HOCHSCHULE
MITTWEIDA
University of Applied Sciences

MASTERARBEIT

Frau
Jenny Maria Felser, B.Sc.

Themenmodellierung in forensischen Kommunikationsdaten

Mittweida, Oktober 2023

Fakultät **Angewandte Computer- und Biowissenschaften**

MASTERARBEIT

Themenmodellierung in forensischen Kommunikationsdaten

Autorin:
Jenny Maria Felser

Studiengang:
Cybercrime/Cybersecurity

Seminargruppe:
CY21wC-M

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
Prof. Dr. rer. nat. Michael Spranger

Einreichung:
Mittweida, 20.10.2023

Verteidigung/Bewertung:
Mittweida, 2023

Faculty of **Applied Computer Sciences and Biosciences**

MASTER THESIS

Topic Modelling for Forensic Communication Data

Author:
Jenny Maria Felser

Course of Study:
Cybercrime/Cybersecurity

Seminar Group:
CY21wC-M

First Examiner:
Prof. Dr. rer. nat. Dirk Labudde

Second Examiner:
Prof. Dr. rer. nat. Michael Spranger

Submission:
Mittweida, 20.10.2023

Defense/Evaluation:
Mittweida, 2023

Bibliografische Beschreibung:

Felser, Jenny Maria:
Themenmodellierung in forensischen Kommunikationsdaten. – 2023. – 110 S.
Mittweida, Hochschule Mittweida – University of Applied Sciences, Fakultät Angewandte Computer- und Biowissenschaften, Masterarbeit, 2023.

Referat:

Die Auswertung von Kurznachrichten, die auf mobilen Endgeräten gespeichert sind, nimmt bei strafrechtlichen Ermittlungen immer mehr an Bedeutung zu. Häufig sind Ermittler hierbei mit umfassenden Nachrichtenmengen konfrontiert. Um einen Überblick zu erhalten, wäre eine kompakte Zusammenfassung der zahlreichen Nachrichten hilfreich. Eine Möglichkeit diese automatisiert zu erhalten, stellt die Themenmodellierung dar. Diese ist allerdings bei forensischen Kommunikationsdaten mit besonderen Herausforderungen verbunden. Zu diesen zählt die Tatsache, dass der Ermittler oft eine Erwartungshaltung an die Themen hat, wobei die für ihn interessanten Themen häufig nur zu einem geringen Anteil in den Daten vertreten sind. Um ihn bei dem Finden von Beweisen zu diesen Themen zu unterstützen, wurden zwei Methoden der halbüberwachten Themenmodellierung und Erweiterungen basierend auf Word Embeddings und paradigmatischen Relationen miteinander verglichen. Insbesondere für umgangssprachliche Kurznachrichten ist die Evaluierung der Themenmodellierung als schwierig anzusehen, da bisherige Studien gezeigt haben, dass gängige quantitative Evaluierungsmaße bei diesen nicht unbedingt die tatsächliche Interpretierbarkeit der Themen widerspiegeln. Daher bestand ein weiteres Ziel der Arbeit darin zu untersuchen, inwieweit die Ergebnisse einer regelmäßig angewendeten automatischen Evaluierungsmethode durch eine Nutzerstudie wiedergegeben werden. Insgesamt konnte festgestellt werden, dass nach der quantitativen Evaluierung die halbüberwachte Themenmodellierung unter Einbeziehung von paradigmatischen Relationen als besonders erfolgversprechend angesehen werden kann, während nach der Nutzerstudie vor allem die Word Embeddings die Ergebnisse der halbüberwachten Themenmodellierung verbessern konnten. Des Weiteren zeigte sich, dass keine Korrelation zwischen den Resultaten der automatischen Evaluierung und der Nutzerstudie vorlag.

Inhaltsverzeichnis

Inhaltsverzeichnis I

Abbildungs- und Tabellenverzeichnis III

1 Einleitung 1

2 Grundlagen und Stand der Forschung 4
2.1 Methodik zur Literaturrecherche . . . 4
2.2 Taxonomie von Themenmodellen . . . 5
2.3 Themenmodellierung in der Forensik . . . 14
2.3.1 Einsatz in der Forensik und im Kontext von Kriminalität . . . 14
2.3.2 Herausforderungen in der Forensik . . . 16
2.4 Lösungsansätze für kurze Texte . . . 19
2.5 Lösungsansätze für verrauschte Texte . . . 29
2.6 Lösungsansätze für die Bedeutung des Kontextes und der Metadaten . . . 31
2.7 Lösungsansätze für die Erwartungshaltung des Ermittlers an die Themen . . . 33
2.8 Lösungsansätze für lückenhafte Kontexte . . . 39
2.9 Lösungsansätze für die hohe Variabilität im Vokabular . . . 42
2.10 Lösungsansätze für mehrsprachige Texte . . . 44
2.11 Evaluierung . . . 47
2.12 Zusammenfassung . . . 52

3 Daten und Methoden 55
3.1 Verwendete Daten . . . 56
3.2 Vorverarbeitung und Aufbereitung der Daten . . . 57
3.2.1 Vorverarbeitung der Daten . . . 57
3.2.2 Aggregation zu Pseudodokumenten . . . 58
3.3 Umsetzung des Seed Guided Topic Modellings . . . 59
3.3.1 Erstellung der Seed-Wortmengen . . . 60
3.3.2 Bestimmung der Themenanzahl . . . 62
3.3.2.1 Semantische Kohärenz . . . 63
3.3.2.2 Frequency-Exclusivity (FREX) . . . 64
3.3.2.3 Kompromiss aus semantischer Kohärenz und FREX . . . 65
3.3.3 Verwendete Algorithmen . . . 66
3.3.3.1 keyword-Assisted Topic Model (keyATM) . . . 67
3.3.3.2 Seeded Latent Dirichlet Allocation (Seeded LDA) . . . 69
3.4 Erweiterung basierend auf CluWords . . . 69
3.4.1 Bestimmung der CluWords unter Verwendung von fastText . . . 72
3.4.2 Bestimmung der CluWords basierend auf paradigmatischen Relationen . 73
3.5 Evaluierung . . . 75
3.5.1 Qualitative und automatische, quantitative Evaluierung . . . 75
3.5.2 Durchführung und Auswertung der Nutzerstudie . . . 75
3.5.3 Vergleich der automatischen Evaluierung und der Nutzerstudie . . . 78

4 Ergebnisse und Diskussion **80**
4.1 Ermittelte optimale Themenanzahl . . . 80
4.2 Qualitative Evaluierung . . . 82
4.2.1 keyATM . . . 82
4.2.2 Seeded LDA . . . 85
4.2.3 Auswirkungen der Cluster of Words (CluWords) auf die Ergebnisse . . . 87
4.2.3.1 CluWords basierend auf Word Embedding Ähnlichkeit . . . 88
4.2.3.2 CluWords basierend auf paradigmatischen Relationen . . . 91
4.2.4 Teilzusammenfassung . . . 94
4.3 Quantitative, automatische Evaluierung . . . 95
4.4 Evaluierung durch die Nutzerstudie . . . 97
4.5 Vergleich zwischen quantitativer Evaluierung und Nutzerstudie . . . 99

5 Zusammenfassung **107**

6 Ausblick **109**

Literaturverzeichnis **111**

Eidesstattliche Erklärung **144**

Abbildungs- und Tabellenverzeichnis

Abbildungsverzeichnis

2.1 Taxonomie von Themenmodellen. 6
2.2 Übersicht über verschiedene Strategien zum Umgang mit kurzen Texten bei der Themenmodellierung 21

3.1 Vorgehensweise zum Vergleich der verschiedenen Ansätze des Seed-Guided Topic Modellings auf forensischen Kommunikationsdaten. 56
3.2 Aufbau des Termbaums als Grundlage zur Erstellung der Seed-Wortmengen. . 60
3.3 Vorgehensweise zur Ermittlung einer geeigneten Anzahl an „Unseeded Topics“. 62
3.4 Skizze des generativen Prozesses von keyATM. 67
3.5 Vorgehensweise zur Anreicherung von Dokumenten mit den CluWords der Topic Labels. 71
3.6 Beispiel für eine Aufgabe des Word Intrusion Tests zur Evaluierung der trainierten Themenmodelle. 77

4.1 Semantische Kohärenz und FREX der Themenmodelle unter Verwendung der Standard-Latent Dirichlet Allocation (LDA) mit verschiedenen Themenanzahlen. 81

Tabellenverzeichnis

2.1 Vor- und Nachteile von ausgewählten algebraischen und probabilistischen Algorithmen der Themenmodellierung. 7
2.2 Vergleich der verschiedenen Unterkategorien von Deep Learning Topic Models (DLTMs). 9
2.3 Vergleich von Embedding-basierten Ansätzen zur Themenmodellierung. 11
2.4 Überblick über die Vor- und Nachteile der graphenbasierten Themenmodelle, der Ansätze des Fuzzy Clusterings, des Exemplar-based Topic Modelling und des Algorithmus Correlation Explanation (CorEx). 13
2.5 Überblick über verwendete Kontextarten bei der Themenmodellierung. 32
2.6 Überblick über Formen des Nutzerinputs bei verschiedenen Ansätzen der halbüberwachten Themenmodellierung. 35
2.7 Ansätze von Lifelong Topic Modelling (LTM) bezüglich der Form und Extraktion des Wissens sowie des Algorithmus zur Themenmodellierung. 40
2.8 Überblick über automatische Evaluierungsmethoden für die Themenmodellierung. 48
2.9 Möglichkeiten zur Evaluierung unter Einbeziehung menschlicher Annotatoren. 52

3.1 Statistische Beschreibung des verwendeten Datensatzes. 57
3.2 Eigenschaften der gebildeten Konversationsdokumente als Eingabe für die Themenmodellierung 59

3.3 Übersicht über die erstellten Seed Wortmengen für die halbüberwachte Themenmodellierung mit keyATM und Seeded LDA. 61
3.4 Verwendete Werte für die Hyperparameter für das Training eines unüberwachten fastText Modells. 72
3.5 Übersicht über die gebildeten CluWords zu den Topic Labels der Seed Wortmengen basierend auf fastText Embedding Ähnlichkeit. 73
3.6 Übersicht über die gebildeten CluWords zu den Topic Labels der Seed Wortmengen basierend auf paradigmatischen Relationen. 75

4.1 Erläuterung der Abkürzungen für die Bezeichnungen der einzelnen Themenmodelle des Seed-Guided Topic Modellings. 80
4.2 Zehn wahrscheinlichste Wörter von ausgewählten „Seeded Topics“ des Modells „Ka-8“ (keyATM bei Einsatz von Standard Collapsed Gibbs Sampling). 84
4.3 Zehn wahrscheinlichste Wörter von ausgewählten „Seeded Topics“ des Modells „KaT“ (keyATM unter Einbeziehung eines Termgewichtungsschemas). 85
4.4 Zehn Wörter mit der höchsten Wahrscheinlichkeit in ausgewählten „Seeded Topics“ des Modells „S-8“, der Seeded LDA. 86
4.5 Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „KaCF-8“ (keyATM unter Einbeziehung von CluWords basierend auf Word Embedding Ähnlichkeit). 89
4.6 Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „„SCF-8“ (Seeded LDA unter Einbeziehung von CluWord basierend auf Word Embedding Ähnlichkeit). 90
4.7 Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „KaCP-8“ (keyATM unter Berücksichtigung von CluWord basierend auf paradigmatischen Relationen). 93
4.8 Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „SCP-8“ (Seeded LDA unter Einbeziehung von CluWord basierend auf paradigmatischen Relationen). 94
4.9 Mittlere semantische Kohärenz aller Themen sowie der „Seeded Topics“ der untersuchten Modelle. 96
4.10 Mean Model Precision der verschiedenen untersuchten Themenmodelle, gemessen mit dem Word Intrusion Test. 98
4.11 Beispiele für Themen mit einer hohen semantischen Kohärenz und geringen Ergebnissen beim Word Intrusion Test. 103
4.12 Beispiele für Themen mit guten Ergebnissen beim Word Intrusion Test und einer geringen semantischen Kohärenz. 105

Abkürzungsverzeichnis

APSUM . . . . . Aspect Summarization Model

AR . . . . . autoregressive

ARI . . . . . Adjusted Rand Index

ART . . . . . Author-Recipient-Topic Model

ASTM . . . . . Attention Segmentation based TM

ATM . . . . . Author Topic Model

BERT . . . . . Bidirectional Encoder Representations from Transformers

BiTTM . . . . . BiTerms-based Topic Model

BTM . . . . . Biterm Topic Modelling

CatE . . . . . Category-Name Guided Text Embedding

ccLDA . . . . . cross-collection LDA

CEW-LDA . . . . . Combined Entropy Weighting-Latent Dirichlet Allocation

CLD2 . . . . . Compact Language Detector 2

CLD3 . . . . . Compact Language Detector 3

CluHTM . . . . . Cluster Hierarchical Topic modelling

CluWord . . . . . Cluster of Word

CNPMI . . . . . Crosslingual Normalized Pointwise Mutual Information

CorEx . . . . . Correlation Explanation

CPLSA . . . . . Contextual Probabilistic Latent Semantic Analysis

CRTM . . . . . Constrained Relational Topic Model

CTI . . . . . Cyber Threat Intelligence

CTM . . . . . Community Topic Model

d-BTM . . . . . Discriminative Biterm Topic Model

DBSCAN . . . . . Density-Based Spatial Clustering of Applications with Noise

DCTM . . . . . Dynamic Community Topic Model

DF-LDA . . . . . Dirichlet Forest-Latent Dirichlet Allocation

DLTM . . . . . Deep Learning Topic Model

DMM . . . . . Dirichlet Multinomial Mixture

DNN . . . . . Deep Neural Network

DREx . . . . . Distributed Representation-based Expansion

DTM . . . . . Dynamic Topic Model

DTV . . . . . Dokument-Themen-Verteilung

DWT-LDA . . . . . Deep Word-Topic Latent Dirichlet Allocation

EC-LDA . . . . . Entity Correlation Latent Dirichlet Allocation

eCDF . . . . . empirical cumulative distribution function

EntLDA ..... Entity-based Topic Model

ES-LDA ..... LDA with Entity based Similarity

ETM ..... Embedded Topic Model

FBI ..... Federal Bureau of Investigation

FBoW ..... Fuzzy Bag of Word

FIM ..... Frequent Itemset Minings

FLATM ..... Fuzzy Logic Approach Topic Model

FREX ..... Frequency-Exclusivity

FTM ..... Fuzzy Topic Modelling

GAN ..... Generative Adversarial Net

GCN ..... Graph Convolutional Network

GPGBN ..... Graph Poisson Gamma Belief Network

GPU ..... Generalized Pólya Urn

GT ..... Groundtruth

GTM ..... Guided Topic-Noise Model

HDBSCAN ..... Hierarchical Density-Based Spatial Clustering of Applications with Noise

HFTM ..... Hierarchical Features-based Topic Model

HSLDA ..... Hybrid Semi-supervised LDA

HSPLSA ..... Hybrid Semi-supervised PLSA

IDF ..... Inverse Dokumentenfrequenz

inverse-AJS ..... inverse Average Jaccard Similarity

ISLDA ..... Interval Semi-supervised LDA

keyATM ..... keyword-Assisted Topic Model

KeyETM ..... Keyword Assisted Embedded Topic Model

KL-Divergenz ... Kullback-Leibler-Divergenz

L-LDA ..... Labelled-LDA

LC-LDA ..... Local Context-Aware LDA Model

LDA ..... Latent Dirichlet Allocation

LIMTopic ..... Link Importance based Topic Model

LSA ..... Latent Semantic Analysis

LTM ..... Lifelong Topic Modelling

M-BERT . . . . . Multilingual BERT

MC-LDA . . . . . M-set and C-set-Latent Dirichlet Allocation

MI . . . . . Mutual Information

mLDA . . . . . Multi-contextual LDA

MoNA . . . . . Mobile Network Analyzer

MP . . . . . Modell Precision

MTM . . . . . Multilingual Topic Model

NMF . . . . . Non-Negative Matrix Factorization

NMI . . . . . Normalized Mutual Information

NPMI . . . . . Normalized Pointwise Mutual Information

NTM . . . . . Neural Topic Model

OOV . . . . . Out-of-Vocabulary

OSDATM . . . . . Ordering-sensitive and Semantic-aware Dynamic Author Topic Model

PCA . . . . . Principal Component Analysis

PDMM . . . . . Poisson DMM

PLSA . . . . . Probabilistic Latent Semantic Analysis

PMI . . . . . Pointwise Mutual Information

PTM . . . . . Pseudo-document-based Topic Modelling

R4WSI . . . . . Random 4 Word Set Intrusion

RART . . . . . Role-Author-Recipient-Topic Model

RMATM . . . . . Reward-Modulated Adversarial Topic Modelling

RTM . . . . . Relational Topic Model

SATM . . . . . Self-Aggregation based Topic Modelling

Seeded LDA . . . . . Seeded Latent Dirichlet Allocation

sLDA . . . . . streaming LDA

sLDA . . . . . Supervised LDA

SMS . . . . . Short Message Service

SMTM . . . . . Seed-guided Multi-label Topic Model

STM . . . . . Social based Topic Model

STM . . . . . Structural Topic Model

STRuFSP . . . . . Semantic-based Topic Representation using Frequent Semantic Patterns

STTM ..... Short Text Topic Modelling

SVM ..... Support Vector Machine

SW ..... Summation over Words

SWB ..... Special Words with Background

SWH ..... Streaming with Heart

TAM ..... Targeted Analysis Model

TATM ..... Targeted Aspects Oriented Topic Modelling

TDM ..... Term-Dokument-Matrix

TND ..... Topic-Noise Discriminator

TOT ..... Topics over Time

TTM ..... Targeted Topic Modelling

TWINT ..... Twitter Intelligence Tool

TWV ..... Themen-Wort-Verteilung

UCR ..... Uniform Crime Reporting

UTOPIAN ..... User-driven Topic modelling based on Interactive Nonnegative Matrix Factorization

VAE ..... Variational Autoencoder

VF-DTM ..... Foreground Dynamic Topic Model

VGATM ..... Variational Graph Author Topic Model

VI ..... Variational Inference

WETC ..... Word Embedding based Topic Coherence

wLDA ..... weighted Latent Dirichlet Allocation

WLM ..... Wikipedia Link-based Measure

WNTM ..... Word Network Topic Model

1 Einleitung

Die mobile Kommunikation über Short Message Service (SMS) und Messenger-Dienste wie WhatsApp, Signal und Telegram erfreut sich heutzutage großer Beliebtheit. Dies wird beispielsweise anhand einer Studie des Digitalverbands Bitcom aus dem Jahr 2021 deutlich [1]. Demnach erhält ein deutscher Handynutzer durchschnittlich 13 Textnachrichten (SMS und Nachrichten von Messenger-Diensten) pro Tag. Hochgerechnet ergibt sich damit für das Jahr 2021 eine geschätzte Gesamtzahl von 300 Milliarden versendeten und empfangenen Textnachrichten in Deutschland.

Mobiltelefone dienen jedoch nicht nur der alltäglichen Kommunikation, sondern bieten auch Kriminellen die Möglichkeit, sich über diese zu Straftaten zu verabreden, diese zu planen, in Auftrag zu geben oder durchzuführen [2]. Dementsprechend gelten Kurznachrichten, die auf mobilen Endgeräten gespeichert sind, als wichtige Beweisquelle für forensische Untersuchungen. Die Auswertung der immensen Menge an Nachrichten, die auf dem Mobilfunkgerät eines Tatverdächtigen gespeichert sind, stellt den Ermittler jedoch vor eine große Herausforderung. Insbesondere bei organisierter Kriminalität und Bandenkriminalität ist zudem häufig die Untersuchung der Nachrichten von einer Vielzahl von Mobilfunkgeräten erforderlich [3]. Die manuelle Analyse der umfangreichen Nachrichtenmengen kann dementsprechend mit einem hohen Zeitaufwand verbunden sein und ist wahrscheinlich nicht effizient durchführbar.

Eine Möglichkeit, dem Ermittler automatisiert einen Überblick über die Inhalte, die in den zahlreichen Nachrichten diskutiert wurden, zu vermitteln und diese möglichst kompakt zusammenzufassen, ist die Themenmodellierung, die auch unter dem englischen Begriff „Topic Modelling“ bekannt ist. Unter der Themenmodellierung wird ein häufig angewendetes Verfahren des Text Minings verstanden, das darauf abzielt, latente Themen in einem Datensatz beziehungsweise einer Dokumentensammlung zu finden [4]. Der Begriff Thema ist hierbei nicht einheitlich definiert, kann jedoch nach Zhai und Massung [5] als der Grundgedanke verstanden werden, der in den Textdaten besprochen wird.

Die Themenmodellierung wurde in verschiedensten Anwendungsbereichen eingesetzt, die von der Unterstützung von Kunden und Herstellern bei der Auswertung von Produktbewertungen [z.B. 6–8], der Identifizierung neuer technologischer Trends aus Patentdokumenten [9] über die Zusammenfassung von medizinischen Forschungsartikeln [z.B. 10] bis hin zur Auswertung von Antworten bei Umfragen und Interviews in den Sozialwissenschaften [11–13] reichen. Hingegen wurde sie bislang in der Forensik kaum genutzt, insbesondere nicht für die Auswertung von forensischen Kurznachrichten.

Die Themenmodellierung in forensischen Kommunikationsdaten wird durch die vorliegende Arbeit adressiert. Ein wesentliches Ziel besteht darin, aufzuzeigen, welche Herausforderungen bei der Themenextraktion aus den forensischen Kurznachrichten bestehen und welche Lösungsstrategien sich anbieten, wofür eine systematische Literaturrecherche durchgeführt

wurde. Unter anderem wird hierbei auf die geringe Länge der Nachrichten, ihre mangelhafte sprachliche Qualität und die Tatsache, dass es sich teilweise um mehrsprachige Datensätze handelt, eingegangen.

Eine weitere Herausforderung ist darin, zu sehen, dass der Ermittler bei der Auswertung der Kurznachrichten nicht nur die reine Datenexploration bezweckt. Stattdessen hat er meist eine gewisse Vermutung über Themen, die in den Nachrichten diskutiert wurden und sucht nach Beweisen oder Indizien, die diese Hypothese stützen oder widerlegen. Jedoch erweist es sich hierbei als erschwerend, dass die fallrelevanten, für den Ermittler interessanten Themen nur zu einem geringen Anteil in dem Datensatz auftreten, während die Gespräche vor allem durch irrelevante Smalltalk-Themen dominiert werden. Die letztgenannte Herausforderung steht im Fokus des praktischen Teils dieser Arbeit. Hierbei werden verschiedene Ansätze der halbüberwachten Themenmodellierung [14, 15] auf realen Falldaten miteinander verglichen, die darauf abzielen, Themen zu extrahieren, die der Erwartungshaltung des Ermittlers entsprechen. Zusätzlich können diese ebenfalls neue Themen in dem Datensatz entdecken, was es dem Ermittler ermöglicht, nicht offensichtliche Zusammenhänge wie beispielsweise bisher unvermutete Verbindungen zu bestimmten Personen oder über die Motivation des Verbrechens zu erschließen.

Konkret werden die beiden halbüberwachten Algorithmen keyATM, das von Eshima u. a. [14] entwickelt wurde und die von Watanabe und Baturo [15] vorgeschlagene Seeded LDA untersucht und mit der unüberwachten LDA als Baseline-System verglichen. Zudem wird eine Erweiterung der beiden halbüberwachten Verfahren unter Einsatz des von Viegas u. a. [16] beschriebenen Ansatzes CluWords vorgeschlagen, der Informationen über die semantische Ähnlichkeit von Wörtern in die Dokumentenrepräsentation integriert.

Die Evaluierung erfolgt sowohl qualitativ als auch quantitativ. Jedoch ist ein grundlegendes Problem bei der quantitativen Evaluierung darin zu sehen, dass automatische Evaluierungsmaße nicht zwangsläufig die tatsächliche menschlich wahrgenommene Interpretierbarkeit von Themen widerspiegeln [17, 18]. Daher besteht ein weiteres Ziel dieser Arbeit darin, die Themenqualität gemäß der semantischen Kohärenz [19] als häufig eingesetztes Evaluierungsmaß mit den Resultaten einer Nutzerstudie auf der besonderen Domäne der forensischen Kommunikationsdaten zu vergleichen.

Die Arbeit ist folgendermaßen aufgebaut: Zunächst werden in Kapitel 2 die theoretischen Grundlagen und der Stand der Forschung betrachtet. Hierbei werden eine Taxonomie der verschiedenen Ansätzen zur Themenmodellierung und bisherige Arbeiten zum Einsatz der Themenmodellierung im Bereich der Forensik vorgestellt, bevor der Fokus auf die mit den forensischen Kommunikationsdaten verbundenen Herausforderungen und entsprechende Lösungsansätze gelegt wird. Anschließend werden in Kapitel 3 die verwendeten Daten und die eingesetzten Methoden beschrieben. Dies umfasst neben der Beschreibung des Datensatzes und der Aufbereitung dieser Daten die Erläuterung der Vorgehensweise zur halbüberwachten Themenmodellierung sowie die Methoden zur Evaluierung. Die Ergebnisse werden in Kapitel 4 dargestellt und diskutiert. In diesem Kapitel wird auf die Resultate der qualitativen und quantitativen Evaluierung sowie auf die Ergebnisse des Vergleichs zwischen der auto-

matisch bestimmten Themenqualität und der Nutzerstudie eingegangen. Die Arbeit endet mit einer kurzen Zusammenfassung in Kapitel 5 sowie einem Ausblick auf die zukünftige Forschung in Kapitel 6.

2 Grundlagen und Stand der Forschung

Das folgende Kapitel präsentiert und diskutiert bisherige Ansätze zur Themenmodellierung. Hierbei wird zunächst auf die verwendete Methodik zur Literaturrecherche eingegangen.

2.1 Methodik zur Literaturrecherche

Um einen Überblick über die verschiedenen Kategorien von Themenmodellen und Verfahren zu erhalten, wurden zunächst Übersichtsartikel wie [20–24] herangezogen. Auf diese Weise sollte vermieden werden, dass bekannte und häufig verwendete State of the Art Methoden nicht berücksichtigt werden.

Für die weitere Literaturrecherche wurde die von Döring [25] vorgeschlagene Vorgehensweise angewendet, die eine systematische Abfrage elektronischer Datenbanken mit dem Schneeballverfahren kombiniert. Hierbei wurde eine sogenannte eingegrenzte Recherche durchgeführt, da diese besonders geeignet für Forschungsgebiete wie die Themenmodellierung ist, zu der bereits eine Vielzahl von Publikationen erschienen sind [25]. Als Literaturdatenbanken wurden IEEE Xplore Digital Library, ACM Digital Library und ScienceDirect ausgewählt. Um die Suche einzugrenzen, wurde der primäre Suchbegriff, worunter ein geeigneter Oberbegriff für ein Forschungsthema wie hier die Themenmodellierung verstanden wird, über eine bool'sche UND-Verknüpfung mit spezifischen sekundären Suchbegriffen kombiniert [25]. Als primärer Suchbegriff wurde die bool'sche ODER-Kombination aus den Synonymen „topic modelling", „topic analysis" und „topic extraction" verwendet. Mithilfe der sekundären Suchbegriffe wurde der Fokus beispielsweise auf den Einsatzbereich der Forensik oder auf eine bestimmte Herausforderung gelegt, die mit der Themenmodellierung in forensischen Kommunikationsdaten verbunden ist, wie die geringe Länge der Kurznachrichten. Beispielsweise diente als sekundärer Suchbegriff für die Herausforderung, dass die forensischen Kommunikationsdaten als verrauscht gelten, die boolsch'sche ODER-Verknüpfung der Begriffe „noisy data", „statistical noise" und „noise".

Bezüglich aller Suchanfragen wurde vorgegeben, dass die Suchbegriffe in den Metadaten der Datenbankeinträge, genauer gesagt im Titel, dem Abstract oder den von den Autoren ausgewählten Keywords, vorkommen müssen. Zudem wurde die Suche auf Veröffentlichungen im Zeitraum von 2015 bis 2023 begrenzt. Ergab diese Suche weniger als fünf als relevant erachtete Treffer, wurde nach dem sekundären Suchbegriff nicht nur in den Metadaten, sondern auch im Volltext gesucht. Darüber hinaus wurde in einem solchen Fall der Zeitraum auf 2005 bis 2023 ausgedehnt. Publikationen wurden als relevant betrachtet und in die Literaturübersicht aufgenommen, wenn sie einen neuen Ansatz oder Algorithmus zur Themenmodellierung präsentierten beziehungsweise eine existierende Methode erweiterten. Ausgeschlossen wurden hingegen alle Arbeiten, die ein Verfahren ohne weitere Verbesserung auf einen konkreten, für die Forensik irrelevanten Anwendungsfall übertrugen, wie beispielsweise der Einsatz von Themenmodellen für Produktempfehlungen im kommerziellen Bereich.

