial-Strength Natural Language Processing. URL: https://spacy.io; beide Adressen wurden zuletzt am 18.06.2024 aufgerufen.

Abb. 6 Die Vorbereitung der Textdaten kann modular aus verschiedenen Schritten zusammengesetzt werden. Sie ist abhängig von den Charakteristika und der Qualität der Daten sowie vom Erkenntnisziel.

deutung sind, gehören sie nicht zwingend zu den bedeutungstragenden Wörtern und könnten daher die Arbeit mit dem Topic-Modell erschweren. Je nach Sprache gibt es verschiedene Stoppwortlisten, die über Programmierbibliotheken wie NLTK nachgenutzt und auch manuell um weitere korpusspezifische Terme erweitert werden können. Es empfiehlt sich stets, die Inhalte dieser Listen zu prüfen, um sicherzugehen, dass für das eigene Vorhaben relevante Wörter nicht unbesehen entfernt werden.

Ein gegenüber Stoppwortlisten systematischeres Vorgehen zur Selektion der relevanten und bedeutsamen Wörter können Methoden wie *TF-IDF* oder *Part-of-Speech-Tagging* (POS-Tagging) bieten. Mit TF-IDF können wir Token identifizieren, die für ein bestimmtes Dokument oder eine Gruppe von Dokumenten charakteristisch sind und demgegenüber Wörter, die besonders häufig in sehr vielen Dokumenten vorkommen (wie Funktionswörter, aber auch andere korpusspezifische Terme), geringer gewichten und entsprechend herausfiltern (Klinke 2017, 274 f.). Beim POS-Tagging wiederum werden automatisiert die Wortarten der lexikalischen Einheiten ermittelt. Auf diese Weise können spezifische Wortarten für die Analyse ausgewählt werden, von denen auszugehen ist, dass sie eine bedeutungstragende Funktion in Texten übernehmen (wie bspw. Nomen, Verben oder Adjektive; siehe z. B. Schöch 2017, 17).

Um die Variationen in den Wortformen zu minimieren und dadurch die Modellierung kohärenterer Topics zu ermöglichen, aber auch die Datenverarbeitung zu erleichtern, hat sich neben dem *Lowercasing* (Kleinschreibung aller Token) insbesondere die *Lemmatisierung,* also die Reduktion der flektierten Einzelwortformen auf ihre Grundform (heiliger → heilig, gingen → gehen), als effizient erwiesen. Insbesondere in englischsprachigen Kontexten ist auch das *Stemming* nicht unüblich; hierbei werden die einzelnen Wörter auf ihren Stamm oder die Wurzel reduziert, indem etwa die Wortendungen abgeschnitten werden (z. B. christlich, Christus → christ, Christentum → christent). Das Ergebnis hiervon sind Token, die nicht notwendigerweise einen gültigen lexikalischen Eintrag in einer Sprache widerspiegeln und daher auch deutlich schwerer zu interpretieren sein können (Schofield & Mimno 2016). Geht es allein um die schnelle Erschließung eines Korpus, mag das letztgenannte Vorgehen sinnvoll sein, für auf Interpretation zielende Forschungsprojekte ist die Lemmatisierung aber die zu bevorzugende Variante.

Jeder einzelne dieser und weiterer Verarbeitungsschritte wirkt sich unmittelbar auf das jeweilige Modellierungsergebnis aus und damit auf das, was wir zu interpretieren gedenken.[18] Das Vorgehen sollte daher nicht nur dokumentiert, sondern auch unter Berücksichtigung der Fragestellung und Erkenntnisziele in die Evaluation der Topic-Modelle integriert werden.