Um zu vermeiden, dass Publikationen, die nicht in den ausgewählten Datenbanken veröffentlicht wurden, übersehen werden, wurde zusätzlich das Schneeballverfahren angewendet, worunter ein Suchverfahren verstanden wird, bei dem die Referenzen von ausgewählten Zeitschriftenartikeln nach weiterer verwandter Literatur durchsucht werden [26]. Als Ausgangswerke wurden als besonders relevant betrachtete Publikationen gewählt, die nach 2014 veröffentlicht wurden, um gewährleisten zu können, dass die mit dieser Methode gefundene Literatur so aktuell wie möglich ist.

Die auf diese Weise identifizierten Publikationen bilden die Grundlage für die folgende Literaturübersicht, in der die verschiedenen Kategorien von Themenmodellen, der bisherige Einsatz der Themenmodellierung im Bereich der Forensik sowie Herausforderungen bei der Extraktion von Themen aus forensischen Kommunikationsdaten und Lösungsstrategien präsentiert werden.

2.2 Taxonomie von Themenmodellen

Grundsätzlich können verschiedene Formen von Algorithmen zur Themenmodellierung unterschieden werden. Eine Möglichkeit für eine Taxonomie der einzelnen Ansätze und Beispiele für Algorithmen der verschiedenen Kategorien kann Abbildung 2.1 entnommen werden.

Die in Abbildung 2.1 aufgeführten Algorithmen haben gemeinsam, dass sie unüberwacht sind und darauf abzielen in Sammlungen von ungelabelten Dokumenten ausschließlich basierend auf dem Dokumenteninhalt latente Themen zu entdecken [23]. Jedoch sind vor allem für probabilistische Themenmodelle und Deep Learning Topic Models Erweiterungen für die halbüberwachte und überwachte Themenmodellierung [z.B. 15, 27, 28] sowie unter Einbeziehung von zusätzlichen Informationen [z.B. 29, 30] möglich (siehe Abschnitt 2.6 und Abschnitt 2.7).

Algebraische Themenmodelle Die ersten Ansätze zur Themenmodellierung, die auf die frühen 1990er Jahre zurückgehen, können in die Klasse der algebraischen Themenmodelle eingeordnet werden [31, 32]. Ihre Grundidee besteht darin, Themen basierend auf dem gemeinsamen Vorkommen von Wörtern in Dokumenten zu extrahieren [33]. Hierzu faktorisieren sie eine Term-Dokument-Matrix (TDM) und zielen darauf ab, eine niedrig dimensionale Approximation dieser TDM zu finden [31, 32]. Als für die Themenmodellierung relevante Ausgabe sind zwei reellwertige Matrizen zu sehen [31, 32]. Diese können als Wort-Themen-Matrix [31, 32] und Dokument-Themen-Matrix [31] beziehungsweise Themen-Dokument-Matrix [32] interpretiert werden. Die Einträge der Wort-Themen-Matrix geben jeweils an, welche Bedeutung ein Wort für ein Thema aufweist, während die Elemente der Dokument-Themen-Matrix beziehungsweise der Themen-Dokument-Matrix ausdrücken, wie relevant ein Thema für ein Dokument ist [31, 32].

Als bekannte Vertreter der algebraischen Themenmodelle gelten die von Deerwester u. a. [31] beschriebene Latent Semantic Analysis (LSA) und die Non-Negative Matrix Factorization (NMF), die von Paatero [32] entwickelt und von Shahnaz u. a. [34] zur Anwendung für die Themenmodellierung vorgeschlagen wurde. Ein Überblick über die Vor- und Nachteile der

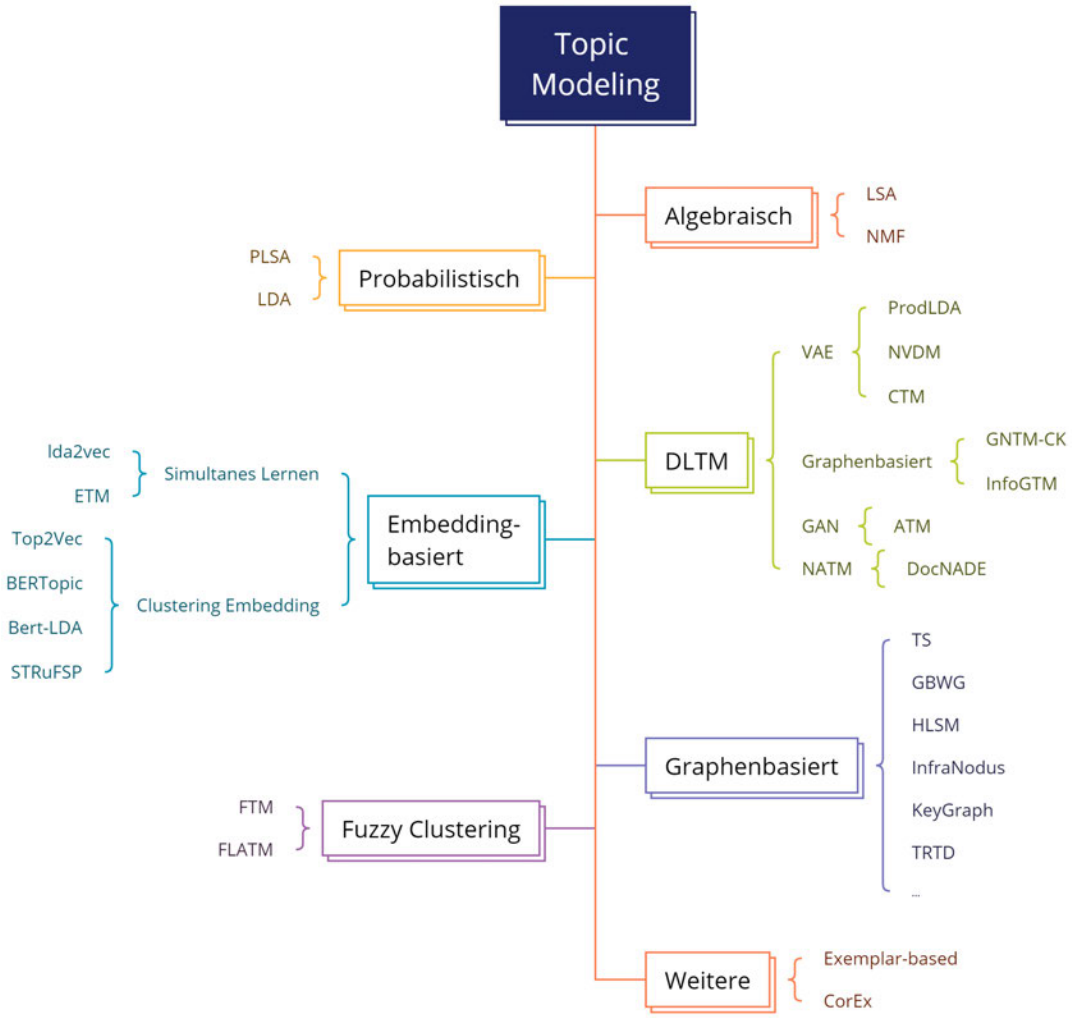

Abbildung 2.1: Taxonomie von Themenmodellen. Die Ansätze der Themenmodellierung wurden in die Oberkategorien *algebraisch, probabilistisch, DLTM, Embedding-basiert, graphenbasiert, Ansätze des Fuzzy Clusterings* und *Weitere Ansätze* aufgeteilt. Für die Oberkategorien DLTM und Embedding-basiert wurden weitere Unterkategorien eingeführt. Für die einzelnen Ober- bzw. Unterkategorien sind exemplarisch einige bekannte Algorithmen aufgeführt. Quelle: Eigene Darstellung.

beiden algebraischen Verfahren kann Tabelle 2.1 entnommen werden. Ein wesentlicher Unterschied bezüglich der beiden Ansätze ist darin zu sehen, dass die Term-Themen-Matrix und Themen-Dokument-Matrix bei NMF im Gegensatz zu der LSA keine negativen Werte enthalten [32, 34]. Dies ist insofern vorteilhaft, da, wie von Chen u. a. [35] und ferner von Shahnaz u. a. [34] und Choo u. a. [36] betont wurde, die Interpretation der negativen Gewichte von Wörtern für Themen häufig schwer fällt, weil es realistischer ist anzunehmen, dass jedes Wort eine, wenn auch geringe, Relevanz für ein Thema besitzt.

Die einzelnen Arbeiten widersprechen sich hinsichtlich der Aussage, ob NMF für kurze und verrauschte Texte wie forensische Kurznachrichten geeignet ist [24, 37–39]. Nach Churchill und Singh [24] kommt NMF gerade für verrauschte Texte infrage, da die durch NMF durchgeführte Dimensionsreduktion Rauschen entfernen könnte. Von Chen u. a. [38] und Albalawi u. a. [39] wurde ebenfalls gezeigt, dass NMF qualitativ und quantitativ hochwertigere Themen auf kurzen, verrauschten Texten wie Tweets hervorbringen kann als probabilistische Themenmodelle wie die nachfolgend erklärte LDA [4]. Jedoch steht dies im Widerspruch zu der experimentellen Studie von Murshed u. a. [23]. Diese verwendeten dasselbe Maß zur Evaluierung wie Chen u. a. [38], nämlich die in Abschnitt 2.11 aufgeführte semantische Kohärenz nach Mimno u. a. [19]. Jedoch kamen sie zu dem Schluss, dass NMF zu deutlich schlechteren Ergebnissen als probabilistische Themenmodelle und Ansätze des Fuzzy Clusterings auf einem Datensatz von Tweets führt. Dementsprechend ist NMF nicht prinzipiell besonders gut für verrauschte Texte geeignet, sondern die Ergebnisse sind neben der Evaluierungsmethode beispielsweise auch abhängig von der Größe des Datensatzes.

Tabelle 2.1: Vor- und Nachteile von ausgewählten algebraischen und probabilistischen Algorithmen der Themenmodellierung. Als algebraische Themenmodelle sind die LSA sowie die NMF in der Tabelle aufgeführt. Zudem sind die Vor- und Nachteile der probabilistischen Algorithmen Probabilistic Latent Semantic Analysis (PLSA) und LDA dargestellt.

Algorithmus	**Vorteile**	**Nachteile**
LSA [31]	Erstes Themenmodell Thema = Rangfolge von Wörtern	Mangelhafte Interpretierbarkeit [35] Ungeeignet für kurze Texte [40] Kein Umgang mit Polysemie [41] Hohe Rechenkomplexität [42, 43]
NMF [32]	Bessere Interpretierbarkeit als LSA Thema = Rangfolge von Wörtern Mehrere Themen je Dokument Hohe Laufzeiteffizienz [36]	Ungeeignet für kurze Texte [23, 44] z.T. ungeeignet für verrauschte Texte [38]
PLSA [41]	Verbesserung von LSA Umgang mit Polysemie Gute Interpretierbarkeit Mehrere Themen je Dokument	Hohe Anzahl an Parametern [4] Neigt zu Overfitting [4] Keine Vorhersage für neue Dokumente [4] Ungeeignet für kurze Texte [45]
LDA [4]	Gleiche Vorteile wie PLSA i.G.z. zu PLSA generatives Modell Ermöglicht Vorhersagen	Ungeeignet für kurze Texten [45] Geringe Adaption ⇒ Anpassung Inferenzverfahren [46]

Probabilistische Themenmodelle Häufig werden zudem probabilistische Algorithmen zur Themenmodellierung [z.B. 4, 41] eingesetzt. Die wesentliche Idee dieser Ansätze besteht darin, ein Dokument als eine Wahrscheinlichkeitsverteilung von latenten Themen zu beschreiben, während ein Thema als eine Wahrscheinlichkeitsverteilung über Wörter charakterisiert wird. Wie die algebraischen Algorithmen analysieren sie zur Themenextraktion Wortkookkurrenzen auf Dokumentenebene [33]. Die von Blei u. a. [4] beschriebene LDA gilt als einer der populärsten Algorithmen zur Themenmodellierung. Bei dieser handelt es sich im Gegensatz zu der PLSA um ein generatives Modell, das Vorhersagen für neue Dokumente treffen kann, die nicht zum Lernen des Themenmodells herangezogen wurden [4]. Dies wird dadurch erreicht, dass bei der LDA eine Dirichlet-Verteilung über die Modellparameter, genauer gesagt über die Dokument-Themen-Verteilung (DTV) und die Themen-Wort-Verteilung (TWV), eingeführt wurde. Im Rahmen des generativen Prozesses wird zunächst für jedes Wort w_{di} eines Dokuments d eine Themenzuweisung $z_{di} \in [1, ..., K]$ bei einem Themenmodell mit K Themen aus der Themenverteilung des Dokuments θ_d gezogen und anschließend das Wort w_{di} mit der Wortverteilung des entsprechenden Themas ϕ_k generiert. Das Ziel beim Training der LDA besteht darin, die latenten Modellparameter basierend auf dem beobachteten Vorkommen der Wörter im Dokument zu schätzen. Hierzu muss die Posterior-Inferenz berechnet werden, was jedoch als rechnerisch aufwendig gilt [4], weshalb üblicherweise Approximationsverfahren wie Collapsed Gibbs Sampling [47] oder Variational Inference (VI) [4] eingesetzt werden. Eine umfassendere Darstellung der Vor- und Nachteile der PLSA und LDA kann den letzten beiden Zeilen von Tabelle 2.1 entnommen werden.

Deep Learning basierte Themenmodellierung Darüber hinaus erfuhren seit 2015 die DLTMs zunehmende Beliebtheit [33]. Nach Zhao u. a. [48] werden unter diesen Themenmodelle verstanden, bei denen die Themen mit Deep Neural Networks (DNNs) gelernt werden. Diese Ansätze adressieren vor allem das von Srivastava und Sutton [46] hervorgehobene Problem, dass die Inferenzverfahren zur Parameterschätzung der LDA bei geringster Änderung der Modellarchitektur erneut hergeleitet werden müssen. Die einzelnen Verfahren können bezüglich des DNNs, mit dem die latenten Parameter gelernt werden, kategorisiert werden, wobei beispielsweise Variational Autoencoder (VAE) [z.B. 46, 49], graphenbasierte DLTMs [z.B. 50], Generative Adversarial Nets (GANs) [51, z.B.] und autoregressive (AR) DLTMs [52] herangezogen wurden. Die DLTMs, die auf VAEs [z.B. 46, 49] beziehungsweise graphenbasierten DLTMs [50, 53] beruhen, zeichnen sich dadurch aus, dass sie den generativen Prozess der LDA erweitern und dementsprechend dieselben Verteilungen wie die gewöhnliche LDA, d.h. eine DTV und eine TWV schätzen. Für eine detailliertere Gegenüberstellung der Ansätze in den verschiedenen Unterkategorien wird auf Tabelle 2.2 verwiesen.

Tabelle 2.2: Vergleich der verschiedenen Unterkategorien von DLTMs. Dargestellt sind die Vor- und Nachteile der verschiedenen Unterkategorien von Ansätzen zur Themenmodellierung unter Einbeziehung von DNNs. Die Unterteilung beruht auf dem verwendeten DNN.

Unter-kategorie	Vorteile	Nachteile	Referenz
VAE	• Generatives Modell • Gleiche Ausgabe wie LDA • Adressiert Inferenzproblem • Höhere Flexibilität • Gezielte Optimierung z.B. nach Themenkohärenz	• z.T. geringe Stabilität • Dirichlet-Vtl. nicht möglich → Alternativen z.T. unzureichend [51] • Hohe Laufzeit [33] • Optimierungskriterium ggf. ungeeignet [54]	[46, 49, 55]
Graphen-basiert	• Gleiche Vorteile wie VAE • Dokument = Graph → Erfassen von versch. Wortbeziehungen	• Gleiche Nachteile wie VAE	[50, 53]
GAN	• Adressiert Inferenzproblem • Liefert TWV • Lernt simultan Word Embeddings • Dirichlet-Prior möglich	• z.T. keine DTV • Ziel: Fehlerverringerung bei Dokumentenrekon-struktion → nicht zwangsläufig interpretierbare Themen [54]	[51]
AR	• Thema = Rangfolge von Wörtern • Mehrere Themen je Dokument • Simultanes Lernen von Word Embeddings	• Bisher geringer Einsatz	[52]

Embedding-basierte Ansätze Darüber hinaus wurden Algorithmen vorgeschlagen, die Themen basierend auf der Ähnlichkeit von Word Embeddings oder Dokument Embeddings lernen [z.B. 56, 57]. Unter einem Word Embedding wird im Allgemeinen eine Abbildung eines Wortes auf eine Vektorrepräsentation in einem niedrig dimensionalen Raum verstanden [5, 58]. Diese wird anhand der Kontexte gelernt, in denen ein Wort in einem Korpus auftritt [59]. Wörter mit einer hohen semantischen Ähnlichkeit werden auf nahe beieinander liegende Word Embeddings projiziert, sodass die Ähnlichkeit von Begriffen ermittelt werden

kann, indem diese Wortvektoren beispielsweise mit der Kosinusähnlichkeit verglichen werden [60]. Dokument Embeddings entsprechen den Darstellungen von Dokumenten als dichte Vektorrepräsentationen [61].

Eine Zusammenfassung der Vor- und Nachteile der verschiedenen Embedding-basierten Ansätze der Themenmodellierung kann Tabelle 2.3 entnommen werden. Grundsätzlich können zwei Arten von Embedding-basierten Ansätzen unterschieden werden: das simultane Lernen von Word Embeddings und Themen sowie die sogenannten Cluster Embedding Ansätze.

Zum einen wurden Verfahren wie lda2vec [57] und Embedded Topic Model (ETM) [62] entwickelt, die gleichzeitig Word Embeddings basierend auf dem Skip-Gram Ansatz von word2vec [63] und Themen lernen. Beide genannten Verfahren bauen auf dem generativen Prozess der LDA auf und liefern dementsprechend als Resultat eine TWV und eine DTV. Die grundlegende Idee besteht darin, Wörter, Themen und im Falle von lda2vec zusätzlich Dokumente als dichte Vektoren in einem gemeinsamen Embedding Raum zu repräsentieren. Im Gegensatz zu gewöhnlichen probabilistischen Themenmodellen werden die Themen nicht basierend auf Wortkookkurrenzen, sondern ausschließlich basierend auf der semantischen Ähnlichkeit nach Embeddings gelernt. Es muss darauf hingewiesen werden, dass ETM ebenfalls als DLTM aufgefasst werden kann, da dieses zur Schätzung der Posterior-Inferenz VAE einsetzt [55].

Zum anderen wurden sogenannte Cluster Embedding Ansätze wie BERTopic [64], BERT-LDA [43], Top2Vec [65] und Semantic-based Topic Representation using Frequent Semantic Patterns (STRuFSP) entwickelt. Sowohl BERTopic [64] als auch die Erweiterung BERT-LDA [43] clustern die Dokumente des Datensatzes basierend auf der Ähnlichkeit von Dokument Embeddings, die mithilfe des Sprachmodells Bidirectional Encoder Representations from Transformers (BERT) [66] gelernt wurden, und fassen die Cluster von Dokumenten als Themen auf. Als Clusteralgorithmus dient hierbei der von McInnes u. a. [67] entwickelte Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), der verhindert, dass Dokumente, die keinen anderen Dokumenten ähneln und vermutlich irrelevant sind, einem Cluster beziehungsweise Thema zugewiesen werden [43, 64]. Um die Themen mithilfe von Wörtern beschreiben zu können, werden mit einem adaptierten TF-IDF-Maß möglichst relevante Wörter aus den Dokumenten jedes Clusters extrahiert [64]. Der von Angelov [65] vorgeschlagene Algorithmus Top2Vec führt ebenfalls ein Clustering von Dokumenten durch, definiert als Thema jedoch nicht das Cluster von ähnlichen Dokumenten, sondern den Zentroiden dieses Clusters. Wörter werden dem Thema über die Ähnlichkeit zwischen dem Themenvektor und den Embedding Repräsentationen der Wörter zugewiesen. Das von Geeganage u. a. [56] beschriebene Verfahren STRuFSP unterscheidet sich insofern von den anderen Embedding-basierten Ansätzen, dass nicht Dokumente, sondern Wörter basierend auf ihrer Word Embedding Ähnlichkeit geclustert werden und diese Cluster als Thema deklariert werden. Dementsprechend beschreibt dieser Ansatz nicht, inwiefern die Themen in den einzelnen Dokumenten vertreten sind.

Tabelle 2.3: Vergleich von Embedding-basierten Ansätzen zur Themenmodellierung. Aufgezeigt werden die Vor- und Nachteile von lda2vec und ETM, die beide auf einem simultanen Lernprozess von Themen und Word Embeddings basieren. Ebenfalls sind BERTopic, BERT-LDA, Top2Vec und STRuFSP als Cluster Embedding Ansätze aufgeführt.

Algorithmus	Vorteile	Nachteile
lda2vec [57]	• Gleiche Ausgabe wie LDA • Lernt gleichzeitig Word Embeddings	• Hohe Laufzeit [68] • Vielzahl an Parametern [68]
ETM [62]	• Kombiniert Vorteile von LDA, WordEmbeddings und VAE • Geeignet für großes Vokabular • Keine Stoppwortentfernung nötig	• Ungeeignet für kleine Datensätze [33] • Ungeeignet für kurze Texte [33]
Top2Vec [65]	• #Themen automatisch ermittelt • Keine Vorverarbeitung nötig • Automatischer Ausschluss irrelevanter Dokumente • Thema als Rangfolge von Wörtern	• Nur ein Thema je Dokument [69] • Hohe Modellkomplexität [70] • z.T. Ausschluss von zu vielen Dokumenten [69]
BERTopic [64]	• Gleiche Vorteile wie Top2Vec • Besonders geeignet für Polysemie und Synonymie [33]	• Gleiche Nachteile wie Top2Vec • z.T. redundante, ähnliche Wörter in Themen • Vortrainierte Embeddings ggf. ungeeignet
BERT-LDA [43]	• Berücksichtigt zusätzlich thematische Dokumentenähnlichkeit nach LDA	• Gleiche Nachteile wie BERTopic
STRuFSP [56]	• Weniger irrelevante, bedeutungslose Wörter in Themen • Thema als Rangfolge von Wörtern	• Keine Themenzuweisung/ -verteilung von Dokumenten

Graphenbasierte Ansätze Eine weitere Kategorie von Verfahren der Themenmodellierung bilden graphenbasierte Ansätze [71–80]. Die wichtigsten Vor- und Nachteile dieser Ansätze können der ersten Zeile von Tabelle 2.4 entnommen werden. Die meisten der Vertreter dieser

Oberkategorie definieren ein Thema als eine ungeordnete, disjunkte Menge von semantisch ähnlichen Begriffen [z.B. 72–74]. Um diese Begriffe zu identifizieren, wird ein Wortnetzwerk erstellt. Die Knoten dieses Netzwerkes repräsentieren die Wörter des Vokabulars des Datensatzes. Bei den meisten Ansätzen werden diese über Kanten miteinander verbunden, wenn sie gemeinsam in einem festgelegten Kontext wie einem Dokument [72, 73, 75, 77, 78] oder einem Paragraphen [71] vorkommen. Als Kantengewichte dienen hierbei beispielsweise die Anzahl der Dokumente, in denen die beiden Wörter gemeinsam auftreten [z.B. 72, 75, 78]. Jedoch ermöglichen es graphenbasierte Ansätze auch andere Kriterien als das gemeinsame Auftreten von Wörtern zur Themenmodellierung heranzuziehen, sodass beispielsweise von Tolegen u. a. [79] Wörter basierend auf ihrer Word Embedding Ähnlichkeit verbunden wurden. Es ist darauf hinzuweisen, dass in einigen Arbeiten nur eine Verbindung zwischen Wörtern in dem Netzwerk hergestellt wurde, wenn beispielsweise ihr gemeinsames Auftreten [72, 81], ein Signifikanzmaß [z.B. 71] oder ein Ähnlichkeitsmaß [z.B. 77] über einem bestimmten Schwellenwert lag, wodurch irrelevantes beziehungsweise zufälliges gemeinsames Vorkommen von Wörtern bei der Themenmodellierung zur Eliminierung des Rauschens nicht berücksichtigt wurde.

Im Anschluss an die Errichtung des Wortnetzwerkes werden meist Algorithmen der sogenannten Community Detection durchgeführt [71–74, 76–79]. Eine Community ist nach Hamm und Odrowski [72] als eine Gruppe von Knoten des Graphen definiert, innerhalb derer eine hohe Kantendichte vorliegt, während nur wenige Kanten zu anderen Gruppen existieren. Eine im Rahmen der Themenmodellierung häufig angewendete Möglichkeit, um diese Communities zu finden, besteht darin, den von Blondel u. a. [82] entwickelten Louvain-Algorithmus wie beispielsweise in [71, 74, 76, 78, 79, 83] oder den Leiden Algorithmus [84] wie in [72] einzusetzen. Die mithilfe dieser Algorithmen gefundenen Communities können entweder direkt als Themen aufgefasst werden [71, 72, 78, 79] oder es werden anschließende Nachverarbeitungsschritte wie die Zusammenführung von mehreren Communities zu einem Thema durchgeführt [73, 85].

Fuzzy Clustering In wenigen Arbeiten wurde zudem das Fuzzy Clustering zur Themenmodellierung eingesetzt [10, 86], dessen Vor- und Nachteile der zweiten Zeile von Tabelle 2.3 entnommen werden können. Die wesentliche Idee besteht darin, die Wörter des Datensatzes basierend auf ihrem Vorkommen in Dokumenten zu clustern und die Term-Cluster als Themen aufzufassen. Hierfür wurden Fuzzy-Clustermethoden wie der Fuzzy-c-Means-Algorithmus [87] angewendet. Im Gegensatz zu dem harten Clustering werden die Wörter nicht einem Cluster fest zugewiesen, sondern können zu mehreren Clustern beziehungsweise Themen gehören, wobei die Wörter einen Grad der Zugehörigkeit für jedes Thema erhalten [10, 86]. Die Themenverteilung eines Dokuments kann anschließend aus der Zuweisung seiner Wörter zu Themen abgeleitet werden. Es ist zu betonen, dass der von Rashid u. a. [86] entwickelte Algorithmus Fuzzy Topic Modelling (FTM) und das von Karami u. a. [10] vorgeschlagene Fuzzy Logic Approach Topic Model (FLATM) Verfahren der Dimensionsreduktion wie die Principal Component Analysis (PCA) [88] oder eine iterative Anwendung des Fuzzy Clusterings einsetzten, weshalb sie nach Rashid u. a. [86] gerade für Datensätze mit einer hohen Sparsität und stark verrauschten Texten wie für forensische Kommunikationsdaten geeignet sind.

Tabelle 2.4: Überblick über die Vor- und Nachteile der graphenbasierten Themenmodelle, der Ansätze des Fuzzy Clusterings, des Exemplar-based Topic Modelling und dem Algorithmus CorEx. Die Oberkategorie sowie ihre Vor- und Nachteile und die Referenzen der Algorithmen dieser Oberkategorien sind in der Tabelle aufgeführt.

Oberkategorie	Vorteile	Nachteile	Referenz
Graphenbasiert	• Automatische Bestimmung der #Themen • Geeigneter für kurze Texte • Ausschluss von irrelevanten Wort-verbindungen möglich	• Thema i.d.R. als disjunkte, ungeordnete Wortmenge • Meist fehlender Umgang mit Polysemie • DTV nicht als direkte Ausgabe	[71–80]
Fuzzy Clustering	• Geeignet für kurze Texte • Berücksichtigt Polysemie • Thema als Rangfolge von Wörtern	• Wkt.-Repräsentation eines Themas nur durch Nachverarbeitung • DTV nur nachträglich möglich	[10, 86]
Weitere - Exemplar-based	• v.a. für kurze und verrauschte Texte • Geringe Laufzeit	• Keine DTV als Resultat • Keine „klassische“ Themenrepräsentation durch Wörter	[89]
Weitere - CorEx	• Thema = Rangfolge von Wörtern • Mehrere Themen je Dokument • Geeignet für kurze Texte	• Disjunkte Themen ⇒ Keine Berücksichtigung von Polysemie • Weniger geeignet für lange Dokumente	[90]

Weitere Formen der Themenmodellierung Zuletzt ist auf zwei Ansätze einzugehen, die nicht einer der Hauptkategorien zugeordnet werden können und in den letzten beiden Zeilen von Tabelle 2.4 aufgeführt sind. Hierzu zählt das von Elbagoury u. a. [89] erläuterte Exemplar-Based Topic Modelling, das sich speziell der Themenextraktion aus Twitter-Datensätzen widmet. Die Autoren gingen davon aus, dass gerade für kurze Texte Themen besser durch ein repräsentatives Dokument, in diesem Fall einen einzigen Tweet, dargestellt werden können, als durch eine Menge von Wörtern. Dieses sogenannte Exemplar wurde mithilfe eines ite-

rativen Verfahrens basierend auf der Untersuchung der Kosinusähnlichkeit zwischen den Tweets ausgewählt [89]. Der ausgewählte Tweet sollte eine möglichst hohe Ähnlichkeit zu einigen Tweets aufweisen und möglichst unähnlich zu allen anderen Tweets sein.