4.3 Evaluation

Die Evaluation von Topic-Modellen ist unerlässlich, um die Qualität und Relevanz der generierten Cluster sicherzustellen. Dies ist umso ratsamer, als bei solchen Verfahren die Gefahr besteht, einem *confirmation bias* zu erliegen, also den Modellierungsprozess und die zu analysierenden Daten so lange zu bearbeiten bis ein gewünschtes oder erwartetes Ergebnis erzielt wird (Shadrova 2021, 5. 16 f.). Auch wenn es im Allgemeinen bei Topic Modeling, wie Maria Antoniak hervorgehoben hat, nicht darum geht, die eine „richtige" Sichtweise auf das Textkorpus zu generieren, sondern darum, eine qualitative Untersuchung zu unterstützen, indem eine von vielen möglichen „interpretative lenses" entdeckt wird, durch die unsere Quellen betrachtet werden können (Antoniak 2022), empfiehlt es sich, zusätzlich zur qualitativen Prüfung der Topics hinsichtlich ihrer Interpretierbarkeit und Repräsentativität, auch einige mathematische Evaluationsmetriken hinzuzuziehen (einen guten Einstieg bieten: Boyd-Graber et al. 2014, 233 f., 237–243; Churchill & Singh 2022, 5–9).

Hinsichtlich der qualitativen Bewertung bietet sich *Blended Reading* als Evaluationsmodus an (Stulpe & Lemke 2016). Hierbei kombinieren wir die Ergebnisse des

18 Neuere Verfahren wie *BERTopic* versprechen dagegen, durch die Nutzung neuester Sprachmodelle auf das Preprocessing weitestgehend verzichten zu können. Hier muss sich allerdings erst noch zeigen, wie gut dies für historische Sprachstufen funktioniert.

Machine-Learning-Prozesses *(Distant Reading)* mit der menschlichen Lektüre und Interpretation *(Close Reading).* Durch das Lesen repräsentativer Dokumente oder Textpassagen für jedes Topic können wir ihre Interpretierbarkeit und Repräsentativität bewerten oder durch das Vergleichen der wichtigsten Wörter und Phrasen, die den Topics zugeordnet sind, die Granularität des Modells.[19] Sinnvoll kann es auch sein, zu prüfen, ob es eine *Ground Truth* für das Textkorpus gibt oder es handhabbar ist, eine zu erstellen. Gemeint ist damit bspw. eine bereits vorhandene manuelle Klassifikation, mit der das Modell verglichen werden kann.[20] In diesem Fall können Metriken wie Genauigkeit *(accuracy),* Präzision *(precision), Recall* und *F-Score* verwendet werden, um die Leistung des Modells zu bewerten (Churchill & Singh 2022, 5–9; Klinke 2017, 269 f.).

Da beim Einsatz von Topic Modeling in der Regel keine derartige *Ground Truth* verfügbar ist, sind einige weitere Metriken entwickelt worden, die für die Evaluation der Modellierungsergebnisse herangezogen werden können (Churchill & Singh 2022, 6–8; Boyd-Graber et al. 2014, 233f., 237–243):

- Über Kohärenzmaße *(coherence)* kann bspw. gemessen werden, wie gut die (Top-)Wörter, die den Topics zugewiesen wurden, zusammenpassen. Je höher der Kohärenzwert, desto semantisch kohärenter sind die Topics theoretisch, was gleichermaßen mit einer besseren Interpretierbarkeit der Wortcluster einhergehen sollte. Tools wie *Gensim* in Python bieten Funktionen zur Berechnung der Kohärenz, die auch genutzt werden können, um zu ermitteln, welche Topic-Anzahl sinnvoll für einen gegebenen Forschungsgegenstand ist.[21]
- Mit der Perplexität *(perplexity)* wiederum kann eingeschätzt werden, wie gut das Topic-Modell neue, nicht gesehene Dokumente vorhersagen kann. Ein niedrigerer Perplexitätswert ist in der Regel besser, aber dieser Wert allein ist oft nicht ausreichend, um die Qualität eines Modells zu beurteilen.
- Auch kann die Exklusivität bzw. Einzigartigkeit der den Topics zugewiesenen (Top-)Wörter für die jeweiligen Topics berechnet werden, um die Unterscheidbarkeit der Wortcluster zu beurteilen.

Dies sind nur einige der verfügbaren Möglichkeiten, um Topic-Modelle zu evaluieren. Da diese Evaluationsmetriken nicht immer positiv mit den menschlichen Beurtei-

19 Insbesondere die Topic-Anzahl wirkt sich auf die Granularität des Modells aus. Eine zu hohe Zahl führt potenziell zu sich überschneidenden, redundanten Clustern, während eine zu niedrige zu heterogene hervorbringt (vgl. auch Schöch 2017, 20, Anm. 7).

20 Bei H-Soz-Kult gibt es bspw. eine manuelle Klassifikation nach Themen, Regionen und Epochen, die eine gute Orientierung für die Evaluation der Modellierungsergebnisse bietet.

21 Mit *Hierarchical Dirichlet Process* (HDP), einer Erweiterung von LDA, wurde ein Verfahren entwickelt, das es ermöglicht, die Anzahl der zu modellierenden Cluster aus den Korpusdaten abzuleiten (Teh et al. 2006).

lungen der Modellierungsergebnisse korrelieren (z. B. Hoyle et al. 2021; Uglanova & Gius 2020), sollten sie stets ergänzend zur qualitativ-manuellen Einordnung durch die Forscher*innen mit ihrem Domänenwissen und unter Berücksichtigung der spezifischen Forschungsfrage sowie der Eigenheiten des Textkorpus eingesetzt werden.

5. Schlussbemerkung

Abschließend lässt sich konstatieren, dass Topic Modeling eine lohnenswerte Ergänzung für die Methodenlandschaft der Theologie darstellen kann. Als „statistische Linse", die das Wissen der Theologie, ihre Theorien und Annahmen formalisiert (nach Blei 2012a, 8), kann sie neue datengetriebene Perspektiven auf die Quellen und Forschungsdebatten eröffnen. Zwar liefern Topic-Modelle nicht an sich Schlussfolgerungen auf konkrete Fragestellungen, sie können aber erfolgreich eingesetzt werden, um Hypothesen zu explorieren und letztlich wieder an den einzelnen Quellen zu überprüfen. Damit werden texthermeneutische Ansätze nicht ersetzt, sondern das Studium der Quellen vielmehr um ein weiteres Instrumentarium erweitert. Mit Topic Modeling wird somit eine Brücke zwischen traditionellen hermeneutischen Ansätzen und modernen datenbasierten Methoden geschlagen, womit sie das analytische Repertoire der Geisteswissenschaften bereichert und neue Wege für die systematische und kritische Auseinandersetzung mit umfangreichen Textkorpora eröffnet. Es bleibt zu hoffen, dass dieser Ansatz als Ausgangspunkt für weiterführende Untersuchungen und Diskussionen innerhalb der Theologie dienen wird.

Literaturverzeichnis

Althage, M. (2022). Potenziale und Grenzen der Topic-Modellierung mit Latent Dirichlet Allocation für die Digital History. In K. D. Döring, S. Haas, M. König & J. Wettlaufer (Hrsg.), *Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft* (S. 255–277). Berlin/Boston: De Gruyter Oldenbourg [= *Studies in Digital History and Hermeneutics,* 6]. https://doi.org/10.1515/9783110757101-014 [zuletzt aufgerufen am 18.06.2024].

Dies. (2023). Digitale Methoden kritisch reflektieren. Die Erweiterung des Werkzeugkastens der Historiker:innen. In P. Trilcke, A. Busch & P. Helling (Hrsg.), *DHd 2023. Open Humanities Open Culture. 9. Tagung des Verbands Digital Humanities im deutschsprachigen Raum.* Trier/Luxemburg: Zenodo. https://doi.org/10.5281/zenodo.7711522 [zuletzt aufgerufen am 18.06.2024].

Antoniak, M. (2022). Topic Modeling for the People. In *Blog von M. Antoniak*. URL: https://maria-antoniak.github.io//2022/07/27/topic-modeling-for-the-people.html [zuletzt aufgerufen am 18.06.2024].