Darüber hinaus wurde von Gallagher u. a. [90] vorgeschlagen, den Algorithmus CorEx [91], der auf der Informationstheorie basiert, zur Themenmodellierung einzusetzen. Die grundlegende Idee ist darin zu sehen, die Wörter und Themen als diskrete Zufallsvariablen zu betrachten. Basierend auf der Mutual Information (MI) [92] werden im Rahmen eines iterativen Algorithmus die Wörter Themen so zugewiesen, dass die Abhängigkeiten beziehungsweise Korrelationen von Wörtern in Dokumenten durch latente Themen maximal erklärt werden. Ein Thema wird als Rangfolge von Wörtern dargestellt, wobei die Wörter basierend auf der MI sortiert werden. Jedes Wort wird hierbei exakt einem Thema zugewiesen.

2.3 Themenmodellierung in der Forensik

Der folgende Abschnitt präsentiert Arbeiten, die die Themenmodellierung im Bereich der Forensik anwendeten, und zeigt Herausforderungen für die Themenmodellierung in forensischen Kommunikationsdaten auf.

2.3.1 Einsatz in der Forensik und im Kontext von Kriminalität

Bisher wurde die Themenmodellierung nur in wenigen Arbeiten [93–96] mit dem Ziel eingesetzt, Datensätze im Rahmen von forensischen Untersuchungen kompakt zusammenzufassen. Zudem fokussierten sich diese nicht auf die Analyse von Themen in forensischen Kommunikationsdaten. Beispielsweise schlugen de Waal u. a. [94] vor, die LDA einzusetzen, um einen Überblick über die Themen in sämtlichen textuellen Daten, die für einen Fall untersucht werden müssen, zu erhalten. Zu diesen zählten neben Kurznachrichten beispielsweise auch E-Mails und Notizen in Textdokumenten. Noel und Peterson [93] zogen ebenfalls die LDA heran, um Themen in allen Word Dokumenten, die aus einem Festplattenspeicher extrahiert wurden, zu finden.

Darüber hinaus kam Topic Modelling zum Einsatz, um die Datenanalyse in bestimmten Deliktbereichen beziehungsweise bei der Auswertung konkreter Fälle zu unterstützen. Beispielsweise wurde von Li u. a. [95] das probabilistische Themenmodell Biterm Topic Modelling (BTM) [97], auf das in Abschnitt 2.4 näher eingegangen wird, gewählt, um Themen in Gesprächen über Korruption auf Twitter zu extrahieren und noch nicht bekannt gewordene Korruptionsfälle aufzudecken. Zudem wurde von Busso u. a. [96] das Themenmodell Structural Topic Model (STM) [11] eingesetzt, um Themen in einer Reihe von rassistischen und beleidigenden Briefen zu erkennen, die in den Jahren 2007 bis 2009 an Privatpersonen und Personen des öffentlichen Lebens verschickt worden waren [98].

Neben dem Einsatz der Themenmodellierung für die Zusammenfassung von großen Datenmengen, wurde es von Joseph und Viswanathan [99] auch zur Unterstützung der Suche in umfassenden Mengen an Dateien herangezogen, die zur forensischen Analyse aus beliebigen Festplattenspeichern extrahiert wurden. Die wesentliche Idee bestand darin, mit einer

adaptierten LDA eine Themenanalyse durchzuführen und anschließend aus den erhaltenen Themen manuell geeignete Suchbegriffe auszuwählen, um beweiserhebliche Informationen in der Datenmenge zu finden.

Zudem wurde die Themenmodellierung von Okolica u. a. [100] und Yang u. a. [101] eingesetzt, um relevante Personen im Enron Skandal identifizieren zu können. Konkret wurde das von Rosen Zvi u. a. [102] vorgestellte Author Topic Model (ATM), auf das in Abschnitt 2.6 eingegangen wird, auf dem Enron Email Korpus angewendet. Okolica u. a. [100] verfolgten hierbei das Ziel Whistleblower anhand der Auswertung der E-Mails erkennen zu können. Ihnen gelang es durch die Ergebnisse der Themenmodellierung sowie der Konstruktion von mehreren Kontaktnetzwerken den tatsächlichen Whistleblower in dem Enron Skandal ausfindig zu machen. Von Yang u. a. [101] wurden mithilfe von ATM Personen des Unternehmens identifiziert, die in ihren E-Mails über Themen diskutierten, die für den Enron Skandal relevant waren.

Darüber hinaus wurde Topic Modelling für präventive Zwecke eingesetzt [103, 104]. Beispielsweise wurde von Bérubé u. a. [103] vorgeschlagen, mit der LDA Tweets nach einem bestimmten Ereignis wie einem Terroranschlag zu untersuchen, um die Reaktionen in der Gesellschaft zu verfolgen. Hierbei wurde bezweckt, rechtzeitig präventive Maßnahmen zu ergreifen, falls das Ereignis bedenkliche Auswirkungen wie Fremdenfeindlichkeit zur Folge hat. Zudem wurde von Rule u. a. [104] die Themenmodellierung angewendet, um rechtzeitig erkennen zu können, wenn sich Diskussionen zu bestimmten Themen auf sozialen Medien wie Twitter emotional aufschaukeln, da diese in potenziell gewalttätigen Ereignissen wie Protesten und öffentlichen Versammlungen münden können. Hierzu wurde vorgeschlagen, täglich Tweets auszuwerten und brisante Themen durch die Ergebnisse der Themenanalyse mittels der LDA in Verbindung mit einer Sentiment Analyse zu detektieren.

Ein weiterer Anwendungsbereich war im Bereich der Cyber Threat Intelligence (CTI) zu sehen, unter dem nach Suryotrisongko u. a. [105] das Sammeln und Analysieren von Informationen über Bedrohungen verstanden wird, mit dem Ziel diese Informationen zur Abwehr der Gefahren zu verwenden. Konkret kam die Themenmodellierung zum Einsatz, um Themen zu erkennen, die in Hacker-Foren und sozialen Plattformen im Dark Web diskutiert wurden [105–110]. Auf diese Weise sollte ein Einblick in neue Vorgehensweisen, Techniken und Kommunikationskanäle von Hackern gewonnen werden, um aktuelle Sicherheitskontrollen verbessern zu können [106, 109] und IT-Sicherheits- und Reaktionsteams auf potenzielle Angriffe vorbereiten zu können [106]. Die meisten Arbeiten [106–109] setzten hierfür die LDA ein, während Suryotrisongko u. a. [105] und Vahedi u. a. [110] BERTopic [64] und top2vec [65] als Embedding-basierte Ansätze wählten.

Zuletzt wäre hinzuzufügen, dass von Kuang u. a. [111] und Pandey und Mohler [112] sogenannte Crime Topic Models zur Analyse von Polizeiberichten im amerikanischen Raum entwickelt wurden. Sie adressierten hierbei das Problem, dass kriminelle Ereignisse für Meldezwecke in Verbrechenskategorien nach dem Uniform Crime Reporting (UCR) Kategorisierungssystems des Federal Bureau of Investigation (FBI) eingeteilt werden, die jedoch als zu grob und ungenau empfunden wurden, wodurch Details über das Tatgeschehen verloren gehen. Um eine geeignetere Kategorisierung von kriminellen Ereignissen zu finden, führten Kuang u. a. [111] eine hierarchische Version der NMF und Pandey und Mohler [112] die LDA

auf kurzen Beschreibungen des kriminellen Ereignisses durch, die von Polizeibeamten verfasst wurden und den Polizeiberichten beigelegt waren. Die auf diese Weise extrahierten Themen wurden als neue Verbrechenskategorien aufgefasst.

Zusammenfassend kann gesagt werden, dass Topic Modelling das Potential besitzt, die forensische Untersuchung und die Polizeiarbeit vielfältig zu unterstützen. Allerdings fand es bisher kaum Anwendung in forensischen Kommunikationsdaten. Dies kann damit begründet werden, dass diese eine Reihe von Herausforderungen für die Themenmodellierung mit sich bringen, auf die im Folgenden eingegangen wird.

2.3.2 Herausforderungen in der Forensik

Kurze Texte Als problematisch für die Themenmodellierung erweist sich die geringe Länge von SMS und Nachrichten von Messenger-Diensten [113]. Beispielsweise ist die Länge von SMS auf 160 Zeichen begrenzt [114] und die durchschnittliche Länge einer WhatsApp Nachricht liegt bei nur 508 Zeichen [115]. Dementsprechend weisen Korpora von Kurztexten eine hohe Sparsität auf, womit gemeint ist, dass der Korpus über ein hohes Vokabular verfügt, aber die einzelnen Texte nur aus wenige Wörtern bestehen [116]. Dies erweist sich für traditionelle Ansätze als problematisch [97]. Die Ursache ist darin zu sehen, dass die meisten klassischen Verfahren, insbesondere probabilistische Ansätze wie die häufig angewendete LDA, aber auch algebraische oder graphenbasierte Methoden auf Kookkurrenzen von Wörtern innerhalb von Dokumenten beruhen [117]. Die geringe Anzahl von Wörtern in einer Kurznachricht führt jedoch dazu, dass den Themenmodellen nicht ausreichend Informationen über das gemeinsame Auftreten von Wörtern auf Dokumentenebene zur Verfügung stehen [45]. Dementsprechend zeigten beispielsweise Tang u. a. [118], dass die LDA selbst bei großen Datensätzen zu mangelhaften Ergebnissen führt, wenn die Dokumente eine geringe Länge besitzen.

Informelle und verrauschte Texte Von dem Problem der geringen Länge von Texten sind neben Kurznachrichten ebenfalls beispielsweise die Titel von Zeitungsartikeln oder wissenschaftliche Artikel [119] und Web Snippets [120] betroffen. Im Gegensatz zu diesen erweist sich jedoch bei (forensischen) Kurz- und Chatnachrichten zusätzlich als herausfordernd, dass es sich um informelle Texte mit mangelhafter sprachlicher Struktur handelt [3]. Xu u. a. [121] verdeutlichten anhand einer umfassenden quantitativen Evaluierung unter Verwendung von Maßen wie der semantischen Kohärenz [19] Abschnitt 2.11), dass die Ergebnisse der häufig angewendeten LDA stark von der Qualität der Eingabedaten abhängen. Relevante Themen konnten vor allem aus formell geschriebenen Texten wie Konferenzartikeln extrahiert werden, während die Themen von Tweets überwiegend als bedeutungslos anzusehen waren.

Ähnlich wie Tweets verfügen forensische Kurznachrichten über eine mangelhafte grammatikalische Struktur und enthalten emotional motivierte Buchstabenwiederholungen wie beispielsweise in „sooo sehr“ und zahlreiche Rechtschreib- und Tippfehler [3]. Letztere stellen insbesondere ein Problem dar, da trotz geläufiger Vorverarbeitungsschritte wie Stemming Wörter mit derselben Bedeutung aufgrund von orthographischen Fehlern nicht unbedingt als derselbe Begriff aufgefasst werden [94, 122]. Dies kann beispielsweise zu fälschlichen Aussagen über das gemeinsame Vorkommen von Wörtern in Dokumenten führen, was sich

beispielsweise für probabilistische Verfahren als erschwerend erweisen kann [94]. Darüber hinaus vergrößert sich hierdurch wiederum das Vokabular, wodurch das Problem der hohen Sparsität noch gravierender ist [122].

Hinzu kommt, dass Kurznachrichten eine hohe Anzahl von sogenannten Rauschwörtern, eher bekannt unter dem englischen Begriff „Noise Words“, enthalten, worunter nach Likhitha u. a. [20] bedeutungslose Begriffe beziehungsweise Tokens wie Slang-Wörter, z.B. „hahaha“, „geil“ und „diggi“, Akronyme wie „LOL“ oder „OMG“ und Web-Links verstanden [73] werden. Li u. a. [123] machte darauf aufmerksam, dass vor allem irrelevante umgangssprachliche Begriffe nicht zwangsläufig über eine auffallend niedrige oder hohe Frequenz verfügen müssen, sondern ihre Vorkommenshäufigkeit teilweise der Frequenz von informativen Wörtern ähneln kann, was es erschwert, diese mit häufigkeitsbasierten Ansätzen vor der Themenmodellierung herauszufiltern.

Darüber hinaus treten nach Churchill und Singh [124] zwei Arten von hochfrequenten Noise Words - kontextfreie und kontextspezifische Noise Words - in Tweets auf, die ebenfalls in Kurznachrichten vorkommen können. Unter kontextfreien Rauschwörtern werden Stoppwörter wie Artikel und Konjunktionen verstanden, die ausschließlich eine grammatikalische Funktion erfüllen [125] und unabhängig vom jeweiligen Kontext als bedeutungslos betrachtet werden [124]. Während diese mit gängigen Stoppwortlisten entfernt werden können, sind kontextspezifische Noise Words schwerer zu detektieren. Zu diesen zählen zum einen Begriffe, die zwar für die besprochenen Themen des Datensatzes relevant sind, bei denen es sich jedoch um sehr allgemeine und wenig aussagekräftige Wörter handelt [124]. Übertragen auf forensische Kurztexte könnten beispielsweise bei einem Fall, der von der finanziellen Unterstützung einer terroristischen Vereinigung handelt, Wörter wie „Geld“ oder „Euro“ kontextspezifische Noise Words darstellen. Darüber hinaus gehören zu den kontextspezifischen Noise Words ebenfalls Begriffe, die für einen bestimmten Datensatz, einen Fachbereich oder Themenbereich nicht von Bedeutung sind, aber dennoch in dem Datensatz eine hohe Frequenz aufweisen und beispielsweise von Akhtar u. a. [126] oder Li u. a. [123] auch als domänenspezifische Stoppwörter bezeichnet werden. Im Kontext von forensischen Kommunikationsdaten könnten zum Beispiel Begriffe wie „Kino“, „schlafen“ oder „essen“, die typisch für alltägliche Smalltalk-Gespräche sind, als kontextspezifische Noise Words aufgefasst werden. Wie von Akhtar u. a. [126] und Churchill und Singh [73] betont wurde, tritt vor allem bei probabilistischen Themenmodellen das Problem auf, dass hochfrequente Noise Words in nahezu jedem Thema eine hohe Wahrscheinlichkeit aufweisen, was die Themen schwer interpretierbar und unterscheidbar macht [127].

Bedeutung der Kontextinformationen und der Metadaten Es muss beachtet werden, dass üblicherweise Kurznachrichten mit einem bestimmten Kontext verbunden sind. Als Kontext können nach Mei und Zhai [128] verfügbare Metadaten eines Dokuments aufgefasst werden. Bezogen auf forensische Kommunikationsdaten können beispielsweise der Absender und Empfänger einer Nachricht beziehungsweise die Gruppe, in der die Nachricht ausgetauscht wurde, sowie zeitliche Informationen Kontexte darstellen. Ferner könnte beispielsweise auch die Stimmung beziehungsweise Emotion hinter einer Nachricht, die automatisch mit einer Sentiment Analyse erkannt werden kann, als Kontext betrachtet werden. Wie von Mei und Zhai [128] betont wird, werden die Themen eines Dokuments beziehungsweise einer Kurznachricht oft durch ihren Kontext beeinflusst. Beispielsweise ist es bei forensi-

schen Kommunikationsdaten, die häufig Gespräche über einen längeren Zeitraum umfassen, realistisch anzunehmen, dass sich die diskutierten Themen über die Zeit verändern. Dementsprechend kann es sinnvoll sein, die Kontextinformationen in die Themenmodellierung mit einzubeziehen.

Erwartungshaltung des Ermittlers an die Themen Der Ermittler hat bei der Themenextraktion aus forensischen Kommunikationsdaten häufig eine gewisse Erwartungshaltung an die Themen. Üblicherweise kennt der Ermittler zumindest den Deliktbereich, von dem sein Fall handelt. Darüber hinaus kann er Vernehmungen oder der Fallakte Informationen über das Tatgeschehen entnehmen [129]. Sein Ziel besteht somit oftmals darin, eine Hypothese zu überprüfen und Beweise zu vermuteten Themen in dem Datensatz zu finden. Ein Problem ist jedoch darin zu sehen, dass, wie in Abschnitt 2.2 erläutert wurde, Algorithmen der Themenmodellierung in ihrer Grundform unüberwacht sind [24]. Diese können gegebenenfalls ein Thema, das für den Ermittler von Interesse ist, nicht identifizieren, falls das gewünschte Thema nur zu einem geringen Anteil in den Daten vertreten ist [130]. Nach Jagarlamudi u. a. [27] und Zuo u. a. [131] sind vor allem probabilistische Themenmodelle von diesem Problem betroffen, da diese darauf abzielen die Wahrscheinlichkeit zu maximieren, dass die beobachteten Dokumente durch den generativen Prozess der Themenmodellierung erzeugt wurden [131]. Nach Zuo u. a. [131] tendieren sie hierdurch dazu, dass sie seltene Themen, die sich in wenigen Dokumenten widerspiegeln, ignorieren [131]. Dieses Problem ist im Kontext von forensischen Kommunikationsdaten besonders ausgeprägt, da irrelevante Smalltalk-Themen häufig die Konversationen im Gegensatz zu eigentlich interessanten, fallrelevanten Themen dominieren [2].

Lückenhafte Kontexte durch verteilte Kommunikation über verschiedene Kanäle Eine weitere Herausforderung, auf die beispielsweise auch Xi u. a. [132] aufmerksam machten, ergibt sich dadurch, dass die mobile Kommunikation über verschiedene Kanäle und Mobilfunkgeräte übertragen wird, was zu segmentierter Information und lückenhaften Kontexten führen kann. Besonders im Bereich von Bandenkriminalität und organisierter Kriminalität können sich fallrelevante Informationen auf den mobilen Endgeräten mehrerer tatverdächtiger Personen befinden, die dementsprechend bei der forensischen Analyse ausgewertet werden müssen [3]. Darüber hinaus findet die mobile Kommunikation meist über verschiedene Messenger-Dienste wie Telegram oder WhatsApp statt [132]. Demzufolge können Informationen über die Planung oder Durchführung einer Straftat über die Nachrichten von mehreren Mobilfunkgeräten und Messenger-Diensten verstreut sein.

Wenn die Themen für jeden Messenger-Dienst auf jedem Gerät einzeln analysiert werden, könnten möglicherweise Zusammenhänge nicht ausreichend erfasst werden und Informationen verloren gehen. Allerdings wäre es ebenfalls nicht ratsam, die Themen aus dem gesamten Datensatz, der die gespeicherten Nachrichten auf allen Mobilfunktelefonen und über alle Messenger-Dienste hinweg beinhaltet, zu extrahieren. Dies liegt daran, dass man dadurch die Gespräche in zahlreichen Chat-Gruppen, die keine Relevanz für den Fall haben, berücksichtigen würde. Somit wäre das Rauschen im Datensatz und das Problem, dass Smalltalk die Gespräche dominiert, besonders schwerwiegend.

Es wäre daher wünschenswert, wenn es eine Möglichkeit gäbe, bei der Extraktion von Themen aus den Daten eines Messenger-Dienstes die Erkenntnisse zu berücksichtigen, die bei der Extraktion von Themen aus einem anderen Messenger-Dienst oder Gerät gewonnen wurden, die im Rahmen der Ermittlungen zu diesem Fall ausgewertet werden.

Variabilität des Vokabulars Als zusätzliche Herausforderung ist zu beachten, dass das verwendete Vokabular in forensischen Kurznachrichten stark von dem Deliktbereich der Straftat sowie dem Bildungsstand und dem sozialen Hintergrund der Tatverdächtigen beziehungsweise im Allgemeinen der Kommunikationsteilnehmer abhängt [133]. Darüber hinaus kann die Wortwahl in einer Kurznachricht ebenfalls davon beeinflusst werden, an wen die Nachricht gerichtet ist [134]. Hinzu kommt, dass sich unabhängig vom forensischen Kontext Einflussfaktoren wie das Alter oder das Geschlecht von Kommunikationsteilnehmern auf das verwendete Vokabular auswirken können [135]. Dementsprechend kann sich das Vokabular in den Kommunikationsdaten verschiedener Fälle, aber auch in den Nachrichten eines Falls in den verschiedenen Chats stark unterscheiden. Mit dieser hohen Variabilität des Vokabulars ist ebenfalls die Herausforderung der Ambiguität verbunden, da Wörter in verschiedenen Kontexten in unterschiedlichen Bedeutungen verwendet können [136].

Mehrsprachige Texte Zuletzt wäre darauf einzugehen, dass die im Rahmen einer forensischen Untersuchung auszuwertenden Datensätze teilweise Kurznachrichten auf verschiedenen Sprachen umfassen. Beispielsweise können bei grenzüberschreitender organisierter Kriminalität wie illegalem Schusswaffenhandel, international organisiertem Drogenhandel oder Geldwäsche [137] Tatverdächtige Nachrichten auf unterschiedlichen Sprachen ausgetauscht haben. Insbesondere für probabilistische Themenmodelle stellen mehrsprachige Datensätze beispielsweise nach Boyd-Graber und Blei [138], Zhang u. a. [139] und Jagarlamudi und Daumé [140] ein Problem dar. Diese sind meist nur in der Lage monolinguale Themen zu extrahieren, bei denen es sich bei den wahrscheinlichsten Wörtern ausschließlich um Begriffe einer Sprache handelt. Nach Jagarlamudi und Daumé [140] hat dies wiederum zur Folge, dass sich die Themenverteilungen semantisch ähnlicher Dokumente verschiedener Sprachen stark unterscheiden. Dies kann im forensischen Kontext problematisch sein, beispielsweise wenn die umfassende Datenmenge auf Chatnachrichten eingeschränkt werden soll, die sich mit einem ausgewählten, fallrelevanten Thema befassen. Der mangelhafte Umgang von probabilistischen Themenmodellen mit mehrsprachigen Datensätzen kann damit begründet werden, dass diese auf dem gemeinsamen Vorkommen von Wörtern auf Dokumentenebene beruhen [139, 140]. Jedoch treten meist innerhalb eines Dokuments, gerade in kurzen Nachrichten, nur die Wörter einer Sprache auf [138]. Wünschenswert wäre es somit, semantisch kohärente Themen extrahieren zu können, bei denen zusammengehörige Wörter eine hohe Wahrscheinlichkeit aufweisen, unabhängig davon, in welcher Sprache sie geschrieben wurden.

2.4 Lösungsansätze für kurze Texte

Um mit der hohen Sparsität der forensischen Kurznachrichten umgehen zu können, kommen Ansätze des sogenannten Short Text Topic Modelling (STTM) infrage, die speziell für die Themenmodellierung in Datensätzen von kurzen Texten entwickelt wurden. Hierfür wurden unterschiedliche Strategien vorgeschlagen, die in Abbildung 2.2 dargestellt werden. Ebenfalls

sind exemplarisch einige Algorithmen aufgeführt, die die jeweilige Strategie anwenden. Es ist an dieser Stelle darauf hinzuweisen, dass mehrere Arbeiten die Integration von Word Embeddings zusammen mit anderen Strategien kombinierten [z.B. 120, 141–143].

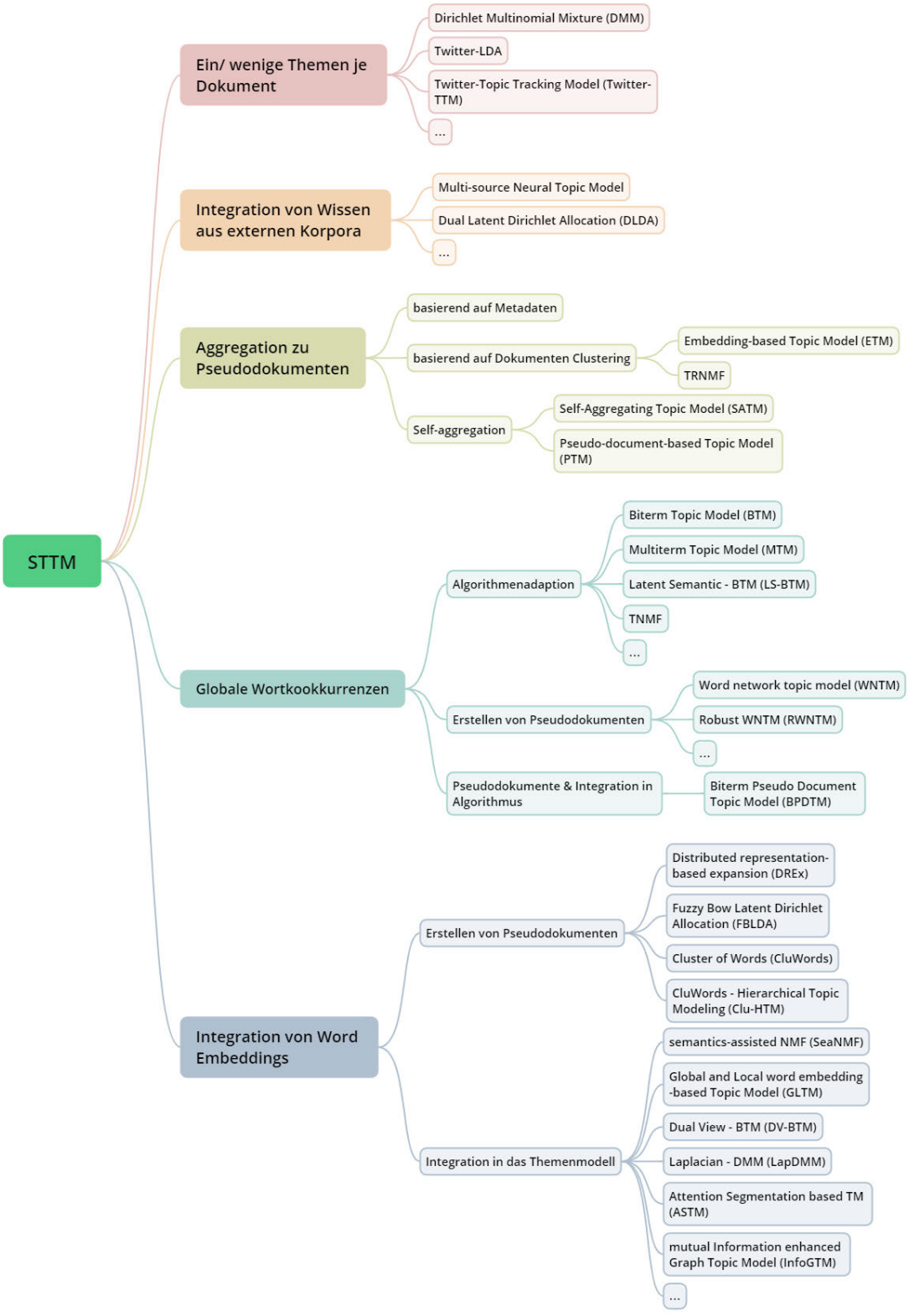

Abbildung 2.2: Übersicht über verschiedene Strategien zum Umgang mit kurzen Texten bei der Themenmodellierung. Fünf Strategien können unterschieden werden, um mit der hohen Sparsität von Datensätzen aus kurzen Texten bei der Themenmodellierung umgehen zu können: der Einsatz von Algorithmen, die davon ausgehen, dass ein Dokument nur aus einem oder wenigen Themen besteht, die Integration von Wissen aus externen Datenquellen, die Aggregation der Texte zu Pseudodokumenten, die Themenmodellierung basierend auf globalen Wortkookkurrenzen und die Einbeziehung von Word Embeddings. Zu jeder Strategie sind gegebenenfalls Unterstrategien und einige Algorithmen, die diese Strategie heranziehen, aufgeführt. Quelle: Eigene Darstellung.

Themenmodelle mit einem Thema je Dokument Eine erste Strategie des STTM besteht darin, anzunehmen, dass sich kurze Texte nur mit einem einzigen Thema befassen [z.B. 144–147]. Als bekanntester Vertreter ist das probabilistische Themenmodelle Dirichlet Multinomial Mixture (DMM), das ursprünglich von Nigam u. a. [144] entwickelt und von Yin und Wang [146] für die Themenextraktion in Kurztexten erweitert wurde, anzuführen. Dieses weist einem Kurztext zunächst aus einer globalen Themenverteilung ein einzelnes Thema zu. Anschließend werden alle Wörter dieses Dokuments aus der Wortverteilung des entsprechenden Themas gezogen [144–146]. Dieser Ansatz kann dem Problem der hohen Sparsität entgegenwirken, da die Information über das gemeinsame Vorkommen von Wörtern in Dokumenten nicht mehr unter mehreren Themen aufgeteilt werden muss [148]. Darüber hinaus wurde von Zhao u. a. [145] speziell für die Themenmodellierung in Tweets die Twitter-LDA entwickelt. Diese unterscheidet sich von DMM insofern, dass das Thema eines Tweets nicht aus einer globalen Themenverteilung für den gesamten Datensatz, sondern aus einer Themenverteilung für den Autor des Tweets gewählt wird.