Blei, D. M. (2012a). Topic Modeling and Digital Humanities, *Journal of Digital Humanities*, 2(1), 8–11. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei [zuletzt aufgerufen am 18.06.2024].

Ders. (2012b). Probabilistic topic models, *Communications of the ACM* 55(4), 77–84. https://doi.org/10.1145/2133806.2133826 [zuletzt aufgerufen am 18.06.2024].

Ders., & Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning – ICML '06* (S. 113–120). New York: ACM Press. https://doi.org/10.1145/1143844.1143859 [zuletzt aufgerufen am 18.06.2024].

Dies. (2007). A correlated topic model of Science, *The Annals of Applied Statistics*, 1(1), 17–35. https://doi.org/10.1214/07-AOAS114 [zuletzt aufgerufen am 18.06.2024].

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937 [zuletzt aufgerufen am: 21. 07. 2024].

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and Feeding of Topic Models. Problems, Diagnostics, and Improvements. In E. M. Airoldi, D. M. Blei, E. A. Erosheva, & S. E. Fienberg (Hrsg.), *Handbook of Mixed Membership Models and Their Applications* (S. 225–254). London/New York: Chapman and Hall/CRC. https://doi.org/10.1201/b17520-21 [zuletzt aufgerufen am 18.06.2024].

Brett, M. R. (2012). Topic Modeling. A Basic Introduction, *Journal of Digital Humanities*, 2(1), 12–16. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett [zuletzt aufgerufen am 18.06.2024].

Bunout, E., & von Lange, M. (2019). Nibbling at Text. Identifying Discourses on Europe in a Large Collection of Historical Newspapers Using Topic Modelling. In *C2DH | Luxembourg Centre for Contemporary and Digital History*. URL: https://www.c2dh.uni.lu/thinkering/nibbling-text-identifying-discourses-europe-large-collection-historical-newspapers-using [zuletzt aufgerufen am 18.06.2024].

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM. Topic Modeling over Short Texts, *IEEE Transactions on Knowledge and Data Engineering*, 26(12), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872 [zuletzt aufgerufen am 18.06.2024].

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling, *ACM Computing Surveys*, 54(10s), 1–35. https://doi.org/10.1145/3507900 [zuletzt aufgerufen am 18.06.2024].

Crane, G. (2006). What Do You Do with a Million Books?, *D-Lib Magazine*, 12(3), o. S. https://doi.org/10.1045/march2006-crane [zuletzt aufgerufen am 18.06.2024].

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces, *Transactions of the Association for Computational Linguistics*, 8, 439–453. https://doi.org/10.1162/tacl_a_00325 [zuletzt aufgerufen am 21. 07. 2024].

Graham, Sh., Milligan, I., & Weingart, S. (2016). *Exploring Big Historical Data. The Historian's Macroscope.* London: Imperial College Press.

Grant, Ph., Sebastian, R., Allassonnière-Tang, M., & Cosemans, S. (2021). Topic Modelling on Archive Documents from the 1970s. Global Policies on Refugees, *Digital Scholarship in the Humanities,* 36(4), 886–904. https://doi.org/10.1093/llc/fqab018 [zuletzt aufgerufen am 18.06.2024].

Graves, M. (2022). Computational Topic Models for Theological Investigations, *Theology and Science,* 20(1), 69–84. https://doi.org/10.1080/14746700.2021.2012922 [zuletzt aufgerufen am 18.06.2024].

Grootendorst, M. (2022). BERTopic. Neural topic modeling with a class-based TF-IDF procedure. Online: *arXiv.* https://doi.org/10.48550/arXiv.2203.05794 [zuletzt aufgerufen am 18.06.2024].

Guldi, J. (2019). Parliament's Debates about Infrastructure. An Exercise in Using Dynamic Topic Models to Synthesize Historical Change, *Technology and Culture,* 60(1), 1–33. https://doi.org/10.1353/tech.2019.0000 [zuletzt aufgerufen am 18.06.2024].

Hodel, T., Möbus, D., & Serif, I. (2022). Von Inferenzen und Differenzen. Ein Vergleich von Topic-Modeling-Engines auf Grundlage historischer Korpora. In S. Gerlek, S. Kissler, Th. Mämecke, & D. Möbus (Hrsg.), *Von Menschen und Maschinen. Mensch-Maschine-Interaktionen in digitalen Kulturen* (S. 185–209). Hagen: Hagen University Press. https://doi.org/10.57813/20220623-153139-0 [zuletzt aufgerufen am 21.07.2024].

Horstmann, J. (2018). Topic Modeling, *ForText. Literatur Digital Erforschen,* 1–16. URL: https://fortext.net/routinen/methoden/topic-modeling [zuletzt aufgerufen am 18.06.2024].

Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., & Resnik, Ph. (2021). Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence, *Advances in Neural Information Processing Systems,* 34 (S. 2018–2033). URL: https://proceedings.neurips.cc/paper_files/paper/2021/hash/0f83556a305d789b1d71815e8ea4f4b0-Abstract.html [zuletzt aufgerufen am 18.06.2024].

Jelodar, H., Wang, Y., Yuan, Ch., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic Modeling. Models, Applications, a Survey, *Multimedia Tools and Applications,* 78(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4 [zuletzt aufgerufen am 18.06.2024].

Jockers, M. L. (2013). *Macroanalysis. Digital Methods and Literary History.* Urbana/Chicago/Springfield: University of Illinois Press.

Klinke, H. (2017). Information Retrieval. In F. Jannidis, H. Kohle, & M. Rehbein (Hrsg.), *Digital Humanities. Eine Einführung* (S. 268–278). Stuttgart: J. B. Metzler. https://doi.org/10.1007/978-3-476-05446-3_19 [zuletzt aufgerufen am 18.06.2024].

Küsters, A., & Garrido, E. (2020). Mining PIGS. A Structural Topic Model Analysis of Southern Europe Based on the German Newspaper Die Zeit (1946–2009),

Journal of Contemporary European Studies, 28(4), 477–493. https://doi.org/10.1080/14782804.2020.1784112 [zuletzt aufgerufen am 18.06.2024].

Lafferty, J. D., & Blei, D. M. (2005). Correlated Topic Models, *Advances in Neural Information Processing Systems,* 18, (S. 147–154). URL: https://proceedings.neurips.cc/paper_files/paper/2005/hash/9e82757e9a1c12cb710ad680db11f6f1-Abstract.html [zuletzt aufgerufen am 18.06.2024].

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization, *Nature,* 401, 788–791. https://doi.org/10.1038/44565 [zuletzt aufgerufen am 21. 07. 2024].

Li, W., & McCallum, A. (2006). Pachinko Allocation. DAG-Structured Mixture Models of Topic Correlations. In *Proceedings of the 23rd International Conference on Machine Learning* (S. 577–584). New York: Association for Computing Machinery. https://doi.org/10.1145/1143844.1143917 [zuletzt aufgerufen am 21. 07. 2024].

Luhmann, J., & Burghardt, M. (2021). Digital Humanities. A Discipline in Its Own Right? An Analysis of the Role and Position of Digital Humanities in the Academic Landscape, *Journal of the Association for Information Science and Technology,* 73(2), 148–171. https://doi.org/10.1002/asi.24533 [zuletzt aufgerufen am 18.06.2024].

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research. Toward a Valid and Reliable Methodology, *Communication Methods and Measures,* 12(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754 [zuletzt aufgerufen am 18.06.2024].

McCallum, A. K. (2002). MALLET. A Machine Learning for Language Toolkit. URL: http://mallet.cs.umass.edu [zuletzt aufgerufen am 18.06.2024].

Mimno, D. (2012). Computational Historiography. Data Mining in a Century of Classics Journals, *Journal on Computing and Cultural Heritage,* 5(1), 1–19. https://doi.org/10.1145/2160165.2160168 [zuletzt aufgerufen am 18.06.2024].