Li u. a. [60] zeigten jedoch anhand von Beispielen aus Datensätzen mit Fragen und Web Snippets, dass selbst kurze Texte mit einer durchschnittlichen Dokumentenlänge von circa vier beziehungsweise elf Wörtern, sich mit mehreren Themen beschäftigen können. Werden alle Wörter fälschlicherweise demselben Thema zugewiesen, kann dies nach Li u. a. [60] somit zu Informationsverlust führen. Darüber hinaus sind Themenmodelle wie DMM nach Li u. a. [149] mit dem sogenannten „Sensitivity Problem“ konfrontiert. Dieses kommt dadurch zustande, dass das gelernte Thema eines Dokuments hauptsächlich durch die Themengewichte seiner Wörter bestimmt wird. Bei einer geringen Anzahl an Wörtern hat hierbei jedes Wort einen hohen Einfluss auf die Themenzuweisung eines Dokuments. Überwiegen in einem kurzen Text hochfrequente Wörter, die in allen Themen hohe Wahrscheinlichkeiten aufweisen und dementsprechend wenig diskriminativ für Themen sind, kann dies dazu führen, dass das Dokument nicht seinem korrekten Thema zugewiesen wird. Dieses Problem ist bei Kurznachrichten besonders schwerwiegend, da in diesen viele hochfrequente Noise Words wie Slang-Wörter (siehe Abschnitt 2.3.1 und Abschnitt 2.5) enthalten sind.

Um diese Probleme zu adressieren, wurde von Li u. a. [60] eine Erweiterung von DMM entwickelt, bei der sich jeder Kurztext aus einer geringen Anzahl an Themen zusammensetzt. Die genaue Anzahl wurde hierbei für jedes Dokument im Rahmen des generativen Prozesses basierend auf einer Poisson-Verteilung bestimmt. Die Autoren zeigten, dass ihr Poisson DMM (PDMM) signifikant bessere Ergebnisse lieferte als das Standard-DMM unter Verwendung von verschiedenen Evaluierungsmetriken wie der semantischen Kohärenz [150] auf mehreren Datensätzen von kurzen Texten wie Web Snippets.

Integration von Wissen aus externen Korpora Weitere Ansätze des STTM zielten darauf ab, die Themenmodellierung auf Datensätzen von kurzen Texten zu verbessern, indem externe Korpora mit Dokumenten ausreichender Länge mit einbezogen wurden [113, 151, 152]. Beispielsweise wurde von Phan u. a. [113] vorgeschlagen, auf einem möglichst umfassenden Datensatz mit langen Dokumenten zu verschiedensten Themen ein Themenmodell mittels der LDA zu trainieren, um anschließend die Themenverteilung von Kurztexten vorherzusagen. Allerdings lag der Fokus hierbei nur darauf, die Themenabdeckung von Kurztexten zu bestimmen, während dieser Ansatz nicht darauf abzielte, auch eine geeignete TWV für einen Datensatz von Kurztexten zu erhalten. Zudem wurde von Jin u. a. [151] an dieser Vorge-

hensweise kritisiert, dass der externe Datensatz nicht unbedingt für den aktuellen Datensatz thematisch relevant sein muss. Sie entwickelten ein Themenmodell basierend auf der LDA, dass stattdessen beim Training die langen Dokumente nur zusätzlich zu den kurzen Texten berücksichtigt. Hierzu wählten sie möglichst thematisch ähnliche Texte wie beispielsweise Blogs und Webdokumente, auf die URLs in kurzen Texten wie Tweets verwiesen. Der Ansatz von Gupta u. a. [152] unterschied sich insofern von den anderen beiden Arbeiten, dass dieser nicht probabilistische Themenmodelle um das Wissen aus externen Korpora erweiterte, sondern das von ihnen entwickelte Multi-source Neural Topic Model in die Kategorie der DLTMs einzuordnen ist. Durch Regularisierung der Zielfunktion des autoregressiven DLTM DocNADE [52] wurde bezweckt, dass sich das DLTM bei dem Lernen von Themen aus dem Datensatz von Kurztexten an Wortverteilungen von Themen orientierte, die aus Datensätzen von langen Dokumenten extrahiert wurden.

Für die Forensik würde eine Anwendungsmöglichkeit dieser Ansätze darin bestehen, dass als lange Dokumente für den Datensatz von Kurznachrichten beispielsweise die Fallakte oder Vernehmungsprotokolle herangezogen werden. Allerdings ist hierbei zu beachten, dass auf diese Weise eher ein allgemeiner Überblick über den Fall vermittelt wird, als konkret die Themen zu analysieren, die in den Chatnachrichten besprochen wurden. Insbesondere bei dem von Jin u. a. [151] vorgestellten Ansatz würden die Themen hierbei auch Wörter enthalten, die in den Kurznachrichten nicht vorkommen.

Aggregation zu Pseudodokumenten Eine alternative Möglichkeit des STTM ist darin zu sehen, mehrere möglichst ähnliche, kurze Texte zu längeren Pseudodokumenten zu verbinden [45, 119, 143, 153–157]. Vor allem in früheren Arbeiten [45, 153, 154] wurden kurze Texte, genauer gesagt Tweets, basierend auf ihren Metadaten zu Pseudodokumenten aggregiert. Anschließend wurden auf diesen Pseudodokumenten traditionelle Algorithmen der Themenmodellierung wie die LDA angewendet. Beispielsweise wurden alle Tweets, die von demselben Autor [45, 153, 154] oder in demselben Zeitraum [153] verfasst wurden, als ein Pseudodokument betrachtet. Weitere Möglichkeiten bestanden darin, Tweets zusammenzufassen, die über denselben Hashtag [153] verfügten oder in denen dieselben Begriffe vorkamen [45]. Für forensische Kommunikationsdaten würde es sich jedoch als schwierig erweisen, geeignete Metadaten auszuwählen. Eine Möglichkeit bestünde darin, alle Nachrichten, die in demselben Einzel- oder Gruppenchat in einem kurzen Zeitfenster wie einem Tag verfasst wurden, als ein Pseudodokument zu betrachten. Allerdings muss es sich bei diesen nicht zwangsläufig um semantisch kohärente Nachrichten handeln, da die Gespräche an einem Tag nicht unbedingt in Zusammenhang stehen müssen.

Stattdessen könnte, wie von Qiang u. a. [155] und Yi u. a. [156] beschrieben wurde, zur Bildung von Pseudodokumenten ein Dokument Clustering vor der Themenmodellierung durchgeführt werden. Die gebildeten Pseudodokumente dienten erneut als Eingabe für Standard-Verfahren der probabilistischen [155] und algebraischen Themenmodellierung [156].

Jedoch ist nach Quan u. a. [157] ein Nachteil dieser Vorgehensweise darin zu sehen, dass die Kurztexte eines Clusters sich nicht zwangsläufig mit demselben Thema befassen müssen. Daher schlugen sie vor, die Zusammenführung von kurzen Texten zu Pseudodokumenten in den Prozess der Themenmodellierung zu integrieren. Das von den Autoren vorgestellte probabilistische Themenmodell Self-Aggregation based Topic Modelling (SATM) führt einen

zweistufigen Prozess durch. In der ersten Phase werden mit dem zuvor vorgestellten Themenmodell DMM die Kurztexte zu latenten Pseudodokumenten verknüpft, während in der zweiten Phase die eigentliche Themenmodellierung basierend auf der LDA auf diesen Pseudodokumenten durchgeführt wird. Allerdings gilt dieser Algorithmus als rechenintensiv, weshalb von Zuo u. a. [119] das probalistische Themenmodell Pseudo-document-based Topic Modelling (PTM) entwickelt wurde. Dieses lernt im Rahmen des generativen Prozesses in einer einzigen Phase sowohl die Zuordnung der kurzen Texte zu latenten Pseudodokumente als auch die Themen. Hierzu wurde eine zusätzliche Multinomialverteilung von Kurztexten über Pseudodokumente in die gewöhnliche LDA eingeführt. Allerdings zeigten Qiang u. a. [22] bei einer umfassenden Untersuchung bezüglich der Effizienz von verschiedenen Verfahren des STTM, dass PTM zwar eine geringere Laufzeit als SATM aufweist, jedoch mit höheren Laufzeiten und einer schlechteren Konvergenz als Ansätze basierend auf den anderen Strategien des STTM verbunden ist. Zudem machten Qiang u. a. [22] und Liang u. a. [158] darauf aufmerksam, dass SATM und PTM über eine hohe Anzahl an Hyperparametern wie die Anzahl an zu bildenden Pseudodokumenten verfügen, die einen großen Einfluss auf die Qualität der gelernten Themen haben.

Insgesamt sind nach Qiang u. a. [22] und Shi u. a. [44] die Ergebnisse der Themenextraktion bei Verfahren basierend auf der Aggregation von kurzen Texten stark von der Qualität der gebildeten Pseudodokumente abhängig. Enthalten diese irrelevante Texte, können hierdurch weniger bedeutungsvolle Themen generiert werden.

Globale Wortkookkurrenzen Weitere Arbeiten im Bereich des STTM adressierten das Problem des Mangels an zuverlässigen Informationen über das gemeinsame Auftreten von Wörtern in einem kurzen Text, indem sie stattdessen die Themen auf der Grundlage globaler Wortkookkurrenzmuster ableiteten [z.B. 97, 131, 159]. Unter globalen Wortkookkurrenzen wird hierbei verstanden, dass das gemeinsame Vorkommen von Wörtern in einem kurzen Kontext wie dem Kurztext selbst oder einem Sliding Window - einer festgelegte Anzahl an aufeinanderfolgenden Wörtern [131] - im gesamten Korpus untersucht wird. Von Murshed u. a. [23] und Zuo u. a. [131] wurde empfohlen bei kurzen Texten mit weniger als zehn Wörtern das Dokument selbst als Kontext zu betrachten und bei Texten mit einer höheren Wortanzahl ein Sliding Window einzusetzen. Diese Idee wurde sowohl für probabilistische Themenmodelle [97, 131] als auch für algebraische Algorithmen [159] aufgegriffen. Es können zwei grundlegende Vorgehensweisen unterschieden werden: Entweder wurden existierende Algorithmen der Themenmodellierung modifiziert, um Themen basierend auf globalen Wortkookkurrenzen ableiten zu können [z.B. 97, 120, 159–163] oder es wurden zunächst Pseudodokumente gebildet und anschließend traditionelle Verfahren der Themenmodellierung eingesetzt [75, 131, 164].

Das von Yan u. a. [97] vorgestellte BTM als bekanntes probabilistisches Themenmodell des STTM basiert auf der erstgenannten Strategie. Die grundlegende Idee bei BTM und seinen Varianten [120, 160–163, 165–168] besteht darin nicht mehr die Entstehung von Dokumenten, sondern von globalen Wortkookkurrenzen zu modellieren. Hierfür werden zunächst aus dem gesamten Datensatz sogenannte Biterme extrahiert. Unter diesen werden zwei Wörter, die gemeinsam in dem betrachteten Kontext wie dem Kurztext auftreten, verstanden. Anschließend wird im Rahmen des generativen Prozesses für die beiden Wörter ein Thema aus einer globalen Themenverteilung gewählt und die beiden Wörter aus der entsprechen-

den Wortverteilung dieses Themas gezogen. Zu Varianten von BTM zählen beispielsweise die von Yan u. a. [161] und Chen u. a. [165] vorgestellten Ansätze, die Anwendungsfälle für die Themenmodellierung speziell in Tweets adressierten.

Jedoch ist BTM mit zwei Problemen verbunden [160, 166, 168]. Zum einen werden nach Xia u. a. [160] eine große Anzahl von Bitermen extrahiert, von denen viele als irrelevant anzusehen sind, da sie beispielsweise sehr allgemeine Wörter enthalten, die in den Themen nicht erwünscht sind. Dies führt dazu, dass die Effizienz bei der Themenmodellierung sowie die Qualität der extrahierten Themen beeinträchtigt werden kann. Zum anderen wurde von Wu und Li [166] und Zhen u. a. [168] an BTM kritisiert, das die beiden Wörter eines Biterms demselben Thema zugewiesen werden, obwohl zwischen diesen nicht zwangsläufig eine semantische Verbindung bestehen muss. Dieses Problem kann nach Zhen u. a. [168] zu weniger kohärenten Themen führen.

Dem ersten Problem wurde adressiert, indem beispielsweise bei dem Themenmodell Discriminative Biterm Topic Model (d-BTM) möglichst aussagekräftige Biterme basierend auf der Dokumentenfrequenz von Wörtern zur Themenmodellierung extrahiert wurden [160] oder eine Hintergrund-Verteilung eingeführt wurde [165] (siehe Abschnitt 2.5). Um das zweite Problem anzugehen, wurden mithilfe einer Dependenzanalyse nur Wörter, zwischen denen eine starke semantische Beziehung besteht, als Biterm definiert [168]. Ein weiterer Ansatz bestand darin, anstatt Wortpaare Nominalphrasen beliebiger Länge heranzuziehen, da die Autoren davon ausgingen, dass die Wörter einer Nominalphrase stärker miteinander korreliert sind [166].

Als weiteres bekanntes probabilistisches Modell des STTM gilt Word Network Topic Model (WNTM) [131], das den zweitgenannten Ansatz verfolgt, um Themen basierend auf globalen Wortkookkurrenzen zu modellieren. Die grundlegende Idee besteht darin, für jedes Wort des Datensatzes Pseudodokumente zu bilden, die jeweils alle Wörter enthalten, die mit ihm gemeinsam in einem kurzen Text vorkommen. Mit diesen Pseudodokumenten als Input wird die gewöhnliche LDA durchgeführt.

Schließlich wurden von Jiang u. a. [169] die beiden Ansätze von BTM und WNTM kombiniert, indem zunächst ähnlich wie bei WNTM Pseudodokumente basierend auf Wortkookkurrenzen erstellt wurden, die anschließend als Eingabe für BTM dienten.

Integration von Word Embeddings Es muss betont werden, dass die Themenmodellierung basierend auf globalen Wortkookkurrenzen zwar nach Yan u. a. [97] und Wang u. a. [164] geeigneter für Short Text Topic Modelling (STTM) ist, da diese zusätzliche Wortkookkurrenzinformationen explizit mit einbezieht, jedoch setzen diese Ansätze nach Lu u. a. [162] und Li u. a. [170] weiterhin voraus, dass Wörter in den kurzen Texten gemeinsam auftreten müssen, damit diese denselben Themen zugewiesen werden. Jedoch sollten semantisch ähnliche Wörter nach Lu u. a. [162] und Li u. a. [170] eine hohe Wahrscheinlichkeit in einem kohärenten Thema aufweisen, auch wenn ihr gemeinsames Vorkommen in dem Datensatz von kurzen Texten nur sehr gering ist. Daher wurde die semantische Ähnlichkeit von Wörtern, die mithilfe von Word Embeddings ermittelt wurde, im Rahmen von Verfahren des STTM berücksichtigt [z.B. 16, 44, 60, 141, 156, 170, 171].

Ähnlich wie bei der zuvor vorgestellten Strategie wurde hierfür entweder die Eingabe für das Themenmodell verändert [16, 126, 171, 172] oder der Algorithmus zur Themenmodellierung adaptiert [z.B. 44, 156, 158, 173]. Zu der ersten Möglichkeit zählt beispielsweise auch der von Viegas u. a. [16] entwickelte Ansatz CluWords, der in dieser Arbeit ebenfalls angewendet wurde und in Abschnitt 3.4 detaillierter beleuchtet wird, sowie die von Viegas u. a. [171] vorgeschlagene Erweiterung Cluster Hierarchical Topic modelling (CluHTM) und das von Bicalho u. a. [172] vorgestellte Verfahren Distributed Representation-based Expansion (DREx). Die grundlegende Idee dieser Ansätze besteht darin, in die ursprünglichen Kurztexte weitere Wörter des Datensatzes einzufügen, die zu den Wörtern im Kurztext eine hohe Word Embedding Ähnlichkeit aufweisen. Zudem wurde von Akhtar u. a. [126] eine Fuzzy Bag of Word (FBoW) [174] Repräsentation der Kurztexte vorgestellt, die die semantischen Beziehungen basierend auf Word Embeddings zwischen den Wörtern des Datensatzes berücksichtigt. Bei diesen Ansätzen wurden im Anschluss jeweils traditionelle Algorithmen der Themenmodellierung, genauer gesagt NMF bei Viegas u. a. [16], eine Adaption von NMF zur hierarchischen Themenmodellierung bei Viegas u. a. [171] und LDA bei Bicalho u. a. [172] auf den erweiterten Kurztexten durchgeführt. Jedoch haben die von Viegas u. a. [16] und Bicalho u. a. [172] vorgestellten Ansätze den Vorteil, dass sie mit jedem beliebigen Algorithmus der Themenmodellierung kombiniert werden können.

Alternativ wurden Informationen über die semantische Ähnlichkeit direkt in algebraische Themenmodelle wie NMF [44, 156], probabilistische Themenmodelle [z.B. 60, 141, 158] oder DLTMs [53, 173, 175] integriert. Bezüglich der probabilistischen Themenmodelle wurde beispielsweise Collapsed Gibbs Sampling um das sogenannte Generalized Pólya Urn (GPU) Modell [176] erweitert [z.B. 60, 158]. Von Nguyen u. a. [141] und Zhang u. a. [177] wurde für jedes Thema eine Word Embedding Komponente eingeführt. Mit einer binären Indikator-Variable wurde entschieden, ob ein Wort eines Dokuments jeweils mit der gewöhnlichen Themen-Wort-Verteilung oder einer Wortverteilung, die mithilfe der Softmax-Funktion basierend auf den Word Embeddings erstellt wurde, erzeugt wird. Es ist zudem anzumerken, dass neben der LDA probabilistische Themenmodelle um die Word Embedding Ähnlichkeit erweitert wurden, die speziell für STTM entwickelt wurden, wie beispielsweise BTM in [162, 170, 178], WNTM in [142], PTM in [143] und DMM in [141, 149, 170]. Das Ziel bestand somit darin, das Problem der hohen Sparsität möglichst zu vermindern, indem mehrere Strategien kombiniert wurden.

Hinsichtlich der Ansätze basierend auf NMF wurde das Minimierungsproblem, das bei der Zerlegung der TDM gelöst werden muss, so angepasst, dass Wörter nicht nur hohe Werte in dem gleichen Thema aufweisen, wenn sie in demselben Dokument gemeinsam auftreten, sondern auch wenn ihre Word Embeddings über eine hohe Ähnlichkeit verfügen [44, 156]. Um bei Verfahren basierend auf Deep Learning Word Embeddings berücksichtigen zu können, wurden von Zhu u. a. [50] und Ge und Hu [53] graphenbasierte DLTMs speziell für kurze Texte entwickelt. Hierfür wurde der Graph, der dem Prozess der Themenmodellierung der DLTMs zugrunde liegt, basierend auf der Word Embedding Ähnlichkeit erstellt, anstatt Wörter, die häufig gemeinsam auftreten, zu verbinden. Ein andere Herangehensweise wurde von Limwattana und Promon [175] für die Deep Word-Topic Latent Dirichlet Allocation (DWT-LDA) beschrieben, die zunächst mittels einer gewöhnliche LDA ein initiales Themenmodell

erstellten und anschließend die Themenzuweisung für jedes Wort durch ein neuronales Netz aktualisierten. Dabei wurde eine Embedding-Schicht verwendet, um sicherzustellen, dass ähnliche Wörter höhere Wahrscheinlichkeiten für dasselbe Thema aufweisen.

Die einzelnen Ansätze unterscheiden sich neben der genauen Vorgehensweise zur Berücksichtigung der semantischen Ähnlichkeit darin, auf welchem Korpus die Word Embeddings trainiert wurden. In einigen Arbeiten [z.B. 60, 141, 143, 162, 170] wurden vortrainierte Word Embeddings herangezogen, die auf großen, externen Korpora, beispielsweise bestehend aus Google Nachrichten im Falle von word2vec [63] oder Wikipedia Artikeln und Inhalte von Websites bezüglich fastText [179] trainiert worden waren. Diese Word Embeddings ermöglichen es Informationen über die semantische Ähnlichkeit von Wörtern im allgemeinen Sprachgebrauch zu erlangen [60]. Wie jedoch beispielsweise von Liang u. a. [158], Shi u. a. [44] und Liu u. a. [120] betont wird, treffen die auf den externen Korpora gewonnenen Informationen über die Wortähnlichkeit nicht zwangsläufig auf den Trainingsdatensatz der Themenmodelle zu. Vor allem im Kontext von forensischen Kommunikationsdaten kann dieses Problem schwerwiegend sein. Dies kann zum einen darauf zurückgeführt werden, dass sich die Chatnachrichten bezüglich der sprachlichen Struktur und des verwendeten Vokabulars stark von anderen Dokumentenarten wie Zeitungsartikeln unterscheiden [133, 134]. Zum anderen werden in forensischen Kurznachrichten teilweise Hidden Semantics verwendet, worunter verstanden wird, dass Wörter nicht in ihrer üblichen Bedeutung verwendet werden [133], weshalb Erkenntnisse aus externen Korpora für diese nicht mehr geeignet sind.

Ein Lösungsansatz besteht darin, wie beispielsweise in der Arbeit von Shi u. a. [44] und Jiang u. a. [142], stattdessen die Word Embeddings auf demselben Trainingsdatensatz wie das Themenmodell zu trainieren. Allerdings benötigen nach Qiu u. a. [180] Word Embeddings ausreichend große Korpora, die bei vortrainierten Word Embeddings teilweise mehrere Millionen Wörter umfassen [181], sodass nach Liu u. a. [120] und Li u. a. [60] die durch lokal trainierte Word Embeddings erfasste semantische Ähnlichkeit dem Problem der Sparsität nicht ausreichend entgegengewirkt wird. Daher wurde von Liang u. a. [158] sowie von Liu u. a. [120] empfohlen als semantisch ähnliche Begriffe diejenigen Wörter zu betrachten, die sowohl auf dem Trainingsdatensatz des Themenmodells als auch nach auf einem externen Datensatz erlernten Word Embeddings eine hohe Ähnlichkeit aufweisen. Zur Bestimmung der semantischen Ähnlichkeit auf dem Trainingsdatensatz können beispielsweise, wie von Liu u. a. [120] vorgeschlagen wurde, statt Word Embeddings Kookkurrenzen zweiter Ordnung eingesetzt werden.

Von Wang u. a. [182] wurde hervorgehoben, dass vor allem die Art, wie die Word Embeddings für das STTM genutzt werden, entscheidend dafür ist, ob sich inkorrekte Informationen über die semantische Bedeutung von Wörtern negativ auf die Themen auswirken. Sie schlugen das Themenmodell Attention Segmentation based TM (ASTM) vor, bei dem die Word Embeddings nicht direkt zum Einsatz kamen, um semantisch ähnliche Wörter demselben Thema zuzuweisen. Stattdessen wurde zunächst eine gewöhnliche LDA durchgeführt, um anschließend vortrainierte Word Embeddings und die erhaltenen Themen-Wort-Verteilungen zu kombinieren, um die Bedeutung von Wörtern zu einem Thema zu bestimmen.

Vergleichende Untersuchungen Qiang u. a. [22] und Murshed u. a. [23] führten beide umfassendere Experimente durch, um verschiedene Strategien des STTM - die Aggregation zu Pseudodokumenten, Themenmodelle, die jedem Dokument nur ein Thema zuordnen, die Berücksichtigung von globalen Wortkookkurrenzen und die Integration von Word Embeddings - zu vergleichen. Die Evaluierung erfolgte in beiden Studien auf gelabelten Datensätzen basierend auf der Dokumentenklassifikation, wobei die Themenverteilung der Kurztexte als Feature-Repräsentation für eine Support Vector Machine (SVM) herangezogen wurde, sowie mit Maßen zur Evaluierung des Dokument-Clusterings, nämlich der Purity [183] und dem Normalized Mutual Information (NMI) [184]. Auf diese Evaluierungsmethoden wird in Abschnitt 2.11 genauer eingegangen.

Murshed u. a. [23] führte die Evaluierung auf zwei selbst erstellten annotierten Datensätzen von Tweets über verschiedene Pandemien und unterschiedliche Arten von Beleidigungen und Mobbing durch. Sie kamen zu dem Resultat, dass BTM und WNTM, die die Themen aus globalen Wortkookkurrenz-Mustern ableiten, bezüglich aller Metriken und beider Datensätze überwiegend bessere Resultate erzielten als Verfahren, die automatisch die Pseudodokumente während der Themenmodellierung bilden wie SATM und PTM. Insbesondere SATM brachte durchgehend niedrige Resultate hervor. Dennoch konnten mit SATM in allen Metriken bessere Ergebnisse erzielt werden als mit traditionellen Ansätzen für lange Dokumente wie der LDA und NMF. Ebenfalls übertraf SATM in allen Experimenten die Twitter-LDA, die auf der Annahme beruht, dass sich ein kurzes Dokument nur aus einem einzigen Thema zusammensetzt. Vielversprechende Ergebnisse, die vergleichbar mit den Ansätzen basierend auf globalen Kookkurrenzen waren, konnten hingegen durch Verfahren des STTM hervorgebracht werden, die Word Embeddings integrierten [158].

Diese Resultate wurden im Wesentlichen von der von Qiang u. a. [22] durchgeführten Studie bestätigt. Im Gegensatz zu Murshed u. a. [23] dienten als Datengrundlage für die Experimente nicht nur Tweets, sondern weitere Formen von kurzen Texten wie Bildunterschriften [185] und Google News Schlagzeilen, wobei alle sechs Datensätzen jeweils mit Themen annotiert waren. Bezüglich aller Metriken lieferten die Themenmodelle, die auf der automatischen Aggregation von Kurztexten beruhten, schlechtere Resultate als andere Strategien des STTM. Erneut brachte insbesondere SATM niedrige Ergebnisse hervor, die zudem auf allen Datensätzen unter den Resultaten von Standard-Verfahren der Themenmodellierung, genauer gesagt der LDA, lagen. Bessere Resultate wurden wiederum mit probabilistischen Themenmodellen basierend auf globalen Wortkookkurenzen, BTM und WNTM, erreicht, wobei die Ergebnisse jedoch stark von dem jeweiligen Datensatz abhingen. Im Gegensatz zu der Studie von Murshed u. a. [23] wurde je nach Datensatz und Evaluierungsmethode DMM als vielversprechendstes Modell betrachtet. Dieses geht wie die Twitter-LDA, die bei den Experimenten von Murshed u. a. [23] schlechtere Ergebnisse als andere Verfahren des STTM hervorbrachte, davon aus, dass jedes Dokument nur aus einem Thema besteht. Jedoch muss angemerkt werden, dass auf dem Datensatz von Tweets DMM in allen Metriken die zweitschlechtesten Ergebnisse nach SATM hervorbrachte. Die Tatsache, dass DMM und die Twitter-LDA bei den beiden Untersuchungen von Qiang u. a. [22] und Murshed u. a. [23] auf informellen Texten geringe Resultate hervorbrachten, könnten auf das zuvor beschriebene Problem zurückgeführt werden, dass die Qualität dieser Ansätze stark von einer hohen Anzahl an Noise Words in den Kurztexten beeinträchtigt wird. Darüber hinaus stellten Qiang u. a. [22] wie bereits Murshed u. a. [23] fest, dass die Integration von vortrainierten Word Embeddings erfolgreich

dem Problem der Sparsität entgegenwirken kann. Bei der Evaluierung basierend auf der Dokumentenklassifikation wurden diese jedoch vor allem als nützlich betrachtet, wenn die Word Embeddings auf thematisch ähnlichen Datensätzen wie das Themenmodell trainiert wurden.

2.5 Lösungsansätze für verrauschte Texte

Das Problem, dass die Kurznachrichten über eine mangelhafte sprachliche Qualität und insbesondere über eine hohe Anzahl von orthographischen Fehlern verfügen, kann, wie insbesondere von Gorro u. a. [186] sowie Churchill und Singh [187] betont wird, durch eine umfassende Vorverarbeitung adressiert werden. Beispielsweise besteht nach Al-Ani und Fasli [122] eine Möglichkeit, mit Rechtschreibfehlern und Wortwiederholungen umzugehen, darin, vor der Themenmodellierung die verschiedenen teilweise fehlerhaften Schreibweisen von Wörtern unter Verwendung des Approximate String Matching Algorithmus [188] auf ein Wort abzubilden.

Eine größere Herausforderung ist hingegen in den in Abschnitt 2.3.2 beschriebenen kontextspezifischen Noise Words zu sehen. Wie in Abschnitt 2.2 und in Abschnitt 2.3.2 erläutert wurde, werden vor allem die Themen von probabilistischen Ansätzen der Themenmodellierung im Gegensatz zu Embedding-basierten Verfahren [z.B. 62, 65] häufig von hochfrequenten Begriffen dominiert [124, 126]. Dementsprechend konzentrierten sich mehrere bisherige Arbeiten darauf, probabilistische Themenmodelle wie die LDA für einen robusteren Umgang mit Noise Words zu adaptieren.