Ders., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (S. 880–889). Singapore: Association for Computational Linguistics. URL: https://aclanthology.org/D09-1092 [zuletzt aufgerufen am 21. 07. 2024].

Niekler, A., Kahmann, Ch., Burghardt, M., & Heyer, G. (2023). The Interactive Leipzig Corpus Miner. An Extensible and Adaptable Text Analysis Tool for Content Analysis. *Publizistik,* 68, 325–354. https://doi.org/10.1007/s11616-023-00809-4 [zuletzt aufgerufen am 18.06.2024].

Nunn, Ch. (2022). Das Thema patristischer Ethik. Versuch einer Annäherung durch Distanz am Beispiel der Briefe des Augustinus von Hippo, *Journal of Ethics in Antiquity and Christianity,* 4, 31–51. https://doi.org/10.25784/jeac.v4i0.1011 [zuletzt aufgerufen am 18.06.2024].

Piper, A. (2018). *Enumerations. Data and Literary Study.* Chicago/London: The University of Chicago Press.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (S. 46–50). Valletta, Malta: ELRA.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses, *American Journal of Political Science,* 58(4), 1064–1082. https://doi.org/10.1111/ajps.12103 [zuletzt aufgerufen am 18.06.2024].

Rockwell, G., & Sinclair, S. (2016). *Hermeneutica. Computer-Assisted Interpretation in the Humanities.* Cambridge, MA/London: The MIT Press.

Rosen-Zvi, M., Griffiths, Th., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. In *UAI '04. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (S. 487–494). Arlington: AUAI Press. https://dl.acm.org/doi/10.5555/1036843.1036902 [zuletzt aufgerufen am 18.06.2024].

Schmidt, B. M. (2012). Words Alone. Dismantling Topic Models in the Humanities, *Journal of Digital Humanities,* 2(1), 49–65. URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt [zuletzt aufgerufen am 18.06.2024].

Schöch, Ch. (2017). Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama, *Digital Humanities Quarterly,* 11(2), 1–53. URL: https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [zuletzt aufgerufen am 19. 07. 2024]

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops. Rethinking Stopword Removal for Topic Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2. Short Papers* (S. 432–436). URL: https://www.aclweb.org/anthology/E17-2069 [zuletzt aufgerufen am 18.06.2024].

Schofield, A., & Mimno, D. (2016). Comparing Apples to Apple. The Effects of Stemmers on Topic Models, *Transactions of the Association for Computational Linguistics,* 4, 287–300. https://doi.org/10.1162/tacl_a_00099 [zuletzt aufgerufen am 18.06.2024].

Schwandt, S. (2018). Digitale Methoden für die Historische Semantik. Auf den Spuren von Begriffen in digitalen Korpora, *Geschichte und Gesellschaft. Zeitschrift für historische Sozialwissenschaft,* 44(1), 107–134. https://doi.org/10.13109/gege.2018.44.1.107 [zuletzt aufgerufen am 18.06.2024].

Shadrova, A. (2021). Topic Models Do Not Model Topics. Epistemological Remarks and Steps towards Best Practices, *Journal of Data Mining & Digital Humanities,* 1–28. https://doi.org/10.46298/jdmdh.7595 [zuletzt aufgerufen am 18.06.2024].

Simmler, S., Vitt, Th., & Pielström, S. (2019). Topic Modeling with Interactive Visualizations in a GUI Tool. In *Proceedings of the Digital Humanities Conference.*

Utrecht. Tool: https://dariah-de.github.io/TopicsExplorer [zuletzt aufgerufen am 18.06.2024].

Snickars, P. (2022). Modeling Media History. On Topic Models of Swedish Media Politics 1945–1989, *Media History,* 28(3), 403–424. https://doi.org/10.1080/13688804.2022.2079484 [zuletzt aufgerufen am 18.06.2024].