Ein Ansatz, der beispielsweise von Li u. a. [127], Wilson und Chew [189] und Yang u. a. [190] vorgeschlagen wurde, bestand darin, zusätzlich zur gewöhnlichen Entfernung von Stoppwörtern Termgewichtungsschemen in Inferenz-Methoden der LDA zu integrieren. Sowohl die von Wilson und Chew [189] entwickelte weighted Latent Dirichlet Allocation (wLDA) als auch die von Li u. a. [127] beschriebene Combined Entropy Weighting-Latent Dirichlet Allocation (CEW-LDA) zielen darauf ab, informative Wörter bei der Themenextraktion zu belohnen, wozu Maße der Informationstheorie bei der Durchführung von Collapsed Gibbs Sampling [47] berücksichtigt werden. Wilson und Chew [189] gewichteten basierend auf der Pointwise Mutual Information (PMI) [191] Wörter ab, die in sehr vielen Dokumenten auftreten, während Li u. a. [127] die Conditional Entropy [192] heranzogen und Wörter als informativ betrachteten, die mit wenigen anderen Begriffen gemeinsam in einem Dokument vorkommen. Von Yang u. a. [190] wurde hingegen vor allem an der von Wilson und Chew [189] vorgeschlagenen Vorgehensweise kritisiert, dass es weniger bedeutend wäre, in wie vielen Dokumenten die Wörter enthalten sind, sondern Termgewichtungsschemen für den Einsatz in der Themenmodellierung Wörter basierend darauf gewichten sollten, inwiefern diese zur Unterscheidung von Themen beitragen. Um zu erreichen, dass diskriminative Wörter ein höheres Gewicht erhalten und Wörter, die in nahezu allen Themen mit einer hohen Wahrscheinlichkeit vorkommen, abgewichtet werden, wurde in Gibbs Sampling [193] ein überwachtes Termgewichtungsschema [194] mit einbezogen. Dieses basiert jedoch auf den Ergebnissen einer vorherigen Durchführung eines unüberwachten Themenmodells wie der Standard-LDA [190], sodass bei diesem Ansatz das Training von zwei Themenmodellen vorausgesetzt wird, was bei größeren Datensätzen mit einer hohen Laufzeit einhergehen kann.

Eine Alternative kann in der Einführung einer Hintergrundverteilung in den generativen Prozess von probabilistischen Themenmodellen gesehen werden [z.B. 123, 124, 160, 195, 196]. Es ist an dieser Stelle darauf hinzuweisen, dass die Hintergrundverteilung teilweise auch als Rauschverteilung [124] oder allgemeines Thema [123] bezeichnet wird. Die Hintergrund-Verteilung dient dazu, Noise Words zu bündeln [123], die in dieser eine hohe Wahrscheinlichkeit erhalten [195]. Da probabilistische Themenmodelle die Maximierung der Likelihood bezwecken, werden in den eigentlichen Themen die Wörter bevorzugt, die durch die Hintergrundverteilung nicht erklärt werden, sodass in diesen bedeutsame Begriffe im Gegensatz zu Noise Words höhere Wahrscheinlichkeiten zugewiesen bekommen [5, 195]. Solche Hintergrundverteilungen wurden in weit verbreitete probabilistische Algorithmen wie in die PLSA durch Mei und Zhai [195] und die LDA, beispielsweise bei dem von Chemudugunta u. a. [196] beschriebenen Algorithmus Special Words with Background (SWB) integriert.

Da es sich bei verrauschten Texten wie Chats oder Posts sozialer Medien meist ebenfalls um Texte geringer Länge handelt [123], widmeten sich weitere Arbeiten [123, 147, 149, 160, 165] konkret der Erweiterung von probabilistischen Themenmodellen um Hintergrundverteilungen, die speziell für kurze Texte entwickelt wurden. Li u. a. [123] und ferner Sasaki u. a. [147] und Li u. a. [149] adressierten hierbei das in Abschnitt 2.4 beschriebene Problem, dass gerade probabilistische Themenmodelle, die jedem Kurztext nur ein Thema zuweisen, unzureichend mit Noise Words umgehen können. Li u. a. [123] zeigten, dass das Hinzufügen von mehreren Hintergrundverteilungen zu DMM kohärente Themen, gemessen mit dem Normalized Pointwise Mutual Information (NPMI) [197], auf einem Datensatz aus zwei Millionen Tweets hervorbrachte als das gewöhnliche DMM. Interessanterweise wurde hingegen keine höhere Themenkohärenz erzielt, wenn die Noise Words mithilfe von Vorverarbeitungsschritten wie der Entfernung von Wörtern mit einer sehr hohen und sehr niedrigen Dokumentenfrequenz oder basierend auf dem Algorithmus TextRank [198] herausgefiltert wurden. Diese Ergebnisse wurden qualitativ durch die Analyse von ausgewählten Themen durch Li u. a. [123] bestätigt und verdeutlichen den Nutzen der Hintergrundverteilungen.

Speziell für stark verrauschte Texte wurde von Churchill und Singh [124] vorgeschlagen, dass die Rauschfilterung durch Hintergrundverteilungen noch weiter verbessert werden kann, indem Word Embeddings integriert werden. Bei der von ihnen vorgeschlagenen Adaption der LDA, dem Topic-Noise Discriminator (TND), wurde ähnlich wie in Abschnitt 2.4 Collapsed Gibbs Sampling adaptiert, um zu bezwecken, dass Wörter mit einer hohen Word Embedding Ähnlichkeit im selben Thema gruppiert werden. Im Kontext der Rauschfilterung ist die Integration von Word Embeddings vor allem vorteilhaft, da die Zuweisung eines Noise Words zur Hintergrundverteilung dazu führt, dass weitere semantisch ähnliche Noise Words dieser Verteilung ebenfalls zugeordnet werden. Dies hat zur Folge, dass die Noise Words aus den eigentlichen Themen herausgefiltert werden. Jedoch kamen Churchill und Singh [124] in einer qualitativen Evaluierung auf einem großen Twitter-Datensatz zu dem Ergebnis, dass TND Themen hervorbrachte, die für menschliche Annotatoren kaum interpretierbar waren und zudem überwiegend nicht mit bekannten Groundtruth-Themen des Datensatzes übereinstimmen. Hingegen war es möglich, menschlich interpretierbare Themen zu erzielen, indem nur die Hintergrundverteilung von TND herangezogen wurde, um Noise Words aus Themen eines traditionellen Themenmodells wie der LDA herauszufiltern. Dementsprechend ist der von Churchill und Singh [124] vorgeschlagene Ansatz insbesondere bei der Kombination mit anderen probabilistischen Themenmodellen als erfolgversprechend zu betrachten.

Darüber hinaus wurde von Churchill und Singh [73] ein graphenbasiertes Themenmodell speziell für rauschbehaftete Texte entwickelt. Die grundlegende Idee bestand wie bei den in Abschnitt 2.2 beschriebenen Ansätzen in der Errichtung eines Termkookkurrenzgraphens. Anschließend wurden schrittweise Kanten zwischen Wörtern entfernt, die nur in wenigen Dokumenten gemeinsam auftraten, da davon ausgegangen wurde, dass mindestens eines der beiden Wörter als Noise Word zu betrachten war. Die auf diese Weise entstandenen Teilgraphen wurden mithilfe einer adaptierten Clique Percolation Methode [199] zu Themen verbunden. Dieser Ansatz war jedoch mit zwei grundlegenden Problemen verbunden: Zum einen variierten die resultierenden Themen stark in der Größe und umfassten teilweise nur fünf Begriffe [73]. Zum anderen wurde durch diesen graphenbasierten Ansatz zwar die Themenanzahl automatisch bestimmt, diese lag aber teilweise weit unter der vermuteten Anzahl an Themen, was auf einen Verlust von Themen hindeutet [200]. Um das erste Problem zu adressieren, schlugen Churchill und Singh [73] vor, die entstandenen Themen mithilfe von Word Embeddings durch weitere semantisch ähnliche Begriffe zu ergänzen, was jedoch wiederum zu einem Anstieg von Noise Words in den Themen führte [73, 200].

Schließlich wurde von Williams u. a. [201] ein Ansatz entwickelt, der darauf abzielt, konkret das Rauschen der Beiträge von sozialen Netzwerken für die Themenmodellierung anzugehen. Dieser unterscheidet sich von den bisher vorgestellten Ansätzen insofern, dass nicht Noise Words abgewichtet oder entfernt wurden, sondern der Datensatz zur Vermeidung von Rauschen vor der Themenmodellierung nach relevanten Dokumenten gefiltert wurde. Jedoch kam die hierbei angewendete Vorgehensweise nicht für die Anwendung auf forensischen Kommunikationsdaten infrage, da sie auf der Netzwerkstruktur von sozialen Plattformen und spezifischen Nutzerinteraktionen, wie beispielsweise dem Retweeten für das soziale Netzwerk Twitter, basierte.

2.6 Lösungsansätze für die Bedeutung des Kontextes und der Metadaten

Mehrere Arbeiten [z.B. 128, 202] zielten darauf ab, bisherige Themenmodellierungsalgorithmen weiterzuentwickeln, um zu berücksichtigen, dass der Kontext eines Dokuments beziehungsweise einer Nachricht von Bedeutung sein kann. Hierbei wurden meistens spezifische Kontextarten in den Prozess der Themenmodellierung einbezogen [z.B. 203–205]. Eine Übersicht über die Art des Kontextes, Algorithmen, die diese Kontextart integrierten sowie der Basisalgorithmus, der für diese adaptiert wurde, sind in Tabelle 2.5 aufgeführt.

Wie in Tabelle 2.5 zu sehen ist, wurden die Kontextinformationen vor allem in probabilistische Algorithmen wie die PLSA [z.B. 128, 202, 216] und die LDA [z.B. 29, 217] integriert. In wenigen Arbeiten [208, 209] wurden zudem Informationen über den Kontext in DLTMs einbezogen. Einige der in Tabelle 2.5 genannten Kontextarten wie das Journal [z.B. 209], der Ort [211] oder die Kollektion [205] wurden für spezielle Anwendungsfälle wie der Themenmodellierung in wissenschaftlichen Artikeln und in Posts sozialer Medien oder zur Themenextraktion aus mehreren Dokumentensammlungen, beispielsweise zur Genre-Analyse, herangezogen. Für den forensischen Bereich ist hingegen die Zeit als Kontext von höherer Bedeutung. Themenmodelle wie Dynamic Topic Model (DTM) [203] und Topics over Time (TOT) [206] sind besonders geeignet, um die Veränderung der Prävalenz von Themen im Laufe der Zeit zu

Tabelle 2.5: Überblick über verwendete Kontextarten bei der Themenmodellierung. Dargestellt sind die verschiedenen Arten von Kontexten, die bei bisherigen Ansätzen in Algorithmen der Themenmodellierung integriert wurden. Die Art des Kontextes, Beispiele für Algorithmen, die diesen Kontext berücksichtigten und der Basisalgorithmus, der hierfür erweitert wurde, werden angegeben.

Kontextart	**Algorithmus**	**Basisalgorithmus**
Zeit	DTM [203]	LDA
	TOT [206]	LDA
	VF-DTM [207]	LDA
Autor	ATM [102]	LDA
Sender, Empfänger	ART [204]	LDA
	RART [204]	LDA
Autor, Zeit	OSDATM [208]	NTM
Autor, Journal (Zeit, Fachkreis)	VGATM [209]	NTM
	CTM [210]	LDA
	DCTM [210]	LDA
Zeit, Ort	Adaptierte sLDA [211]	LDA
Kollektion	ccLDA [205, 212]	LDA
Netzwerk	RTM [213] und CRTM [214]	LDA
	GPGBN [30]	NTM
	netPLSA [215]	PLSA
	LIMTopic [216]	PLSA
	copulaPLSA [202]	PLSA
	LC-LDA [217]	LDA
	STM [218]	LDA
Generisch	CPLSA [128]	PLSA
	mLDA [29]	LDA

untersuchen und lassen somit Rückschlüsse auf einen Zeitraum zu, in dem ein fallrelevantes Thema intensiv diskutiert wurde. Darüber hinaus wurden speziell für die Themenanalyse in E-Mails, Posts sozialer Medien und Kurznachrichten die Themenmodelle Author-Recipient-Topic Model (ART) [204] und Role-Author-Recipient-Topic Model (RART) [204] entwickelt, die als Kontext den Sender und Empfänger mit einbeziehen. Jedes Sender-Empfänger-Paar wird mit einer Themenverteilung assoziiert, wobei ein Thema, wie bei der gewöhnlichen LDA, als Verteilung von Wörtern dargestellt wird. Wie McCallum u. a. [204] zeigten, ermöglichte dieser Algorithmus, Kommunikationsteilnehmer zu ermitteln, die sich vor allem über ein bestimmtes (fallrelevantes) Thema unterhielten.

Zudem wurden von mehreren Arbeiten Netzwerke wie Zitationsnetzwerke oder soziale Netzwerke als Kontext berücksichtigt [30, 202, 213–218]. Hierbei können zwei verschiedene Ansätze unterschieden werden: Zum einen wurden Algorithmen wie Relational Topic Model (RTM) [213], Constrained Relational Topic Model (CRTM) [214] und Graph Poisson Gamma

Belief Network (GPGBN) [30] entwickelt, die im Rahmen der Themenmodellierung Verbindungen zwischen Dokumenten erfassen, um für neue, ungesehene Dokumente ausgehend von den in ihnen diskutierten Themen eine Vorhersage über ihre Verknüpfungen zu anderen Dokumenten treffen zu können. Zum anderen wurden sogenannte *Regularized Topic Models* wie beispielsweise die netPLSA [215], Link Importance based Topic Model (LIMTopic) [216] und die copulaPLSA [202] vorgeschlagen, die Einschränkungen auf die Themenverteilungen von Dokumenten auferlegen. Die meisten Ansätze [215–218] gingen hierbei davon aus, dass verknüpfte Dokumente eine ähnliche Themenverteilung aufweisen sollten. Bei forensischen Kommunikationsdaten wäre es jedoch schwierig, eine geeignete Netzwerkstruktur als Kontext zu identifizieren. Kommunikationsnetzwerke als Kontext, bei denen Nachrichten zwischen Sender und Empfänger verknüpft werden, sind bei Kommunikationsdaten im Gegensatz zu sozialen Netzwerken nicht zielführend, da bei der Auswertung eines einzelnen Mobilfunkgerätes entweder der Empfänger oder der Sender einer Nachricht immer der Besitzer des Mobilfunkgerätes ist. Eine Möglichkeit wäre zum Beispiel die Verknüpfung von Nachrichten durch Erwähnungen mittels des @-Zeichens. Es müsste jedoch untersucht werden, bei welchen Messenger-Diensten dieses zum Einsatz kommt und ob die Erwähnungen in Gesprächen häufig genug verwendet werden, um sinnvolle thematische Verknüpfungen herstellen zu können.

Von besonderem Interesse sind hingegen die von Mei und Zhai [128] vorgestellte Contextual Probabilistic Latent Semantic Analysis (CPLSA) und die von Tang u. a. [29] beschriebene Multi-contextual LDA (mLDA). Diese können als Verallgemeinerung der anderen Algorithmen angesehen werden, da diese es ermöglichen, beliebige Kontextarten zu integrieren. Hierzu führten Mei und Zhai [128] Kontextvariablen in die PLSA und Tang u. a. [29] in die LDA ein, wodurch kontextabhängige Sichten auf ein Thema eingenommen werden konnten. Durch den jeweiligen Kontext wurden sowohl die Wortverteilungen der Themen als auch die Themenverteilung der Dokumente beeinflusst. Darüber hinaus schlug Tang u. a. [29] speziell für kurze Texte ein *Co-Regularization Framework* vor. Sie adressierten hierbei das Problem, dass, wie von Tang u. a. [29] ausführlicher erläutert wird, durch die Wahl einer Kontextart im generativen Prozess bei der CPLSA und der mLDA die hohe Sparsität in Datensätzen von kurzen Texten noch verstärkt wird. Die wesentliche Idee bestand darin, dass die verschiedenen Kontextarten basierend auf einem zentroidbasierten Regularisierungsschema miteinander kooperieren. Hierzu wurde für jede Kontextart Pseudodokumente gebildet und auf diesen kontextspezifische Themen erlernt. Zusätzlich zu den kontextspezifischen Themen wurden sogenannte Konsensthemen eingeführt, die bei verschiedenen Kontextarten auftraten und somit als robust betrachtet wurden. Bei dem Prozess der Themenmodellierung sollte der Abstand zwischen den sichtspezifischen Themen und den Konsensthemen möglichst verringert werden.

2.7 Lösungsansätze für die Erwartungshaltung des Ermittlers an die Themen

Um Themen extrahieren zu können, die für den Ermittler von besonderem Interesse sind, können Ansätze gewählt werden, die das Vorwissen des Ermittlers über die Themen mit einbeziehen. Eine erste Möglichkeit besteht darin, überwachte Ansätze einzusetzen, zu denen

beispielsweise probabilistische Themenmodelle wie die Labelled-LDA (L-LDA) [28] und die Supervised LDA (sLDA) [219] zählen. Für das Training dieser Modelle werden jedoch annotierte Datensätze benötigt, bei denen die einzelnen Dokumente mit ihrem bekannten Thema gelabelt sind. Ein denkbarer Ansatz, um einen annotierten Datensatz zu erstellen, würde darin bestehen, möglichst viele Falldaten aus verschiedenen Deliktbereichen zu sammeln und das bekannte Delikt als Themenlabel zu betrachten. Allerdings müssten hierfür zunächst rechtliche Fragen geklärt werden, inwiefern die Zusammenführung von Daten aus verschiedenen Fällen in Deutschland zulässig ist.

Interessanter sind hingegen halbüberwachte Algorithmen [z.B. 6, 14, 220, 221], die sich vor allem in der Art des Nutzerinputs unterscheiden, den sie benötigen. Exemplarisch sind einige halbüberwachte Algorithmen der Themenmodellierung in Tabelle 2.6 sowie der benötigte Nutzerinput aufgeführt. Ebenfalls ist die in Abbildung 2.1 eingeführte Kategorie von Themenmodellen, in die der Algorithmus einzuordnen ist, sowie gegebenenfalls der Basisalgorithmus, der adaptiert wurde, um Vorwissen zu integrieren, in Tabelle 2.6 angegeben. Es ist zu beachten, dass zwei verschiedene Strategien, nämlich das Targeted Topic Modelling und das Seed-Guided Topic Modelling, existieren, um Vorwissen im Sinne von wenigen Begriffen zu integrieren.

Tabelle 2.6: Überblick über Formen des Nutzerinputs bei verschiedenen Ansätzen der halbüberwachten Themenmodellierung. Neben der Form des Nutzerinputs ist eine Auswahl an halbüberwachten Algorithmen, die diesen Ansatz nutzen, der Basisalgorithmus, der hierfür angepasst wurde und die Kategorie, in die der Algorithmus einzuordnen ist, angegeben. Hierfür wurden die in Abschnitt 2.2 eingeführten Kategorien herangezogen. Für die Embedding-basierten Ansätze wurde kein Algorithmus adaptiert, sondern eine neue Definition für Themenmodelle eingeführt.

Nutzerinput	**Algorithmus (Auswahl)**	**Basisalgorithmus**	**Kategorie**
Wenige gelabelte Dokumente	HSPLSA [220]	PLSA [41]	probabilistisch
	HSLDA [220]	LDA [4]	probabilistisch
Feedback (Interaktives Topic Modelling)	TopicSifter [221]	NMF [32]	algebraisch
	UTOPIAN [36]	NMF [32]	algebraisch
	Ansatz von El-Assady u. a. [222]	LDA [4]	probabilistisch
Must-Links & Cannot-Links	DF-LDA [223, 224]	LDA [4]	probabilistisch
	MC-LDA [225]	LDA [4]	probabilistisch
Einziger Begriff (Targeted Topic Modelling)	TTM [6]	LDA [4]	probabilistisch
	TATM [226]	DMM [144] und BTM [97]	probabilistisch
	TAM [227]	DMM [144]	probabilistisch
	BiTTM [228]	BTM [97]	probabilistisch
	APSUM [7]	DMM [144]	probabilistisch
	HFTM [229]	LDA [4]	probabilistisch
Wenige Begriffe (Seed-Guided Topic Modelling)	SeededLDA [27]	LDA [4]	probabilistisch
	keyATM [14]	LDA [4]	probabilistisch
	SMTM [230]	LDA [4]	probabilistisch
	ISLDA [12]	LDA [4]	probabilistisch
	user-orientedLDA [231]	LDA [4]	probabilistisch
	GTM [130]	LDA [4]	probabilistisch
	SeededLDA [15, 232]	LDA [4]	probabilistisch
	KeyETM [233]	ETM [62]	Embedding-basiert
	RMATM [234]	ATM [51]	DLTM
	CatE [235]	-	Embedding-basiert
	SeaTopic [236]	-	Embedding-basiert
	SeedTopicMine [237]	-	Embedding-basiert
	Anchored CorEx [90]	CorEx [90]	Weitere

Wie Tabelle 2.6 entnommen werden kann, benötigen einige halbüberwachte Algorithmen wie die Hybrid Semi-supervised PLSA (HSPLSA) und die Hybrid Semi-supervised LDA (HSLDA) [220] weiterhin zumindest eine geringe Menge an **gelabelten Dokumenten**. Sind keine annotierten Daten verfügbar, kommen stattdessen beispielsweise Ansätze der **interaktiven Themenmodellierung** in Betracht, die in einem iterativen Prozess dem Nutzer die extrahierten Themen präsentieren und sein Feedback über die Relevanz der Themen entgegennehmen [36, 221, 222]. Das Feedback kann beispielsweise erfolgen, indem der Nutzer die Themen mithilfe eines visuellen Analysesystems direkt als relevant oder irrelevant bewertet [221] oder indem er von mehreren erstellten Themenmodellen dasjenige auswählt, das mehr interessante Themen enthält [222]. Im Anschluss wird erneut ein Themenmodell trainiert, bei dem bezweckt wird, dass die Themen möglichst viele der wahrscheinlichsten beziehungsweise am höchsten gewichtesten Wörter der Themen enthalten, die der Nutzer als relevant eingestuft hat [221, 222]. Jedoch ist dieser „Human-in-the-Loop Prozess“ mit einem hohen Arbeitsaufwand für den Ermittler verbunden und geht zudem mit hohen Laufzeiten durch das mehrmalige Training der Themenmodelle einher.

Relevanter sind Ansätze, die minimalen Nutzerinput in Form von einzelnen Begriffen [z.B. 6, 14, 15] oder Wortpaaren [z.B. 223, 224, 227] benötigen. Das Ziel von letztgenannten besteht darin, das Wissen des Nutzers über thematische Beziehungen zwischen Wörtern bei der Themenmodellierung zu berücksichtigen [223–225]. Beispielsweise wurden hierfür von Andrzejewski u. a. [223] sogenannte **Must-Links** und **Cannot-Links** in die LDA integriert, indem die gewöhnliche Dirichlet a-priori-Verteilung durch eine Dirichlet Forest a-priori-Verteilung ersetzt wurde. Unter einem Must-Link werden nach Andrzejewski u. a. [223] zwei Wörter verstanden, die in Themen mit ähnlichen Wahrscheinlichkeiten auftreten sollten, während ein Cannot-Link ein Paar von Wörtern bezeichnet, die nicht in demselben Thema mit einer hohen Wahrscheinlichkeit vorkommen dürfen [238]. Die von Andrzejewski u. a. [223] vorgeschlagene Dirichlet Forest-Latent Dirichlet Allocation (DF-LDA) wurde von Kobayashi u. a. [224] erweitert, um logische Verknüpfungen wie beispielsweise Disjunktionen zwischen Must-Links und Cannot-Links einbeziehen zu können. Statt Wortpaaren wurden bei der M-set and C-set-Latent Dirichlet Allocation (MC-LDA) Mengen von mehr als zwei Wörtern, die bei den Must-Sets möglichst in denselben Themen und bezüglich der Cannot-Sets in verschiedenen Themen auftreten sollten, einbezogen [224]. Jedoch besteht nach Chen und Liu [239] und Zhai u. a. [240] ein Nachteil dieser Ansätze, darin, dass sie nicht mit einer hohen Anzahl von Cannot-Links beziehungsweise Cannot-Sets umgehen können, da diese je nach konkretem Ansatz entweder zu einer schlechten Performance [223] oder qualitativ mangelhaften Themen führen [225]. Zudem fokussieren sich diese Ansätze nicht darauf, bestimmte, seltene Themen zu finden, die der Nutzer in dem Datensatz vermutet.

Hierfür kommen Ansätze des **Targeted Topic Modellings** [z.B. 6, 226, 228] und des Seed-Guided Topic Modellings infrage. Es muss an dieser Stelle darauf hingewiesen werden, dass die Begriffe Targeted Topic Modelling und Seed-Guided Topic Modelling in der Literatur uneinheitlich verwendet werden. In dieser Arbeit werden unter dem Begriff Targeted Topic Modelling probabilistische Algorithmen wie Targeted Aspects Oriented Topic Modelling (TATM) [226], Targeted Analysis Model (TAM) [227] und BiTerms-based Topic Model (BiTTM) [228] verstanden, die darauf abzielen, möglichst feingranulare Themen zu extrahieren, die sich ausschließlich mit einem bestimmten Aspekt befassen. Zur Beschreibung dieses Aspekts, der aus verschiedenen Blickwinkeln beleuchtet werden soll, gibt der Nutzer einen einzelnen

charakteristischen Begriff vor. Bisher wurden diese Algorithmen vor allem angewendet, um Kunden bei der Auswertung von Produkt Reviews zu unterstützen und ihnen Themen zu einem bestimmten Aspekt eines Produkts aufzuzeigen [6–8, 229]. Im forensischen Kontext könnten sie beispielsweise dazu eingesetzt werden, verschiedene Unterthemen zu finden, die sich ausschließlich mit dem untersuchten Delikt befassen, wie beispielsweise der Drogenkriminalität.

Die grundlegende Idee dieser Algorithmen besteht darin, den Datensatz auf Dokumente [6, 226, 227, 229], Wortpaare [228] oder Wörter [7] zu reduzieren, die für den Aspekt relevant sind. Die Einschränkung auf relevante Dokumente wurde beispielsweise durch Reinforcement Learning [227], durch eine vorherige, unüberwachte Themenextraktion [226] mithilfe einer Adaption des DMM-Algorithmus [144] oder basierend auf dem Vorkommen des aspektspezifischen Begriffs beziehungsweise von Wörtern, die mit ihm in syntagmatischer Relation standen, realisiert [6, 229]. Insbesondere von Wang u. a. [228] wurde betont, dass durch das Ausschließen von ganzen Dokumenten bei der Themenmodellierung gegebenenfalls sinnvolle Themen verloren gehen, wenn zumindest ein Teil des Dokuments für den Aspekt relevant gewesen wäre. Um dieses Problem zu adressieren, wurden Ansätze vorgeschlagen, die die Relevanz von einzelnen Wörtern [7] bestimmen oder auf BTM [97] basieren und die Relevanz von Bitermen schätzen [228]. Zur Ermittlung der Relevanz wurde bei den Bitermen unter anderem berücksichtigt, ob diese ein aspektspezifisches Wort enthielten [228], während die Relevanzbestimmung von einzelnen Wörtern beispielsweise auf ihrem Vorkommen in einem externen Korpus basierte [7]. Die als irrelevant eingestuften Wörter beziehungsweise Wörter der irrelevanten Biterme und Dokumente wurden in einigen Arbeiten aus einer einzigen Wortverteilung eines irrelevanten Themas gezogen, das somit als Ausschuss-Thema dient und anschließend ignoriert werden kann [z.B. 6, 7, 228]. Alternativ wurden irrelevante Dokumente bereits vor der Themenmodellierung entfernt [226, 227]. Für diese Arbeit wären vor allem Ansätze wie TATM [226], TAM [227] und BiTTM [228] interessant, da diese auf Algorithmen wie DMM [144] und BTM [97] basieren, die besonders für kurze Texte geeignet sind. Dennoch ist zu beachten, dass der Erfolg dieser Ansätze im wesentlichen von der Qualität der Relevanzbestimmung abhängt und eine falsche Schätzung der Relevanz zum Verlust von wichtigen, fallrelevanten Informationen führen kann.

Hingegen berücksichtigen die Ansätze des **Seed-Guided Topic Modellings** den gesamten Datensatz [z.B. 14, 15, 235]. Wie Tabelle 2.6 entnommen werden kann, wurden für diese Verfahren erneut vor allem probabilistische Ansätze, genauer gesagt die LDA, adaptiert [z.B. 15, 27, 130]. Sie erfordern als Vorwissen, dass der Benutzer für jedes gewünschte Thema einige Begriffe, meist bezeichnet als „Seeds“, vorgibt, bei denen er vermutet, dass sie für das entsprechende Thema charakteristisch sind. Im Gegensatz zu dem Targeted Topic Modelling besteht das Ziel nicht darin, eine fokussierte Analyse zu einem einzigen Oberthema durchzuführen, sondern das Themenmodell dazu zu ermutigen, eine vorgegebene Anzahl an Themen zu detektieren, die mit den Seed Wörtern assoziiert werden [27].

Im Rahmen des probabilistischen Seed-Guided Topic Modellings wurde von Eshima u. a. [14] und Jagarlamudi u. a. [27] eine zusätzliche Wortverteilung eingeführt, die ausschließlich Seed Wörter umfasst. Weitere Möglichkeiten bestehen in der Einführung eines asymmetrischen Dirichlet-Priors bezüglich der Themen-Wort-Verteilung [15, 232] und in der Integration des GPU-Modells in den Algorithmus Collapsed Gibbs Sampling [130, 230, 231]. Es ist zu betonen,

dass bei allen Strategien das Modell nur dazu angeregt, aber nicht gezwungen wird, Themen zu extrahieren, die mit den Seed Wörtern in Verbindung stehen [14, 27, 230, 231]. Falls das gewünschte Thema überhaupt nicht in dem Datensatz auftritt, würde das Themenmodell dieses auch nicht hervorbringen [130]. Existiert das vermutete Thema jedoch zumindest zu einem geringen Anteil, bewirken die Ansätze des probabilistischen Seed-Guided Topic Modellings, dass die Seed Wörter des Themas sowie mit diesen in Verbindung stehende Wörter mit einer höheren Wahrscheinlichkeit in dem Thema vorkommen [12, 27, 230]. Darüber hinaus werden die Themenverteilungen von Dokumenten, die Seed Wörter enthalten, beeinflusst, sodass diese eine höhere Prävalenz für Themen aufweisen, die mit den in ihnen vorhandenen Seed Wörtern assoziiert werden [27, 130]. Die von Eshima u. a. [14], Watanabe und Baturo [15] und Churchill u. a. [130] vorgeschlagenen Erweiterungen der LDA unterscheiden sich insofern von den anderen Ansätzen, dass sie es ermöglichen weniger Seed Wortmengen als gewünschte Themen anzugeben. Somit sind diese in der Lage neben den vermuteten Themen, die durch Seed Wörter beschrieben wurden, zusätzlich unüberwachte Themen zu extrahieren.

Darüber hinaus zählen zu dem Seed-Guided Topic Modelling neben probabilistischen Algorithmen ebenfalls Verfahren des DLTM [234], Embedding-basierte Ansätze [235–237] und Verfahren basierend auf der Informationstheorie [90]. Die Algorithmen basierend auf Word Embeddings zeichnen sich von den anderen Ansätzen insofern ab, dass sie eine neue Definition eines Themas einführten [235–237]. Für jedes Thema wird hierbei vom Nutzer ein einzelnes Seed Wort spezifiziert [235–237]. Unter einem Thema wird eine Menge von Wörtern verstanden, die zu dem entsprechenden Seed Wort eine möglichst hohe semantische Ähnlichkeit aufweisen und zu anderen Seed Wörtern möglichst unähnlich sind. Im Gegensatz zu den anderen Ansätzen des Seed-Guided Topic Modellings wird gefordert, dass jedes Wort nur einem einzelnen Thema zugewiesen werden darf [235–237]. Die grundlegende Idee zur Messung der semantischen Ähnlichkeit bestand darin, Word Embeddings basierend auf dem Skip-Gram Ansatz von word2vec [63] zu trainieren, die jedoch nicht nur den lokalen Kontext des betrachteten Wortes, sondern auch das mit ihm assoziierte Seed Wort sowie das Dokument, in dem das Wort auftritt, vorhersagen [235–237].

Meng u. a. [235] und Zhang u. a. [236, 237] verglichen die Word Embedding basierten Ansätze mit probabilistischen halbüberwachten Verfahren basierend auf einer Nutzerstudie, wozu sie verschiedene Datensätze heranzogen, die bei Zhang u. a. [236] ebenfalls umgangssprachliche Tweets umfassten. Sie kamen hierbei zu dem Resultat, dass die Annotatoren bei Ansätzen basierend auf Word Embeddings mehr Begriffe unter den am höchsten gerankten Wörter eines Themas als tatsächlich relevant zu den Seed Wörtern betrachteten als bei anderen Ansätzen [235–237]. Hingegen kam Churchill u. a. [130] bei einer Evaluierung ebenfalls mithilfe einer Nutzerstudie zu dem Ergebnis, dass Category-Name Guided Text Embedding (CatE) [235] als Word Embedding basierter Ansatz insbesondere bei Datensätzen mit einem hohen statistischen Rauschen wie Tweets und dementsprechend auch mobilen Kommunikationsdaten dazu neigt irrelevante und hochfrequente Wörter in Themen hoch zu gewichten. Zudem ist generell ein Nachteil an den Word Embedding basierten Ansätzen darin zu sehen, dass diese nur eine Zuweisung von Wörtern zu Themen vornehmen und hingegen nicht die Abdeckung der Themen in den Dokumenten als Ausgabe liefern [235–237]. Diese Ansätze könnten daher

interessante Anwendungsmöglichkeiten der Themenmodellierung, wie beispielsweise die Reduktion des Datensatzes auf Nachrichten, die sich vor allem mit den durch die Seed Words beschriebenen Themen, beschäftigen, nicht realisiert werden.

2.8 Lösungsansätze für lückenhafte Kontexte

Eine weitere Herausforderung bestand in dem Problem der lückenhaften Kontexte durch die Kommunikation über verteilte Kanäle. Eine Möglichkeit, um diese zu adressieren, ist in der Anwendung von Verfahren des sogenannten LTM zu sehen [239, 241–254]. Das Ziel von LTM besteht darin, im Laufe der Zeit immer kohärentere und qualitativ hochwertigere Themen zu extrahieren, indem das Wissen aus früheren Themenmodellierungen einbezogen wird [243]. Meistens wird das semantische Wissen vollautomatisch aus den erhaltenen Themen aus vergangenen Datensätzen generiert und in einer stetig wachsenden Wissensbasis gesammelt, um dieses bei der Themenmodellierung auf einem neuen Datensatz berücksichtigen zu können [239, 246].

Bisher wurde LTM beispielsweise angewendet, um aktuelle Themen aus Streaming-Daten wie Tweets zu detektieren [z.B. 243] und Themen aus Produktbewertungen zu extrahieren [z.B. 239, 241, 248], wobei die Ergebnisse der Themenmodellierung auf Datensätzen zu verschiedenen Produkten berücksichtigt wurden, um auf einem Datensatz mit Bewertungen zu einem bestimmten Produkt möglichst gute Themen zu erhalten. Ebenfalls könnte es angewendet werden, um im forensischen Kontext das Problem anzugehen, dass fallrelevante Informationen segmentiert vorliegen. Ein potentieller Ansatz ist darin zu sehen, LTM für die Auswertung der Nachrichten eines konkreten Falls einzusetzen, indem die Wissensbasis nach jeder Durchführung von Topic Modelling auf den gespeicherten Nachrichten eines Mobilfunktelefons aktualisiert wird. Anschließend könnte dieses Wissen für die Themenextraktion aus den Nachrichten eines weiteren Mobilfunktelefons, das für den entsprechenden Fall ausgewertet werden muss, herangezogen werden.

Die bisherigen Ansätze von LTM unterschieden sich vor allem in der Art des automatisch erlernten Wissens, der Methode, wie das Wissen aus den gelernten Themen extrahiert wurde und dem angewendeten Algorithmus zur Themenmodellierung. Ein Überblick über die verschiedenen Ansätze wird in Tabelle 2.7 vermittelt.

Wie dieser entnommen werden kann, handelte es sich bei dem automatisch erlernten Wissen bei den meisten Ansätzen um die in Abschnitt 2.7 beschriebenen Must-Links [239, 241–246, 248–250, 252–254]. Einige Arbeiten [239, 241, 242, 245, 253] setzten zusätzlich Cannot-Links ein. Dementsprechend wurde beabsichtigt, die Themenkohärenz zu erhöhen, indem in das Themenmodell Wissen darüber integriert wurde, welche Wörter zusammengehören und welche möglichst nicht gemeinsam in einem Thema auftreten sollten [254].

Zur Extraktion von Must-Links aus den erlernten Themen wurden beispielsweise von Chen u. a. [243] und Lei u. a. [252] Wortpaare aus den Begriffen mit der höchsten Wahrscheinlichkeit in einem Thema gebildet. Alternativ wurden Wörter, die in denselben Themen eine hohe Wahrscheinlichkeit aufwiesen, als Must-Links betrachtet, wobei die Wahrscheinlichkeitsverteilungen der Wörter in den Themen mit dem Skalarprodukt verglichen wurden [248]. Mehrere

Tabelle 2.7: Ansätze von LTM bezüglich der Form und Extraktion des Wissens sowie des Algorithmus zur Themenmodellierung. Die Art des Wissens, die Methode, die zur Extraktion des Wissens aus den erstellten Themen angewendet wurde, sowie der Algorithmus zur Themenmodellierung sind aufgeführt.

	Ansatz
Art des Wissens	Must-Link [239, 241–246, 248–250, 252–254] Cannot-Link [239, 241, 242, 245, 253] Wortverteilungen von bisherigen Themen [251]
Extraktion des Wissens	Wortpaare der wahrscheinlichsten Wörter von Themen [243, 252] Skalarproduktähnlichkeit der Themenverteilungen [248] FIM [239, 244, 249, 254] Ähnlichkeit von Word Embeddings [241, 248] NPMI [242, 245, 250, 253] Community Detection [246]
Algorithmus zur Themenmodellierung	LDA [239, 244–246, 249, 250, 253, 254] NMF [243, 248, 252] DocNADE [251]

Ansätze des LTM integrierten zudem Word Embeddings in die Themenmodelle [241, 248, 251], die von Qin u. a. [248] zudem zur Extraktion von Must-Links und von Xu u. a. [241] zur Bildung von Must-Links und Cannot-Links verwendet wurden. Hierzu wurden die Word Embeddings entweder gemeinsam mit den Themenmodellen lokal auf den einzelnen Datensätzen trainiert [248] oder vortrainierte Word Embeddings wurden als Ausgangsbasis verwendet und basierend auf den Ergebnissen des Topic Modellings verbessert [241, 251]. Insbesondere für Daten mit einer hohen Sparsität, zu denen ebenfalls die Datensätze von forensischen Kurznachrichten zählen, bietet es sich nach Qin u. a. [248] an, nur thematisch ähnliche Wortpaare zu berücksichtigen, deren Word Embeddings ebenfalls eine hohe Ähnlichkeit aufweisen. Diese Vorgehensweise wurde damit begründet, dass die Themen aus Daten mit einer hohen Sparsität teilweise inkohärent sein können und dementsprechend nicht ausreichen, um aus diesen qualitativ hochwertiges Wissen extrahieren zu können und sich die Integration von Word Embeddings, wie in Abschnitt 2.4 erläutert wurde, bereits für einzelne Datensätze mit einer hohen Sparsität als erfolgversprechend erwies.

Die von Chen und Liu [239], Chen u. a. [243], Wang u. a. [249], Lei u. a. [252] und Chen und Liu [254] verwendeten Vorgehensweisen ähnelten sich insofern, dass die Must-Links aus den Themen eines einzigen Themenmodells gebildet wurden und die Wissensbasis nach der Themenmodellierung auf jedem einzelnen Datensatz aktualisiert wurde. Jedoch wurde von Chen und Liu [239], Wang u. a. [249] und Chen und Liu [254] die Ansicht vertreten, dass die Must-Links, die auf der Basis eines Themenmodells auf einem einzelnen Datensatz gewonnen wurden, nicht repräsentativ genug für neue Datensätze sind. Aus diesem Grund bildeten sie stattdessen die Must-Links aus Paaren von Wörtern, die gemeinsam in mehreren Themen auftraten, die aus einigen früheren Datensätzen extrahiert wurden. Diese wurden mit Algorithmen des FIM [255, 256] erkannt. Darüber hinaus wurden in der Arbeit von Chen und Liu [239] zur Extraktion der Cannot-Links ebenfalls die extrahierten Themen aus mehreren Datensätzen herangezogen, indem verglichen wurde, in wie vielen Themen die Wörter gemeinsam mit einer hohen Wahrscheinlichkeit auftraten und in wie vielen unterschiedlichen

Themen sie eine hohe Wahrscheinlichkeit aufwiesen. Jedoch zeigten Khan u. a. [253] in einer quantitativen Evaluierung basierend auf Reviews zu verschiedenen Elektronikartikeln, dass solche Ansätze des LTM, konkret die von Chen und Liu [239, 254] vorgeschlagenen Themenmodelle, nur eine hohe Themenkohärenz [19] liefern können, wenn das Wissen aus gelernten Themen von einer Vielzahl von Datensätzen stammt, während die Themenkohärenz bei einer geringen Anzahl von Datensätzen unter die erreichten Werte der LDA sank.

Im Gegensatz zu den bisher vorgestellten Ansätzen war Khan u. a. [242] der Auffassung, dass bei unüberwachten Verfahren der Themenmodellierung prinzipiell nicht festgestellt werden könnte, ob die aus Themen extrahierten Must-Links und Cannot-Links tatsächlich geeignet wären, um die Themenmodellierung auf einem anderen Datensatz zu verbessern. Ihr Vorgehen unterschied sich in [242] sowie in ihren weiteren Arbeiten [245, 246, 250] von der üblichen Vorgehensweise des LTM insofern, dass die Must-Links und Cannot-Links nicht direkt aus Themen, sondern aus dem Datensatz extrahiert wurden, auf dem das Themenmodell trainiert wurde. Hierzu wurden statistische Signifikanzmaße, genauer gesagt das NPMI [197] in [242, 245, 250, 253] sowie Algorithmen der Community Detection in [246] für die Extraktion der Must-Links und Cannot-Links aus den Datensätzen angewendet.

Wie in Tabelle 2.7 zu sehen ist, wurde das automatisch erlernte Wissen bisher vor allem in die LDA integriert [239, 244–246, 249, 250, 253, 254]. Um zu erreichen, dass die Wörter, die in der Wissensbasis einen Must-Link bildeten, möglichst über hohe Wahrscheinlichkeiten in denselben Themen verfügen und die Wörter eines Cannot-Links hingegen nicht in demselben Thema eine hohe Wahrscheinlichkeit aufwiesen, wurde wie bei den in Abschnitt 2.7 beschriebenen halbüberwachten Themenmodellen der Inferenz-Algorithmus Collapsed Gibbs Sampling der LDA angepasst. Wie jedoch von Chen u. a. [243] gezeigt wurde, erweist sich dieses Vorgehen als rechenintensiv. Hingegen konnte die Laufzeit auf demselben Datensatz deutlich reduziert werden, indem stattdessen das automatisch erlernte Wissen in das algebraische Themenmodell NMF integriert wurde. Hierzu wurde von Chen u. a. [243] sowie von Qin u. a. [248] und Lei u. a. [252] eine zusätzliche Einschränkung auf die Zielfunktion von NMF auferlegt und somit das Optimierungsproblem bei der Matrizenfaktorisierung adaptiert. Zudem wurde von Gupta u. a. [251] das Neural Topic Model (NTM) DocNADE [257] für LTM adaptiert. Als gelerntes Wissen dienten für dieses nicht Must-Links oder Cannot-Links, sondern die kompletten Wortverteilungen von mehreren gelernten Themen, wobei durch Regularisierung der Zielfunktion erreicht wurde, dass die Themen auf einem neuen Datensatz den bisher extrahierten Themen ähnelten.

Zusammenfassend ist zu sagen, dass LTM und insbesondere die Integration von automatisch erlernten Must-Links und Cannot-Links im forensischen Kontext eine Möglichkeit bietet, erkannte Zusammenhänge in die Themenmodellierung für die Auswertung eines konkreten Falls mit einzubeziehen. Wenn beispielsweise der Name einer Person zusammen mit anderen für den Fall relevanten Begriffen wie den Namen bekannter Tatverdächtiger oder verdächtiger Organisationen in einem Thema vorkommt, wäre es ratsam, dieses Wissen in die Themenextraktion aus den Nachrichten auf einem anderen Mobilfunktelefon zu berücksichtigen. Es ist jedoch zu betonen, dass der Einsatz von LTM für forensische Kommunikationsdaten auf die Nachrichten der Mobiltelefone, die im Rahmen eines konkreten Falls analysiert werden müssen, beschränkt ist. Das Übertragen von gelerntem Wissen aus der Themenmodellierung eines vorherigen Falls auf einen aktuellen Fall wäre hingegen als problematisch anzusehen,

da selbst die Daten von Fällen des gleichen Deliktbereichs hoch spezifisch sind. Darüber hinaus wären rechtliche Fragestellungen zu klären, ob Informationen einer Untersuchung auf eine andere übertragen werden dürften, da dadurch die Unvoreingenommenheit bei der Untersuchung eines Falls gefährdet sein könnte.

Jedoch kann auch bei LTM auf den Nachrichten eines Falls das Problem auftreten, dass nicht alle der Must-Links und Cannot-Links, die aus den Themen der Daten eines Mobilfunkgeräts extrahiert wurden, auch tatsächlich für die Themenmodellierung auf den Daten eines anderen Mobilgeräts relevant sind. Dies kann darauf zurückgeführt werden, dass, wie auch von Chen u. a. [244] und Chen und Liu [254] hervorgehoben wird, die Ambiguität von Begriffen eine Herausforderung für LTM darstellt. Dementsprechend muss eine semantische Beziehung in Form eines Must-Links, der auf einem bestimmten Datensatz zwischen zwei Wörtern besteht, nicht zwangsläufig auch zutreffend für einen anderen Datensatz sein, falls eines der Wörter auf diesem in einer anderen Bedeutung verwendet wird. Im forensischen Kontext wird dieses Problem durch die Verwendung von Hidden Semantics noch erhöht, da beispielsweise ein bestimmtes Codewort nur von einer Chatgruppe verwendet werden könnte, aber nicht in den anderen Chats.

Um dieses Problem zu adressieren, sollte, wie auch von Chen und Liu [239], Chen u. a. [244], Khan u. a. [245], Gupta u. a. [251] und Khan u. a. [253] vorgeschlagen wurde, überprüft werden, ob das bisher gesammelte Wissen tatsächlich für einen bestimmten Datensatz passend ist, bevor dieses bei der Themenmodellierung auf einem neuen Datensatz berücksichtigt wird. Beispielsweise wurde von Chen und Liu [239] und Wang u. a. [249] überprüft, ob zwei Wörter auf einem neuen Datensatz ebenfalls einen Must-Link bilden, indem ihre statistische Korrelation mit dem PMI berechnet wurde. Zusätzlich wurden von Chen und Liu [239] und Wang u. a. [249] auf jedem Datensatz vor dem Einsatz von LTM mit der Standard-LDA Themen extrahiert und anschließend im Rahmen des LTM nur die Must-Links berücksichtigt, die aus Themen stammten, die zu mindestens einem unüberwachten Thema auf dem neuen Datensatz eine hohe Kullback-Leibler-Divergenz aufwiesen. Ferner wurde von Khan u. a. [245] die Überschneidung zwischen den häufigsten Wörtern von jedem bisherigen Datensatz mit dem Vokabular des neuen Datensatzes untersucht, um die Eignung eines Datensatzes für die Durchführung von Topic Modelling auf dem neuen Datensatz zu bewerten. In ihrer späteren Arbeit rieten Khan u. a. [247] dazu, gerade bei einem längerfristigen Lernprozess, Must-Links und Cannot-Links, die bei mehreren aufeinanderfolgenden Themenmodellierungen auf verschiedenen Datensätzen als irrelevant erachtet wurden, endgültig aus der Wissensbasis zu entfernen. Allerdings erscheint die Relevanzbestimmung der Autoren als problematisch, da die hochfrequenten Wörter nicht zwangsläufig besonders aussagekräftig für einen Datensatz sein müssen [258] und zudem die semantische Ähnlichkeit von Texten sowie sprachliche Varianten einschließlich Rechtschreibfehlern nicht berücksichtigt werden.

2.9 Lösungsansätze für die hohe Variabilität im Vokabular

Ein Ansatz, um mit der Herausforderung der hohen Variabilität des Vokabulars in forensischen Kommunikationsdaten umgehen zu können, besteht darin, externe Wissensquellen bei der Themenmodellierung mit einzubeziehen [56, 259–261]. Das Ziel ist darin zu sehen,

möglichst viele Bedeutungen von Wörtern abzudecken und Informationen über die Semantik von Begriffen zu integrieren. Als externe Wissensquelle wurde beispielsweise von Van Linh u. a. [262] das semantische Netz WordNet [263] herangezogen. Von den Autoren wurde vorgeschlagen, ein graphenbasiertes DLTM um die Berücksichtigung von Wortrelationen in WordNet zu erweitern, wobei sie bezweckten, mit der Vielfalt des Vokabulars und der variierenden Bedeutung von Begriffen in Streaming-Daten wie Tweets umgehen zu können. Dies wurde realisiert, indem das verwendete Graph Convolutional Network (GCN) [264] die Themen basierend auf einer Graphenstruktur lernte, bei der die Wörter des Datensatzes verbunden wurden, wenn zwischen ihnen eine Synonym- oder Antonym-Relation bestand. Um mit Ambiguität umgehen zu können, wurden die Synonyme und Antonyme für alle verschiedenen Bedeutungen eines Wortes in der Graphenstruktur mit dem Wort verbunden.

Darüber hinaus wurden Ontologien wie beispielsweise DBpedia [265], Graphdatenbanken wie Freebase [266] und Taxonomien wie Probase [267] eingesetzt, um Informationen über Named Entities wie Firmen, Organisationen, Personen oder Orte in die Themenmodellierung integrieren zu können. Sowohl bei der von Wang u. a. [259] vorgestellten Entity Correlation Latent Dirichlet Allocation (EC-LDA) als auch bei der von Song u. a. [261] beschriebenen LDA with Entity based Similarity (ES-LDA) und der von Allahyari und Kochut [260] erläuterten Entity-based Topic Model (EntLDA) wurden Ähnlichkeiten von Named Entities beziehungsweise Beziehungen zwischen diesen in die LDA mit einbezogen. Hierzu wurde zunächst die Ähnlichkeit der Entities beispielsweise von Wang u. a. [259] und Song u. a. [261] basierend auf sogenannten Entity Embeddings, die mittels der Informationen aus Freebase erstellt wurden, oder von Allahyari und Kochut [260] mit dem speziell für DBPedia erstellten Wikipedia Link-based Measure (WLM) [268] bestimmt. Während Wang u. a. [259] und Allahyari und Kochut [260] die LDA mit dem Ziel adaptierten, dass ähnliche Entitäten über ähnliche Wahrscheinlichkeiten in denselben Themen verfügen sollten, gingen Song u. a. [261] davon aus, dass Dokumente, in denen ähnliche beziehungsweise verknüpfte Entitäten vorkamen, von ähnlichen Themen handelten. Daher sollte ihre Themenverteilung eine möglichst hohe Ähnlichkeit aufweisen.

Eine Einschränkung dieser Ansätze ist jedoch darin zu sehen, dass nach Tian u. a. [269] sowie nach Bollegala u. a. [270] externe Wissensquellen wie Thesauri und Taxonomien Begriffe vor allem in ihren üblichen Bedeutungen erfassen. Dementsprechend können sie zwar mit Ambiguität umgehen und die Vielfalt des Wortschatzes zu einem gewissen Teil abdecken, berücksichtigen jedoch nicht die besondere und manchmal versteckte Bedeutung von Wörtern im forensischen Kontext, beispielsweise in Form von Hidden Semantics [3]. Zudem ist insbesondere an dem von Van Linh u. a. [262] beschriebenen Ansatz als problematisch zu sehen, dass dieser sich zur Themenmodellierung ausschließlich auf die Wortrelationen nach externen Wissensquellen verlässt. Dementsprechend besteht die Gefahr, dass Informationen über spezifische, fallrelevante Verknüpfungen, beispielsweise zwischen Tatgegenständen und beteiligten Personen, verloren gehen. Bei der Integration von Ontologien in die Themenmodellierun kommt hinzu, dass sich bereits notwendige Vorbereitungsschritte wie die Extraktion der Named Entitites bei der mangelhaften Qualität der Kommunikationsdaten als problematisch erweisen können [271, 272].

2.10 Lösungsansätze für mehrsprachige Texte

Schließlich wäre als letzte Herausforderung auf die Mehrsprachigkeit von forensischen Kommunikationsdaten einzugehen. Um aus diesen möglichst kohärente Themen zu extrahieren, können Strategien der „Cross-Language Topic Extraction“ angewendet werden. Diese wurden speziell entwickelt, um Themen in multilingualen Datensätzen zu analysieren. Die einzelnen Ansätze unterscheiden sich in der Strategie, die angewendet wurde, um eine Verknüpfung zwischen den verschiedenen Sprachen in einem Datensatz herzustellen [273]. Hierfür existieren die folgenden Möglichkeiten:

- Verknüpfung über Dokumente des Datensatzes (engl. Document Linking) [75, 274–281]
- Verknüpfung über das Vokabular mithilfe von Wörterbüchern (engl. Vocabulary-linking) [138–140, 282–288]
- Kombination aus Document Linking und Vocabulary Linking [289, 290]
- Verknüpfung über Word Embeddings [9, 273, 291–295]

Es sollte darauf hingewiesen werden, dass die Sprache der einzelnen Dokumente für alle Ansätze bekannt sein muss [z.B. 139]. Um dies im forensischen Kontext zu erreichen, könnte entweder Ermittlerwissen darüber herangezogen werden, ob in bestimmten Chats ausschließlich in einer Sprache geschrieben wurde oder es könnte zunächst eine automatische Sprachdetektion durchgeführt werden, beispielsweise mit überwachten Word Embedding Modellen wie fastText [296]. Die im Folgenden vorgestellten Ansätze wurden entweder für die bilinguale Themenmodellierung entwickelt [z.B. 274, 278, 285] oder können für Datensätze in beliebig vielen Sprachen eingesetzt werden [z.B. 139, 275, 276].

Document Linking Die erstgenannte Kategorie von Verfahren, die unter dem englischen Begriff „Document Linking“ bekannt ist, setzt voraus, dass jedes Dokument des Datensatzes in allen Sprachen vorhanden ist [75, 274–281]. Hierbei kann es sich entweder um exakte Übersetzungen handeln oder um Dokumente, die zumindest dasselbe Konzept beschreiben, wie beispielsweise Versionen eines Wikipedia-Artikels in verschiedenen Sprachen. Die meisten dieser Verfahren adaptierten probabilistische Themenmodelle, insbesondere die LDA, mit dem Ziel, dass die verschiedensprachigen Versionen eines Dokuments entweder durch dieselbe Themenverteilung [274–277] oder durch eine ähnliche Themenverteilung [278, 297] beschrieben werden. Ein Thema wurde hierbei durch mehrere Wortverteilungen in den verschiedenen Sprachen repräsentiert [274–277]. Dieser Ansatz kommt für die Anwendung auf forensischen Kommunikationsdaten nicht infrage, da keine Verknüpfung zwischen den Nachrichten verschiedener Sprachen besteht. Aus demselben Grund können hybride Strategien, die Verfahren des „Document Linking“ und „Vocabulary Linking“ kombinieren [289, 290], nicht angewendet werden, da diese ebenfalls davon ausgehen, dass die mehrsprachigen Dokumente als verschiedene Versionen eines einzigen Dokuments betrachtet werden können.

Vocabulary Linking Interessanter sind hingegen die Verfahren des „Vocabulary Linkings“ [138–140, 282–288], unter denen Ansätze der „Cross-Language Topic Extraction“ verstanden werden, die Inhalte verschiedener Sprachen über Wörterbücher verknüpfen [278]. Im Gegensatz zu den zuvor beschriebenen Ansätzen stellen diese keine speziellen Anforderungen an den multilingualen Datensatz, auf dem das Themenmodell trainiert wird [284]. Daher kommen diese eher für forensische Kurznachrichten infrage. Beispielsweise entwickelten Zhang

u. a. [139] und Jiang u. a. [298] ein multilinguales, probabilistisches Themenmodell basierend auf der PLSA, das Themen als eine mehrsprachige Wahrscheinlichkeitsverteilung von Wörtern darstellte. Um zu erreichen, dass Übersetzungen im Wörterbuch ähnliche Wahrscheinlichkeiten in denselben Themen aufwiesen, wurde ein Regularisierer zur Likelihood-Zielfunktion hinzugefügt. Wie von Zhang u. a. [139] gezeigt wurde, kann die mehrsprachige Wortverteilung nach der Themenextraktion in mehrere einsprachige Wortverteilungen aufgeteilt werden, wodurch es möglich ist, zu untersuchen, wie in verschiedenen Sprachen dasselbe Thema diskutiert wurde. Darüber hinaus wurde neben der PLSA die LDA in mehreren Arbeiten für die Anwendung auf mehrsprachige Datensätze erweitert [138, 140, 282, 283]. Hierbei bestand ein häufiger Ansatz darin, ein Thema nicht mehr als Wahrscheinlichkeitsverteilung über Wörter, sondern über Übersetzungspaare darzustellen [138, 140, 282, 283].

Jedoch erweist sich nach Yang u. a. [287] an diesen Ansätzen als problematisch, dass sie davon ausgehen, dass die Texte in den unterschiedlichen Sprachen die gleichen Themen behandeln. Gerade bei Kommunikationsdaten können die verschiedensprachigen Chats mit unterschiedlichen Gesprächspartnern stattfinden. Daher kann nicht zwangsläufig angenommen werden, dass in diesen dieselben Themen besprochen werden. In diesem Fall könnte, wie bei dem von Yang u. a. [287] vorgeschlagenen Multilingual Topic Model (MTM), eigene Themen für jede Sprache im Rahmen des generativen Prozesses gelernt werden. Gleichzeitig werden gewichtete Verknüpfungen zwischen den Themen verschiedener Sprachen gelernt. Dabei bestehen Verbindungen mit hohen Gewichten nur zwischen Themen, bei denen es sich bei den wahrscheinlichsten Wörtern um viele sprachliche Entsprechungen handelt, während es ebenfalls möglich ist, Themen in einer Sprache zu extrahieren, die keinem Thema in einer anderen Sprache ähneln.