Stulpe, A., & Lemke, M. (2016). Blended Reading. Theoretische und praktische Dimensionen der Analyse von Text und sozialer Wirklichkeit im Zeitalter der Digitalisierung. In M. Lemke & G. Wiedemann (Hrsg.), *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse* (S. 17–61). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-07224-7_2 [zuletzt aufgerufen am 21. 07. 2024].

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes, *Journal of the American Statistical Association,* 101(476), 1566–1581. https://doi.org/10.1198/016214506000000302 [zuletzt aufgerufen am 18.06.2024].

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). OCTIS. Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations* (S. 263–270). https://doi.org/10.18653/v1/2021.eacl-demos.31 [zuletzt aufgerufen am 18.06.2024].

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning. Vector Space Models of Semantics, *Journal of Artificial Intelligence Research,* 37(1), 141–188. URL: https://jair.org/index.php/jair/article/view/10640 [zuletzt aufgerufen am 18.06.2024].

Uglanova, I., & Gius, E. (2020). The Order of Things. A Study on Topic Modelling of Literary Texts. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (S. 57–76). URL: https://ceur-ws.org/Vol-2723/long7.pdf [zuletzt aufgerufen am 18.06.2024].

Vayansky, I., & Kumar, S. A. P. (2020). A Review of Topic Modeling Methods, *Information Systems,* 94(101582), 1–15. https://doi.org/10.1016/j.is.2020.101582 [zuletzt aufgerufen am 18.06.2024].

Völkl, Y., Sarić, S., & Scholger, M. (2022). Topic Modeling for the Identification of Gender-Specific Discourse. Virtues and Vices in French and Spanish 18th Century Periodicals, *Journal of Computational Literary Studies,* 1(1), 1–27. https://doi.org/10.48694/jcls.108 [zuletzt aufgerufen am 18.06.2024].

Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA. Why priors matter. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (S. 1973–1981). New York: Curran Associates Inc. https://dl.acm.org/doi/10.5555/2984093.2984314 [zuletzt aufgerufen am 21. 07. 2024].

Wang, X., & McCallum, A. (2006). Topics over Time. A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (S. 424–433). New York: Association for Computing Machinery. https://doi.org/10.1145/1150402.1150450 [zuletzt aufgerufen am 21. 07. 2024].

Wehrheim, L. (2019). Economic History Goes Digital. Topic Modeling the Journal of Economic History, *Cliometrica,* 13(1), 83–125. https://doi.org/10.1007/s11698-018-0171-7 [zuletzt aufgerufen am 18.06.2024].

Ders., Spoerer, M., & Jopp, T. A. (2022). Turn, Turn, Turn. A Digital History of German Historiography, 1950–2019, *The Journal of Interdisciplinary History,* 53(3), 471–507. https://doi.org/10.1162/jinh_a_01871 [zuletzt aufgerufen am 18.06.2024].

Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, Ch., & Zhuang, F. (2021). A Neural Topic Model with Word Vectors and Entity Vectors for Short Texts, *Information Processing & Management,* 58(2), 1–11. https://doi.org/10.1016/j.ipm.2020.102455 [zuletzt aufgerufen am 18.06.2024].

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic Modeling of Short Texts. A Pseudo-Document View. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (S. 2105–2114). New York: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939880 [zuletzt aufgerufen am 18. 06. 2024].

Bildnachweise

Alle Abbildungen wurden von der Autorin selbst erstellt. Die Abb. 1–3 wurden mit dem Gensim-Wrapper für *MALLET* generiert und mit der Python-Bibliothek *WordCloud*[22] visualisiert. Die Abb. 4–6 wurden mit *Miro*[23] erstellt. Die Abb. 5 f. basieren auf Althage (2023). Alle übrigen Abbildungen wurden hier erstveröffentlicht.

22 S. https://amueller.github.io/word_cloud, zuletzt aufgerufen am 18. 06. 2024.

23 S. https://miro.com/de, zuletzt aufgerufen am 18. 06. 2024.