Eine weitere Herausforderung liegt nach Hao und Paul [284], Ma und Nasukawa [285] und Yuan u. a. [288] an den Ansätzen des Vocabulary Linkings darin, dass diese stark von der Qualität der verwendeten Wörterbücher abhängen. Dies wird an einer umfassenden Untersuchung von Hao und Paul [299] deutlich, der auf zehn bilingualen Korpora von Wikipedia-Artikeln die Einflussfaktoren auf die Qualität der Ergebnisse von Verfahren des Vocabulary Linkings [138, 140] untersuchte. Hierzu wurden zwei speziell für die Cross-Language Topic Extraction entwickelten Evaluierungsmethoden angewendet, konkret die Crosslingual Normalized Pointwise Mutual Information (CNPMI) [300] als Maß der Themenkohärenz und die Performance in der sprachübergreifenden Dokumentenklassifikation, wobei die Themenverteilung der Dokumente als Feature-Repräsentation diente [301]. Die Ergebnisse zeigten eine deutliche Verschlechterung der Themenkohärenz und der Performance in der sprachübergreifenden Dokumentenklassifikation, wenn viele der übersetzten Wortpaare im mehrsprachigen Datensatz nicht im Wörterbuch aufgeführt waren [299]. Dies kann damit erklärt werden, dass die multilingualen Erweiterungen der LDA basierend auf dem Vocabulary Linking bei einem bilingualen Wörterbuch mit einer geringen Abdeckung mit dem Vokabular des Datensatzes auf gewöhnliche LDA [285] oder DMM [138] reduziert werden. Für die Anwendung in der Forensik ist dies als problematisch anzusehen, da allgemeine Wörterbücher viele der umgangssprachlichen und ungewöhnlichen Wörter in den Kurznachrichten nicht enthalten [302]. Weitere Probleme ergeben sich nach Ma und Nasukawa [285] und Yuan u. a. [288] daraus, dass Übersetzungen in einem Wörterbuch fehlerhaft oder beispielsweise bei ambigen Wörtern nicht auf den Datensatz zutreffend sein können [285, 288].

Aus diesem Grund schlug Yuan u. a. [288] vor, das in Abschnitt 2.7 erwähnte interaktive Topic Modelling für die Themenextraktion aus multilingualen Datensätzen heranzuziehen. Hierzu wurde das sogenannte Anchor-based Topic Modelling, das speziell für das interaktive Topic Modelling von Arora u. a. [303] und Lund u. a. [304] entwickelt wurde, angepasst. Dem Problem, dass die geringe Qualität von Wörterbüchern zu inkohärenten Themen führen kann, wurde begegnet, indem die Themen, die durch mehrere Wortverteilungen in den verschiedenen Sprachen präsentiert wurden, in einem iterativen Prozess jeweils dem Nutzer gezeigt wurden. Im Anschluss wurden diese durch die Einbeziehung von Nutzerfeedback verbessert. Jedoch ist diese Vorgehensweise, wie bereits in Abschnitt 2.7 erwähnt wurde und zudem von Hao und Paul [299] hervorgehoben wurde, mit einem hohen Zeitaufwand für den Nutzer verbunden.

Darüber hinaus setzten Hao und Paul [284] und Ma und Nasukawa [285] Strategien des *Document Linkings* ein, um das Problem anzugehen, dass die Verfahren des *Vocabulary Linkings* unter einer mangelnden Qualität der Wörterbücher leiden. Von besonderem Interesse ist dabei der von Hao und Paul [284] vorgestellte Ansatz „Softlink“, der in zwei Untersuchungen [284, 299] die Themenkohärenz und Performance bezüglich der sprachübergreifenden Dokumentenklassifikation im Vergleich zu Verfahren des Vocabulary Linkings bereits übertraf, wenn nur 20 % der Einträge eines Wörterbuchs zur Verfügung standen. Die wesentliche Idee besteht darin, dass Dokumente über eine ähnliche Themenverteilung verfügen sollten, wenn sie einen hohen Anteil an überlappenden Wortübersetzungen aufweisen.

Ausgerichtete Word Embeddings Eine weitere Möglichkeit, um trotz eines Mangels von umfassenden Wörterbüchern kohärente, multilinguale Themen zu extrahieren, besteht in dem Einsatz von Word Embeddings [9, 273, 291–295]. Die meisten Ansätze bezogen sogenannte ausgerichtete Word Embeddings bei der Themenmodellierung mittels der LDA [9, 273, 291–294] oder DLTMs [295] mit ein. Ausgerichtete Word Embeddings ermöglichen es, Wörter aus verschiedenen Sprachen in einem gemeinsamen Embedding Raum zu repräsentieren [305]. Hierbei sollten verschiedensprachige Wörter mit einer ähnlichen Bedeutung über möglichst nahe beieinander liegende Word Embeddings verfügen [291]. Im Bereich der Cross-Language Topic Extraction wurden hierzu entweder zunächst monolinguale Word Embeddings auf einsprachigen Datensätzen trainiert und mit verschiedenen Transformationsmethoden [z.B. 306, 307] in den gemeinsamen Embedding Raum abgebildet [9, 273, 291, 292] oder vortrainierte, kontextabhängige Word Embeddings wie Multilingual BERT (M-BERT) [66] eingesetzt, die auf mehrsprachigen Datensätzen gelernt wurden [291, 294, 295]. Es ist darauf hinzuweisen, dass der erste Ansatz für die Transformation ebenfalls eine geringe Anzahl an Einträgen eines bilingualen Wörterbuchs voraussetzt [9, 273, 291, 292].

Die einzelnen Ansätze der Themenmodellierung in diesem Bereich unterscheiden sich vor allem in dem Aspekt, ob zuerst das Vokabular verschiedener Sprachen über die ausgerichteten Word Embeddings verknüpft wurde und anschließend die Themenmodellierung durchgeführt wurde [273, 291, 292, 295] oder die Word Embeddings erst nach der Themenextraktion zum Einsatz kamen [9, 292, 294]. Hinsichtlich der erstgenannten Vorgehensweise wurden beispielsweise die Dokumente auf den verschiedenen Sprachen mithilfe der ausgerichteten Word Embeddings in eine einheitliche Zielsprache übersetzt und anschließend auf den übersetzten Dokumenten die gewöhnliche LDA durchgeführt [292]. Eine weitere Möglichkeit

bestand darin, die ausgerichteten Word Embeddings direkt in den generative Prozess der LDA zu integrieren, indem die Themen-Wort-Verteilung durch die Softmax-Funktion approximiert wurde [273].

Hinsichtlich der zweitgenannten Variante wurden auf allen Dokumenten einer bestimmten Sprache jeweils separate Themenmodelle mit der gewöhnlichen LDA gelernt [292, 294]. Anschließend wurden ähnliche Themen in verschiedenen Sprachen ermittelt [292, 294]. Hierfür wurde die sprachübergreifende Vektorrepräsentation, die mittels der ausgerichteten Word Embeddings generiert wurde, verwendet, um die wahrscheinlichsten Wörter der Themen miteinander zu vergleichen.

Die *Cross-Language Topic Extraction* unter Einbeziehung von ausgerichteten Word Embeddings könnte für die Themenmodellierung in forensischen Kommunikationsdaten erfolgversprechend sein, da diese, wie von Chang und Hwang [273] gezeigt wurde, bei Datensätzen von kurzen Texten deutlich kohärentere Themen - gemessen am CNPMI [300] - als Verfahren des Vocabulary Linkings hervorbringen konnten. Darüber hinaus besteht ein weiterer Vorteil nach Tsou u. a. [9] darin, dass diese, je nach Art der ausgerichteten Word Embeddings, höchstens eine geringe Anzahl an Einträgen eines bilingualen Wörterbuchs als externe Ressource benötigen. Im Gegensatz zu den Verfahren des Vocabulary Linkings können hierfür jedoch allgemeine Wörterbücher herangezogen werden, die nur eine geringe Abdeckung mit dem Vokabular des Datensatzes aufweisen [9, 273]. Dementsprechend zeigten Chang und Hwang [273], dass Ansätze basierend auf ausgerichteten Word Embeddings deutlich bessere Resultate bezüglich der Themenkohärenz sowie der Performance in der sprachübergreifenden Dokumentenklassifikation als Verfahren des Vocabulary Linking erzielten, wenn das Wörterbuch weniger als 10 % der Übersetzungspaare des Datensatzes umfasste.

2.11 Evaluierung

Um die Qualität von verschiedenen Ansätzen der Themenmodellierung bewerten zu können, ist es wichtig zu wissen, wie diese evaluiert werden können. Im Allgemeinen kann, wie beispielsweise von Churchill und Singh [24] erläutert wurde, zwischen qualitativer und quantitativer Evaluierung unterschieden werden. Es muss jedoch an dieser Stelle darauf hingewiesen werden, dass keine einheitliche Kategorisierung von Evaluierungsmethoden existiert. In den meisten Arbeiten wurde unter einer qualitativen Evaluierung verstanden, dass Beispiele von Themen präsentiert und interpretiert werden [z.B. 51, 231, 233, 259]. Hierbei wurden üblicherweise die wahrscheinlichsten Wörter von ausgewählten Themen gezeigt und Terme, die als irrelevant für das Thema erschienen, gekennzeichnet [z.B. 51, 231, 233].

Die quantitative Evaluierung kann nach Hoyle u. a. [18] automatisch oder unter Einbeziehung von menschlichen Annotatoren erfolgen. Eine Übersicht über die verschiedenen Kategorien von automatischen Evaluierungsmethoden und konkrete Maße dieser Kategorien kann Tabelle 2.8 entnommen werden. Ebenfalls ist in dieser aufgeführt, ob der Fokus bei der Evaluierung auf der TWV oder der DTV des Themenmodells liegt und ob zur Evaluierung eine Groundtruth (GT) bezüglich der korrekten Themen der Dokumente benötigt wird. Es ist darauf hinzuweisen, dass bei Themenmodellen wie BTM [97] oder Ansätzen basierend auf DMM

[z.B. 170], die nicht direkt die DTV, aber die TWV ausgeben, die DTV beispielsweise mittels der von beschriebenen Quan u. a. [157] Methode Summation over Words (SW) abgeleitet werden kann.

Tabelle 2.8: Überblick über automatische Evaluierungsmethoden für die Themenmodellierung. Aufgeführt sind die Kategorie von Evaluierungsmethoden beziehungsweise und bekannte Methoden der Kategorie. Ebenfalls ist angegeben, ob der Fokus der Evaluierungsmaße auf der Bewertung der Themen-Wort-Verteilung oder Dokument-Themen-Verteilung liegt und ob ein mit Themen annotierter Datensatz benötigt wird.

Kategorie	Maß/ Methode (Auswahl)	TWV	DTV	GT
Prädiktive Maße	Log Likelihood			
	Perplexity [41]			
Kohärenz	C_{UCI} [150]	x		
	C_{UMass} [19]	x		
	C_V [308]	x		
	C_{NPMI} [309]	x		
	TF-IDF-Kohärenz [12]	x		
	Word Embedding based Topic Coherence (WETC) [55]	x		
Themendiversität	Anteil einzigartige Wörter [62]	x		
	FREX [310, 311]	x		
	inverse Average Jaccard Similarity (inverse-AJS) [273]	x		
	Dice Index	x		
	Kullback-Leibler-Divergenz (KL-Divergenz) [312]	x		
Zusammenfassung mittels Themen	Topic Information Gain [65]	x	x	
	Specifity und Specificity entropy [313]		x	
Downstream Application	Dokument Klassifikation		x	x
	Dokument Clustering		x	x
	Information Retrieval		x	x

Prädiktive Maße Hinsichtlich der automatischen Evaluierung bestand eine häufig angewendete Methode darin, zu bewerten, wie gut ein gelerntes Themenmodell bisher ungesehene Dokumente eines Held-Out Datensatzes vorhersagen kann und somit die Generalisierungsfähigkeit des Modells zu beurteilen [z.B. 10, 46]. Hierzu wurde entweder die Log-Likelihood wie beispielsweise in [10, 314] oder die Perplexity [41] wie zum Beispiel in [46, 77, 216, 314] auf den unbekannten Dokumenten als Evaluierungskriterium herangezogen. Bei der Perplexity handelt es sich um den Kehrwert der durchschnittlichen Log-Likelihood pro Wort [4]. Sie misst, wie überrascht beziehungsweise verwirrt ein Modell über ungesehene Daten ist [48]. Dementsprechend bedeutet eine geringe Perplexity, dass das Themenmodell besser in der Lage ist, die Wahrscheinlichkeitsverteilung der Wörter in neuen Dokumenten zu modellieren. Jedoch ist darauf hinzuweisen, dass Chang u. a. [315] zeigten, dass die Perplexity nicht mit der menschlichen Bewertung der Interpretierbarkeit von Themen korreliert ist und dieser teilweise sogar widerspricht. Darüber hinaus kommt die Perplexity nur zur Evaluie-

rung von Algorithmen infrage, die den generativen Prozess von Dokumenten modellieren [316] und ist daher beispielsweise nicht für algebraische Themenmodelle wie NMF sowie für Varianten von probabilistischen Themenmodellen wie WNTM [131] geeignet.

Kohärenz Alternativ wurden Maße für die Themenkohärenz als automatische Evaluierungsmetrik verwendet. Diese beurteilen im Allgemeinen, wie sehr die wahrscheinlichsten Wörter eines Themas zusammenhängen [308]. Um die Qualität des gesamten Themenmodells beurteilen zu können, kann beispielsweise der Durchschnitt der einzelnen Kohärenzwerte der Themen berechnet werden [z.B. 161, 172]. Für eine detaillierte Beschreibung der einzelnen Maße für die Themenkohärenz wird auf Röder u. a. [308] verwiesen. Bei den meisten Kohärenzmaßen wird ein Bestätigungsmaß für Wortpaare unter den Begriffen mit der höchsten Wahrscheinlichkeit eines Themas berechnet, das angibt, wie sehr sich zwei Begriffe ähneln beziehungsweise wie sehr diese miteinander assoziiert werden [308]. Die einzelnen Werte werden beispielsweise mit dem arithmetischen Mittel zu einem einzigen Kohärenzwert für ein Thema aggregiert [308]. Während bei intrinsischen Kohärenzmaßen wie der semantischen Kohärenz, auch bekannt als C_{UMass} [19], das Bestätigungsmaß auf demselben Datensatz berechnet wird, auf dem auch das Themenmodell gelernt wird, wird es bei extrinsischen Kohärenzmaßen wie C_{UCI} [150] auf einem externen Referenzkorpus ermittelt [317].

Als bekanntes Maß für die Themenkohärenz wäre beispielsweise das von Newman u. a. [150] vorgeschlagene Maß C_{UCI} [150] anzuführen. Dieses verwendet als Bestätigungsmaß das PMI [191], das für jedes Wortpaar auf einem externen Korpus wie Wikipedia geschätzt wird. Aletras und Stevenson [309] schlug vor, das PMI durch das NPMI zu ersetzen. Wie von Lau u. a. [318] gezeigt wurde, ist dieses Kohärenzmaß, das in Tabelle 2.8 als C_{NPMI} bezeichnet wird, im Gegensatz zu der Perplexity mit der menschlichen Bewertung von Themen korreliert. Dementsprechend wurde sie häufig zur Evaluierung von neuen Algorithmen zur Themenmodellierung angewendet, wie beispielsweise von Churchill und Singh [124], Nguyen u. a. [141] und Zhang u. a. [236]. Jedoch zeigten Doogan und Buntine [17], dass im Allgemeinen die Korrelation zwischen Kohärenzmaßen und der menschlichen Interpretierbarkeit stark abhängig davon ist, auf welchem Korpus sie berechnet werden. Zhao u. a. [48] riet deshalb dazu, vor allem für spezielle Datensätze wie Twitter-Daten und dementsprechend auch für Kommunikationsdaten intrinsische Kohärenzmaße zu wählen.

Themendiversität Weitere Evaluierungsmaße bewerten, ob ein Algorithmus in der Lage ist möglichst einzigartige Themen zu finden [62, 273, 312]. Eine Möglichkeit, die beispielsweise von Dieng u. a. [62], Airoldi und Bischof [310] und Bischof und Airoldi [311] angewendet wurde, besteht darin, zu beurteilen, inwiefern sich die wahrscheinlichsten Wörter der Themen überschneiden. Dieng u. a. [62] berechnete beispielsweise den Prozentanteil einzigartiger Wörter unter den wahrscheinlichsten Begriffen aller Themen des Modells. Das von Bischof und Airoldi [311] beschriebene Maß FREX, das detaillierter in Abschnitt 3.3.2 erläutert wird, bezieht zusätzlich den Wahrscheinlichkeitswert der Top Wörter in den Themen mit ein. Eine Alternative zu dem Vergleich der wahrscheinlichsten Wörter besteht darin, die Ähnlichkeit der vollständigen Themen-Wort-Verteilungen zu messen [z.B. 206]. Hierzu wurde zum Beispiel von Wang und McCallum [206] die durchschnittliche Distanz der Wortverteilungen der Themen mit der KL-Divergenz [312] berechnet.

Zusammenfassung mittels Themen Der von Angelov [65] beschriebene „Topic Information Gain“ und die von Yuan u. a. [313] aufgestellte „Specifity“ und „Specifity Entropy“ als weitere Evaluierungsmaße für Themenmodelle bewerten, ob die Dokumente des Datensatzes durch die Themen angemessen beschrieben beziehungsweise zusammengefasst werden, wozu Maße der Informationstheorie eingesetzt wurden. Genauer gesagt misst der „Topic Information Gain“ basierend auf der MI [92], wie informativ die wahrscheinlichsten Wörter eines Themas für ein Dokument sind, dass eine hohe Wahrscheinlichkeit in diesem Thema besitzt. Dieses Maß erfasst sowohl die Qualität der TWV als auch der Zuordnung der Themen zu Dokumenten beziehungsweise der DTV. Dementsprechend nimmt es sowohl bei Themen mit wenigen informativen Wörtern als auch bei einer falschen Zuordnung von Dokumenten zu Themen niedrige Werte an.

Die grundlegende Idee der von Yuan u. a. [313] entwickelten „Specifity“ besteht darin, dass ein Dokument durch eine geringe Anzahl von Themen treffend beschrieben werden kann. Mit anderen Worten sollte bei einem gelungenen Themenmodell das Dokument für wenige Themen eine hohe Abdeckung und für die anderen Themen eine geringe Wahrscheinlichkeit aufweisen. Die Berechnung der „Specifity“ basiert auf dem Vergleich der Themenverteilung eines einzelnen Dokuments mit einer bimodalen Verteilung. Die „Specifity Entropy“ bezieht sich hingegen nicht auf einzelne Dokumente, sondern auf eine Teilmenge von Dokumenten des Datensatzes, die beispielsweise beim Information Retrieval für eine ausgewählte Query zurückgeliefert wurden. Sie entspricht der Entropie [192] der „Specifity“ der einzelnen Dokumente dieser Teilmenge. Diese sollte bei einem gelungen Themenmodell möglichst gering sein, da dies darauf hinweist, dass eine geringe Anzahl an Themen in den Dokumenten der Ergebnismenge mit einer hohen Abdeckung vertreten ist. Themen werden somit als aussagekräftig angesehen, wenn sie nicht über alle Dokumente des Datensatzes mit einer ähnlichen Wahrscheinlichkeit vertreten sind, sondern in einer Teilmenge besonders prominent sind.

Downstream Application Darüber hinaus wurden Themenmodelle anhand der sogenannten „Downstream Application Performance“ bewertet, worunter verstanden wird, dass sie auf eine externe Aufgabe angewendet werden und die Qualität der Themenmodelle anhand ihrer Performance in dieser Aufgabe bewertet wird [18, 48]. Im Gegensatz zu der Themenkohärenz und der Themendiversität fokussiert sich diese Evaluierungsmethode nicht auf die Bewertung der TWV, sondern dient dazu, die DTV zu evaluieren [48]. Diese wird hierbei als die semantische Repräsentation von Dokumenten aufgefasst [48]. Die in Tabelle 2.8 aufgeführten Methoden der Downstream Application haben gemeinsam, dass sie erfordern, dass die Datensätze mit dem tatsächlichen Thema der Dokumente annotiert sind [157, 169, 252, z.B.].

Eine häufig gewählte externe Aufgabe besteht in der Dokumentenklassifikation [z.B. 119, 157, 165]. Auf den gelabelten Daten wird ein Klassifikationsmodell trainiert, wobei die Themenverteilungen als Feature-Vektoren der Dokumente betrachtet werden. Die Qualität der DTV werden anhand von klassischen Evaluierungsmaßen der Klassifikation beurteilt.

Zudem wurden Themenmodelle anhand der Aufgabe des Dokumenten Clusterings bewertet und verglichen. Hierfür existieren zwei grundlegende Möglichkeiten: Die erste Möglichkeit, die beispielsweise von Yan u. a. [97], Duan u. a. [216] und Vosecky u. a. [319] angewendet wurde, ist darin zu sehen, die Themenverteilungen wie bei der Klassifikation als Dokumen-

tenrepräsentation anzusehen und anschließend ein Clustering mit Algorithmen wie k-means [320] und Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [321] durchzuführen. Eine häufigere Vorgehensweise sah vor, jedes Thema als ein Dokumentencluster aufzufassen und ein Dokument dem Thema beziehungsweise Cluster zuzuordnen, für das es die höchste Wahrscheinlichkeit aufwies [z.B. 126, 141, 149, 169, 216, 217, 234]. Da bei beiden Ansätzen davon ausgegangen wird, dass das tatsächliche Thema des Dokuments bekannt ist, können Evaluierungsmaße für das Clustering, die eine Groundtruth benötigen, verwendet werden. Beispielsweise wurde das NMI in [126, 141, 149, 169, 216, 217], die Purity in [126, 141, 169] und der Adjusted Rand Index (ARI) in [169, 234] herangezogen.

Schließlich erfolgte die Evaluierung von Themenmodellen ebenfalls anhand der Aufgabe des Text Retrievals, die beispielsweise von Gupta u. a. [251], Lei u. a. [252] und Larochelle und Lauly [257] gewählt wurde. Hierbei werden Dokumente des Testdatensatzes, auf dem nicht das Themenmodell trainiert wurde, als Queries verwendet, um Dokumente aus dem Trainingsdatensatz zu erhalten. Als Rangfunktion dient die Ähnlichkeit der Themenverteilungen der Dokumente. Ein zurückgeliefertes Trainingsdokument wird als relevant angesehen, wenn es mit demselben Thema gelabelt wie die Query beziehungsweise das Testdokument ist. Auf diese Weise können Evaluierungsmaße des Text Retrievals wie die Precision in [257] und die Precision@k in [251, 252] berechnet werden.

Nutzerstudie Eine Alternative zur automatischen Evaluierung bieten Nutzerstudien [z.B. 225, 315], die zudem ebenfalls eingesetzt wurden, um zu bewerten, ob quantitative Evaluierungsmaße mit der menschlichen Bewertung eines Themas übereinstimmen. Beispielsweise wurde von Chang u. a. [315] die Word Intrusion sowie die Topic Intrusion vorgeschlagen und beispielsweise in Nikolenko u. a. [12], Eshima u. a. [14] und Li u. a. [127] zum Vergleich des entwickelten Themenmodells mit Baseline Methoden angewendet. Während die Word Intrusion darauf abzielt die TWV beziehungsweise die Themenkohärenz zu bewerten, kann mithilfe der Topic Intrusion die DTV beurteilt werden. Die Aufgabe der Probanden bei der Word Intrusion ist es unter einer Menge von präsentierten Begriffen einen Eindringling zu identifizieren, der nicht zu den anderen Wörtern passt. Während es sich bei diesem um ein Wort mit einer geringen Wahrscheinlichkeit in einem bestimmten Thema handelt, weisen die anderen Wörter der Menge eine hohe Wahrscheinlichkeit in diesem Thema auf. Je leichter es den Probanden fällt, das unpassende Wort auszuwählen, als desto kohärenter ist das Thema anzusehen. Bei der Topic Intrusion sollen die Probanden ein Thema bestimmen, das nicht zu einem Dokument passt. Hierzu werden ihnen ein Ausschnitt eines Dokuments und die wahrscheinlichsten Wörter von mehreren erstellten Themen des Datensatzes vorgelegt. Während das Eindringling-Thema mit einer geringen Wahrscheinlichkeit in dem Dokument aufweisen, handelt es sich bei den anderen Themen um diejenigen mit der höchsten Abdeckung in dem Dokument. Analog zum Word Intrusion Test sollte das Erkennen des falschen Themas bei einer gelungenen DTV mit wenig Schwierigkeiten verbunden sein. Aus diesen Nutzerstudien kann ein quantitatives Maß abgeleitet werden, indem beispielsweise der Anteil der Probanden, denen es gelang, den richtigen Eindringling auszuwählen, berechnet wird [12, 315].

Das Topic Labeling und das Word Labeling [z.B. 190, 225, 286] bilden eine weitere Möglichkeit zur Evaluierung von Themenmodellen unter Einbeziehung von menschlichen Annotatoren. Hierbei werden zunächst die Themen, die durch die wahrscheinlichsten Wörter repräsentiert

werden, durch die Annotatoren als kohärent oder inkohärent gelabelt. Anschließend werden die einzelnen Wörter eines kohärenten Themas als korrekt oder inkorrekt gekennzeichnet. Als Evaluierungsmetrik wurden Maße aus dem Information Retrieval wie die $Precision@n$ eingesetzt [190, 225, 286].

Schließlich wurden beispielsweise von Doogan und Buntine [17] und Churchill und Singh [124] Evaluierungsmethoden beschrieben, welche davon ausgehen, dass menschliche Annotatoren für Themen mit hoher Interpretierbarkeit eher einen Namen oder ein Label vergeben können als für jene, die für sie wenig Sinn ergeben. Doogan und Buntine [17] und Churchill und Singh [124] setzten diese Methode ein, um die TWV von Themenmodellen zu beurteilen. Sie zeigten hierfür den Annotatoren die wahrscheinlichsten Wörter eines zu bewertenden Themas, für das sie in [17] frei ein Label wählen konnten und in [124] sich für eines von mehreren zur Verfügung stehenden Labels entscheiden mussten. Von Doogan und Buntine [17] wurde eine weitere Aufgabe zur Bewertung der Dokument-Themen-Verteilung aufgestellt. Im Rahmen dieser mussten die Probanden ebenfalls ein Label aus einer Menge von Kandidaten für ein Thema auswählen, jedoch wurde ihnen nicht die wahrscheinlichsten Wörter gezeigt, sondern die Dokumente, die die höchste Wahrscheinlichkeit für das zu labelnde Thema aufwiesen. Quantitative Maße, die hieraus abgeleitet wurden, umfassten die Anzahl der Probanden, die in der Lage waren, ein Label zu vergeben [17] und die Anzahl der Probanden, bei denen das gewählte Label übereinstimmte [17, 124].

Tabelle 2.9: Möglichkeiten zur Evaluierung unter Einbeziehung menschlicher Annotatoren. Angegeben sind die verschiedenen Kategorien von Methoden, die Themenmodelle mithilfe einer Nutzerstudie evaluieren sowie eine Auswahl der Methoden und Maße der entsprechenden Kategorie. Ebenfalls ist aufgeführt, ob mit den einzelnen Methoden die TWV oder DTV bewertet wurde.

Kategorie	Methode/ Maß (Auswahl)	TWV	DTV
Intrusion Tasks	Word Intrusion [315]	x	
	Topic Intrusion [315]		x
	R4WSI Kohärenz [116]	x	
Direkte Kohärenzbewertung	Topic Labeling [190, 225, 286]	x	
	Word Labeling [190, 225, 286]	x	
Vergabe von Topic Labels	Basierend auf wahrscheinlichsten Wörtern [17]	x	
	Basierend auf repräsentativsten Dokumenten [17]		x

2.12 Zusammenfassung

Zusammenfassend ist zu sagen, dass die Themenmodellierung bisher kaum zur Unterstützung der Datenauswertung im forensischen Kontext eingesetzt wurde [93–96]. Darüber hinaus lag in den bisherigen Arbeiten der Fokus nicht auf den forensischen Kommunikationsdaten. Dies kann darauf zurückgeführt werden, dass diese mit besonderen Herausforderungen verbunden sind. Zu diesen zählen die geringe Länge der Nachrichten, ihre mangelhafte sprachliche Qualität, einschließlich eines hohen Anteils an irrelevanten „noise words“, und

die Tatsache, dass der Kontext der Nachrichten wie der Zeitpunkt, zu dem sie gesendet wurden oder der Sender, die Themen beeinflussen kann. Zudem zeichnen sich die forensischen Kommunikationsdaten durch die verschiedenen Kommunikationsteilnehmer und deren soziodemographische Merkmale über ein vielfältiges Vokabular aus [133, 135]. Hinzu kommt, dass fallrelevante Nachrichten auf verschiedenen Mobilfunktelefonen gespeichert sein können, was mit segmentierter Information und lückenhaften Kontexten einhergeht und zudem dazu führt, dass die Nachrichten teilweise auf mehreren Sprachen verfasst sind. Außerdem besitzt der Ermittler meist Vorwissen über den Fall und erwartet bestimmte Themen im Datensatz, die jedoch durch traditionelle Algorithmen teilweise nicht identifiziert werden können [130].

Die Algorithmen der verschiedenen Oberkategorien von Themenmodellen können die einzelnen Herausforderungen unterschiedlich gut bewältigen. Beispielsweise gelten für kurze und verrauschte Texte gerade Ansätze des Fuzzy Clusterings [10, 86], graphenbasierte Ansätze [71–80] sowie das Exemplar-based Topic Modelling [89] als geeignet. Jedoch können sie zum Beispiel keine Kontextinformationen integrieren oder das Interesse des Ermittlers an bestimmten Themen berücksichtigen. Darüber hinaus wurden vor allem für probabilistische Ansätze wie die LDA, aber auch algebraische Modelle und DLTMs Erweiterungen vorgeschlagen, um eine oder wenige der Schwierigkeiten zu überwinden. Bisher fehlen jedoch Ansätze zur Bewältigung aller Herausforderungen im Zusammenhang mit forensischen Kommunikationsdaten.

Daher ist eine Priorisierung je nach konkretem Anwendungsfall vorzunehmen. So sind beispielsweise gerade bei grenzüberschreitender und organisierter Kriminalität die lückenhaften Kontexte durch die verteilte Kommunikation und die Mehrsprachigkeit als problematisch anzusehen. Das erste Problem kann durch LTM adressiert werden, das die Erkenntnisse aus der Themenmodellierung auf den Kommunikationsdaten eines Mobilfunktelefons in die Themenextraktion auf neuen Kommunikationsdaten einbezieht. Bezüglich der Mehrsprachigkeit von Datensätzen sind insbesondere Ansätze basierend auf sprachübergreifenden, ausgerichteten Word Embeddings als erfolgversprechend anzusehen, da diese weder besondere Anforderungen an den Datensatz stellen noch umfangreiche Wörterbücher voraussetzen. Des Weiteren kann die Berücksichtigung von Kontexten vor allem dann wichtig sein, wenn es um die Beantwortung spezifischer Fragen geht, beispielsweise wann ein fallrelevantes Thema besonders intensiv diskutiert wurde oder welche Nutzer sich vor allem mit diesem befassten.

In dieser Arbeit wurde hingegen der Fokus darauf gelegt, den Ermittler bei zwei grundsätzlichen Zielen bezüglich der Auswertung der umfangreichen Menge von Kurznachrichten zu unterstützen: der Exploration und der Überprüfung bereits aufgestellter Hypothesen. Für diese beiden Ziele können Ansätze des probabilistischen Seed Topic Modellings wie keyATM [14], die Seeded LDA [15] und Guided Topic-Noise Model (GTM) [130] vielversprechend sein. Im Gegensatz zu anderen halbüberwachten Ansätzen, wie beispielsweise des Targeted Topic Modellings [z.B. 6, 227, 228], zwingen sie das Themenmodell nicht dazu, Themen zu finden, die der Erwartungshaltung des Ermittlers entsprechen, sondern ermutigen es nur dazu. Hierdurch eignen sie sich dazu, festzustellen, ob ein vermutetes Thema tatsächlich im Datensatz auftritt. Gleichzeitig ermöglichen sie es weitere unüberwachte Themen zu finden, wodurch neue Erkenntnisse und Zufallsfunde bei Ermittlungen unterstützt werden.

Mit wenigen Ausnahmen [124] wurden die Algorithmen des Seed-Guided Topic Modellings jedoch für lange und sprachlich korrekte Dokumente entwickelt. Aus diesem Grund ist es erforderlich, diese für die Anwendung auf den Kommunikationsdaten mit Strategien zu kombinieren, die vor allem mit der Herausforderung der hohen Sparsität umgehen können. Jedoch erweist sich als erschwerend für die Auswahl einer geeigneten Strategie, dass die einzelnen Arbeiten je nach Datensatz und Evaluierungsmethode verschiedene Ansätze als vielversprechend betrachteten.

Die Evaluierung stellt sich als prinzipielles Problem bei der Themenmodellierung dar. Um verschiedene Algorithmen miteinander vergleichen zu können, sind vor allem quantitative Evaluierungsmaße hilfreich. Hierbei wurden eine Reihe von automatischen Evaluierungsmethoden wie die Berechnung der Themenkohärenz [z.B. 19, 150, 309], der Themendiversität [z.B. 62, 273] und prädiktive Maße wie die Log Likelihood vorgeschlagen [z.B. 41], die sich im Gegensatz zu Methoden der Downstream Application dadurch auszeichnen, dass sie keine annotierten Datensätze benötigen. Insbesondere die Themenkohärenz wurde vielfach zur Evaluierung herangezogen [73, 172, z.B.]. Allerdings widersprechen sich die einzelnen Arbeiten darin, ob diese tatsächlich mit der menschlichen Einschätzung der Interpretierbarkeit von Themen übereinstimmt [17, 18, 318]. Dementsprechend ist es gerade bei Daten wie forensischen Kurznachrichten, die bisher noch nicht für Themenmodellierung herangezogen wurden, erforderlich, die quantitativen Maße mit der menschlichen Einschätzung zu vergleichen.

3 Daten und Methoden

Um eine geeignete Möglichkeit zu finden, den Ermittler bei dem Finden von Beweisen zu vermuteten Themen zu unterstützen, wurden zwei Verfahren des Seed-Guided Topic Modellings miteinander verglichen: das von Eshima u. a. [14] vorgestellte keyATM und die von Watanabe und Baturo [15] beschriebene Seeded LDA. Beide Verfahren sind probabilistische Algorithmen, genauer gesagt Adaptionen der LDA [4]. Diese wurden zum einen gewählt, da sie, wie in Abschnitt 2.7 beschrieben wurde, in der Lage sind, das Vorwissen des Ermittlers zu übergehen [14, 15], falls das Thema überhaupt nicht im Datensatz auftritt, und daher besonders zur Überprüfung von Hypothesen im forensischen Kontext geeignet sind. Zum anderen ermöglichen sie es im Gegensatz zu anderen probabilistischen Ansätzen des Seed-Guided Topic Modellings zusätzliche, unüberwachte Themen zu extrahieren [14, 15], was die gleichzeitige Exploration des Datensatzes und das Erkennen von neuen Zusammenhängen erlaubt. Darüber hinaus wurden zwei Erweiterungen der halbüberwachten Algorithmen untersucht, die die Ähnlichkeit von Wörtern basierend auf Word Embeddings beziehungsweise basierend auf paradigmatischen Relationen berücksichtigen, wobei das wesentliche Vorgehen auf dem von Viegas u. a. [16] entwickelten und auf dem in Abschnitt 2.4 erwähnten Ansatz CluWord basierte. Als Baseline diente die Standard-LDA, wie beschrieben von Blei u. a. [4].

Alle Experimente wurden auf einem Datensatz von Kurznachrichten aus einem realen Fall durchgeführt. Für die untersuchten Algorithmen dienten als Eingabe Pseudodokumente, die basierend auf einer automatischen Konversationsdetektion gebildet wurden [2]. Die Themenanzahl für alle Themenmodelle wurde mithilfe der Standard-LDA und zwei quantitativen Evaluierungsmaßen, der semantischen Kohärenz [19] und der FREX [310, 311], bestimmt. Die Hyperparameter α und β wurden sowohl für die Standard-LDA als auch für die halbüberwachten Adaptionen der LDA und ihre Erweiterungen auf dieselben Werte gesetzt, nämlich $\alpha = 0,1$ und $\beta = 0,01$. Mit der Wahl von kleineren Werten für α und β wurde bezweckt, dass wenige Themen in Dokumenten dominieren und wenige Wörter höhere Wahrscheinlichkeiten in den Themen aufweisen, wodurch möglichst differenzierte Verteilungen gebildet werden können [15]. Darüber hinaus wurden diese Werte ebenfalls von Zuo u. a. [119] und Wang u. a. [164] für kurze Texte empfohlen, da es realistischer ist, anzunehmen, dass sich diese sich vor allem auf einige, wenige Themen konzentrieren, für die zudem nicht alle Wörter des meist umfangreichen Vokabulars eine ähnlich hohe Bedeutung aufweisen [148]. Alle Themenmodelle wurden mit 2000 Iterationen trainiert. Auf die Wahl der Werte für die spezifischen Hyperparameter für die beiden Verfahren des Seed-Guided Topic Modellings wird an entsprechender Stelle eingegangen. Die untersuchten Ansätze zur Themenmodellierung wurden sowohl quantitativ als auch durch eine Nutzerstudie [315] evaluiert und verglichen.

In den nachfolgenden Unterabschnitten werden die verwendeten Daten sowie die eingesetzten Methoden detaillierter beschrieben. Die angewendete Vorgehensweise von der Vorverarbeitung der Nachrichten bis hin zur Evaluierung wird zudem in Abbildung 3.1 dargestellt.

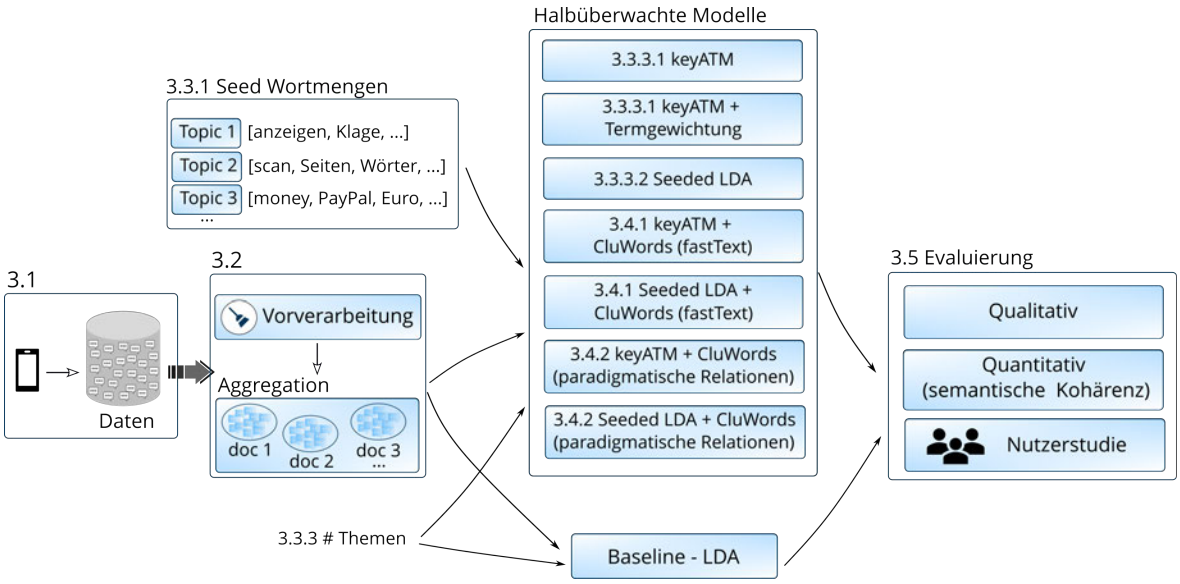

Abbildung 3.1: Vorgehensweise zum Vergleich der verschiedenen Ansätze des Seed-Guided Topic Modellings auf forensischen Kommunikationsdaten. Die Nachrichten wurden für alle Ansätze vorverarbeitet und zu Pseudodokumenten aggregiert. Mit diesen Pseudodokumenten, den gebildeten Seed Wortmengen sowie einer benutzerdefinierten Anzahl an Gesamtthemen als Eingabe wurden zwei halbüberwachte Algorithmen - keyATM und die Seeded LDA - sowie Erweiterungen dieser Ansätze trainiert. Die Evaluierung erfolgte qualitativ, mit einer automatischen quantitativen Evaluierungsmethode sowie mit einer Nutzerstudie. Die in der Abbildung aufgeführten Nummern beziehen sich auf die Abschnitte und Unterabschnitte, in denen die dargestellte Methode erläutert wird.

3.1 Verwendete Daten

Als Datengrundlage für die nachfolgenden Untersuchungen wurden Chats aus einem realen Fall zu dem Delikt der finanziellen Unterstützung einer terroristischen Vereinigung herangezogen. Der Datensatz ist jedoch öffentlich nicht zugänglich. Er umfasst insgesamt 132.669 Nachrichten des Messenger-Dienstes WhatsApp, die auf dem Mobilfunktelefon einer tatverdächtigen Person gespeichert waren. Neben dem eigentlichen Nachrichteninhalt sind ebenfalls Metadaten wie Angaben zu dem Chat, in dem die Nachricht verschickt wurde, dem Absender der Nachricht sowie eine Zeitangabe, die sich aus Datum und Uhrzeit zusammensetzt, vorhanden.

Die Nachrichten wurden in 146 Chats über einen Zeitraum von circa viereinhalb Jahren von Mitte Dezember 2014 bis Mitte Mai 2019 ausgetauscht. Unter den 132.669 Nachrichten befinden sich 96 Nachrichten, die als leer gekennzeichnet sind. Zu den restlichen Nachrichten zählen neben Textnachrichten auch Audionachrichten, Ortsangaben, Bilder, Videos, weitergeleitete Kontakte, versendete Dokumente und verpasste Anrufe. Für die Analysen wurden ausschließlich die insgesamt 118.489 Textnachrichten berücksichtigt, die in 114 Chats verschickt worden waren. Diese sind primär auf deutsch und zu einem geringeren Anteil auf türkisch und arabisch verfasst.

Tabelle 3.1: Statistische Beschreibung des verwendeten Datensatzes. Die Tabelle enthält Angaben zu der Größe des Datensatzes und des Vokabulars sowie zu der durchschnittlichen Frequenz der Wörter in dem Datensatz und der durchschnittlichen Anzahl an Wörtern in einer Nachricht.

Eigenschaft/ Statistik	Ergebnis
# Anzahl Chats	insgesamt 105 (94 Einzelchats, elf Gruppenchats)
# Teilnehmer in Gruppenchats	4-39
# Nachrichten	106.389
Vokabulargröße (# einzigartige Wörter)	36.467
# einzigartige Tokens	39.039
∅ Frequenz der Wörter	22,62
∅ Nachrichtenlänge (in Wörtern)	7,75

Da in dieser Arbeit der Fokus auf monolinguale Themenmodellierung gelegt wurde, wurde der Datensatz mithilfe einer automatischen Sprachdetektion auf deutsche Chats reduziert. Als Verfahren zur Spracherkennung wurden Compact Language Detector 2 (CLD2) [322] und Compact Language Detector 3 (CLD3) [323], die beide zur Integration in den Google Chrome Browser entwickelt wurden, herangezogen. Die Anwendung von zwei Verfahren kann damit begründet werden, dass nach Lui und Baldwin [324] die Kombination der Ergebnisse von verschiedenen Methoden bei kurzen, umgangssprachlichen Texten als zuverlässiger zu betrachten ist als das Ergebnis eines einzelnen Verfahrens. Es wurden die Nachrichten aller Chats entfernt, die sowohl von CLD2 als auch CLD3 als nicht-deutsch eingestuft worden waren. Dies führte zu einer Reduktion des Datensatzes auf 105 Chats, die circa 106.000 Nachrichten umfassten. Eine Übersicht über die Eigenschaften des auf diese Weise extrahierten Datensatzes kann Tabelle 3.1 entnommen werden.

3.2 Vorverarbeitung und Aufbereitung der Daten

Die Nachrichten wurden zunächst bereinigt und anschließend zu Pseudodokumenten aggregiert.

3.2.1 Vorverarbeitung der Daten

Wie von Churchill und Singh [187] in umfangreichen Experimenten auf der Grundlage von Social-Media-Daten mit verschiedenen probabilistischen Themenmodellen gezeigt wurde, hat die Datenbereinigung einen starken Einfluss auf die Qualität der Themen. Sie betonen hierbei, dass gerade für rauschbehaftete Daten, zu denen ebenfalls die Handynachrichten zählen, eine möglichst gründliche Vorverarbeitung für gute Resultate bei der Themenmodellierung von hoher Bedeutung ist. Dies umfasste die Durchführung der folgenden Schritte:

1. Entfernung von überflüssigem Whitespace
2. Entfernung von Weblinks, Email-Adressen und Erwähnungen inklusive des @-Zeichens, da diese nicht zum Inhalt beitrugen
3. Entfernung von Emojis, da diese größtenteils in der TWV ohne weiteren Kontext über eine mangelhafte Interpretierbarkeit verfügen

4. Entfernung von Satz- und Sonderzeichen mit Ausnahme von Währungszeichen, die für den Fall über die finanzielle Unterstützung einer terroristischen Vereinigung relevant sein könnten
5. Entfernung von Zahlen, da diese wie Emojis ohne weiteren Kontext eine geringe Bedeutung besitzen
6. Entfernung von deutschen, türkischen und englischen Stoppwörtern unter Verwendung der von Diaz [325] bereitgestellten Stoppwortlisten. Die Entfernung der türkischen und englischen Stoppwörter sowie der islamischen Redewendungen war trotz der in Abschnitt 3.1 beschriebenen Reduktion des Datensatzes auf deutsche Nachrichten erforderlich, da nicht ausgeschlossen werden konnte, dass türkische und arabische Floskeln oder Anglizismen auch in Nachrichten verwendet wurden, die als deutsch klassifiziert wurden.
7. Entfernung von folgenden weiteren Begriffen, die als Noise Words eingestuft wurden, um möglichst dem in Abschnitt 2.3.2 beschriebenen Problem entgegenzuwirken, dass die Themen durch hochfrequente, bedeutungslose Begriffe dominiert werden:
   - die 200 häufigsten Wörter und die 100 Wörter mit der geringsten Inverse Dokumentenfrequenz (IDF)
   - Wörter, bei denen aus Analysen in [258] bekannt war, dass sie keine Relevanz für den Datensatz besaßen
   - alle Modul- und Auxiliarverben sowie häufig verwendete deutsche Verben, die manuell von [326] ausgewählt wurden
   - Islamische Redewendungen, die [327] entnommen wurden
   - Akronyme des Netzjargons, die aus [328] manuell selektiert wurden
8. Lemmatisierung mithilfe des TreeTaggers [329, 330], um zu ermöglichen, dass gebeugte Formen eines Wortes mit derselben Bedeutung als eine einzelne Einheit repräsentiert werden und zur Verminderung des Problems der hohen Sparsität durch Reduktion des Vokabulars [331]
9. Umwandlung in Kleinschreibung zur Normalisierung
10. Tokenisierung in Terme

Die Tokenisierung in Unigramme als letzter Vorverarbeitungsschritt wurde gegenüber der Tokenisierung in Bigramme oder größere N-Gramme bevorzugt, da letztere zwar teilweise aussagekräftiger sein können, aber diese zu einem größeren Vokabular führen und sich daher laut Churchill und Singh [187] insbesondere bei Datensätzen mit hoher Sparsität negativ auf die Qualität der Themen auswirken können.

3.2.2 Aggregation zu Pseudodokumenten

Nach der Vorverarbeitung bestanden die Nachrichten durchschnittlich aus weniger als drei Wörtern. Um dem Problem der geringen Länge der Nachrichten entgegenzuwirken, wurde die in Abschnitt 2.4 beschriebene Strategie angewendet, die vorverarbeiteten Nachrichten zu längeren Pseudodokumenten mithilfe eines Cluster-Verfahrens zu aggregieren. Hierfür wurde das von Spranger u. a. [2] beschriebene Verfahren eingesetzt, dass die Nachrichten, die innerhalb eines Chats verschickt wurden, basierend auf ihrem zeitlichen Kontext zu möglichst zusammenhängenden Konversationen gruppiert. Wie in Spranger u. a. [2, 3] ausführlicher beschrieben wurde, umfasst eine Konversation c die Nachrichten m eines Chats, die zu den

Zeitpunkten t_i und t_{i+1} fortlaufend ausgetauscht wurden, ohne eine individuell geschätzte Antwortzeit ϵ zu überschreiten. Somit wird eine Konversation formal wie durch Gleichung (3.1) definiert [2].

$$c = (m_1, ..., m_n | t_i^m - t_{i+1}^m <= \epsilon, \forall i = 1...n) \tag{3.1}$$

Aufgrund der Berücksichtigung der individuell bestimmten Antwortzeit wurde diese Strategie als flexibler betrachtet als beispielsweise alle Nachrichten, die an einem Tag ausgetauscht wurden, als ein Pseudodokument zu betrachten. Die Konversationsdetektion wurde durch den Mobile Network Analyzer (MoNA) [2], einem forensischen Tool zur Auswertung und Analyse von mobilen Kommunikationsdaten durchgeführt.

Tabelle 3.2: Eigenschaften der gebildeten Konversationsdokumente als Eingabe für die Themenmodellierung. Die Anzahl der gebildeten Konversationsdokumente und die minimale, maximale und durchschnittliche Anzahl an Nachrichten, aus denen eine Konversation bestand sowie die durchschnittliche Länge der Konversationen nach der Vorverarbeitung werden angegeben.

Eigenschaft	**Angabe**
# Konversationsdokumente	15.625
Minimale Anzahl an Nachrichten in einer Konversation	1
Maximale Anzahl an Nachrichten in einer Konversation	349
∅ Anzahl an Nachrichten je Konversation	≈ 8
Konversationslänge nach Vorverarbeitung (in Wörtern)	13,08

Informationen über die gebildeten Konversationsdokumente können Tabelle 3.2 entnommen werden. Sie dienten als Eingabe für alle Algorithmen der Themenmodellierung, sowohl des Baseline-Modells als auch aller halbüberwachten Ansätze.

3.3 Umsetzung des Seed Guided Topic Modellings

In diesem Abschnitt soll die Umsetzung des Seed Guided Topic Modellings basierend auf den beiden Algorithmen, keyATM [14] und Seeded LDA [15], vorgestellt werden. Beide Algorithmen extrahieren neben den $\tilde{K}$ Themen, die durch die Seed Wortmengen beschrieben werden und in dieser Arbeit als „Seeded Topics“ bezeichnet werden, ebenfalls eine benutzerdefinierte Anzahl an weiteren, unüberwachten Themen, hier bezeichnet als „Unseeded Topics“. Im Folgenden wird zunächst auf die Vorgehensweise zur Erstellung der Seed Wortmengen sowie zur Schätzung einer geeigneten Themenanzahl eingegangen, bevor die beiden verwendeten Algorithmen genauer beschrieben werden.

3.3.1 Erstellung der Seed-Wortmengen

Zunächst wurde für jedes gewünschte Thema eine Seed-Wortmenge erstellt. Diese enthielt möglichst charakteristische Begriffe für das entsprechende Thema und diente für alle durchgeführten Experimente basierend auf den beiden halbüberwachten Algorithmen [14, 15] als Eingabe. Eine häufige Vorgehensweise zur Konstruktion der Seed-Wortmengen, die beispielsweise ebenfalls von Watanabe und Baturo [15] zur Durchführung der Seeded LDA sowie ferner auch durch Zha und Li [230] zur Anwendung des Seed-Guided Topic Modellings eingesetzt wurde, besteht darin, als Seed Wörter für jedes Thema die Wörter mit der höchsten Wahrscheinlichkeit in Themen zu verwenden, die zuvor mit einer unüberwachten LDA auf dem Datensatz gefunden wurden. Jedoch ist ein Nachteil dieser Vorgehensweise darin zu sehen, dass die vorherige Durchführung der LDA auf den teilweise umfangreichen Falldaten mit einem hohen Zeitaufwand verbunden ist. Zudem kann ebenfalls nicht sichergestellt werden, dass es sich tatsächlich um fallrelevante Begriffe handelt, die Themen beschreiben, die für den Ermittler von Interesse sind.

Stattdessen wurden die Seed Wortmengen basierend auf dem von Spranger u. a. [2] detaillierter beschriebenen Termbaum erstellt. Bei dem Termbaum handelt es sich um ein Wissensmodell, das ursprünglich zur automatischen Detektion von fallrelevanten Nachrichten mithilfe eines regelbasierten Ansatzes entwickelt wurde. Dieses beschreibt ein komplexes System von sogenannten Syntagmen, worunter in diesem Kontext fallrelevante Terme verstanden werden, die gemeinsam in einer Nachricht vorkommen [2]. Der Aufbau dieses Termbaums kann Abbildung 3.2 entnommen werden.

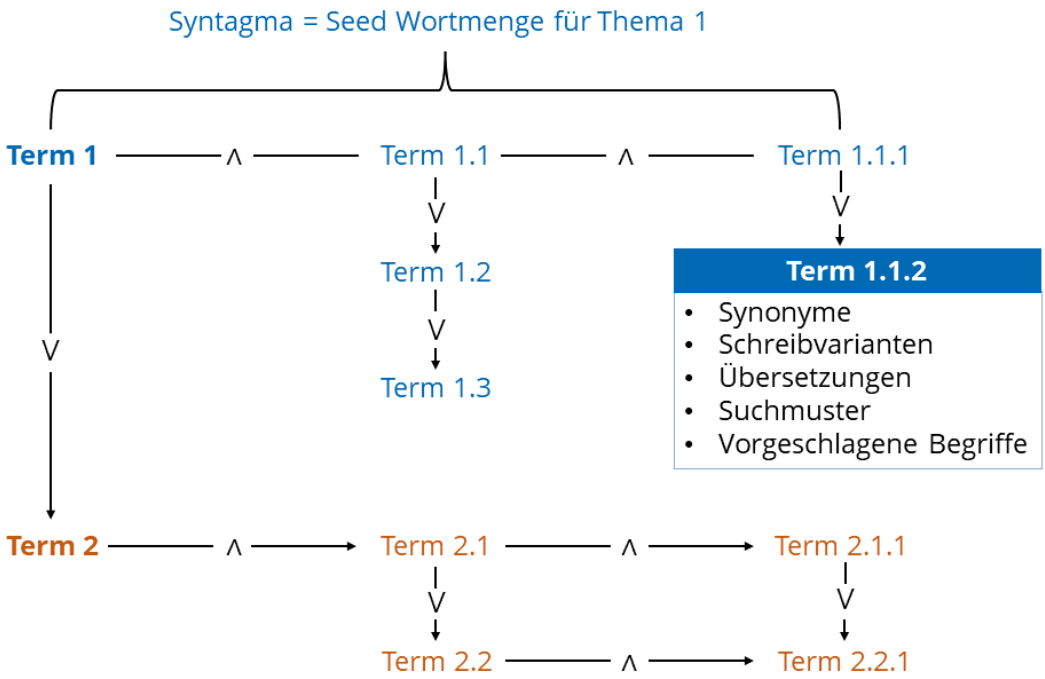

Abbildung 3.2: Aufbau des Termbaums als Grundlage zur Erstellung der Seed-Wortmengen. Die blau und orange hervorgehobenen Begriffe bilden jeweils ein Syntagma in dem Begriffsbaum, das als Seed-Wortmenge eines Themas diente. Bei den fett gekennzeichneten Begriffen „Term 1“ und „Term 2“ handelt es sich um den Startbegriff des Syntagmas, der als Topic Label herangezogen wurde. Quelle: Adaptiert aus Spranger u. a. [2].

Tabelle 3.3: Übersicht über die erstellten Seed Wortmengen für die halbüberwachte Themenmodellierung mit keyATM und Seeded LDA. Angegeben sind die ausgewählten Themenlabel, die Anzahl der Seed Wörter und jeweils fünf ausgewählte repräsentative Seed Wörter aus den insgesamt acht gebildeten Seed Wortmengen. Jede der erstellten Seed Wortmengen diente dazu, die halbüberwachten Algorithmen dazu zu verleiten, ein gewünschtes, fallrelevantes Thema zu identifizieren.

Topic Label	**#Seed Wörter**	**Seed Wörter (Auswahl)**
ankommen	7	ankommen, Condor, Postboten, Nachforschungsauftrag, bhf
Anwalt	18	anzeigen, beschuldigen, Klage, Fachgebiet, ermitteln
Buch	8	scan, Seiten, Wörter, Kapitel, Leseratte
Geld	64	money, PayPal, Euro, überweisen, Zahlung
Polizei	7	rufen, durchleuchten, Absperrbänder, Kripo, Verwarnung
Schwester	7	Nalan, Sister, jung, Stiefbruder, Eheangelegenheiten
Terror	25	Anschlag, Vereinigung, Waffe, Gewalt, Nordirland
Verein	23	Vereinsregister, Rechtspersönlichkeit, Vereinsgründer, Club, Dokumentablagen

Ein Term t wird in dem Wissensmodell nicht durch ein einzelnes Wort repräsentiert, sondern als Vektor $\vec{t} = (w_0, ..., w_n, p_0, ..., p_k)$ dargestellt [2]. Hierbei beschreibt w_i eine Menge an sprachlichen Variationen wie Synonyme, Schreibvarianten, gruppenspezifische Ausdrücke und Übersetzungen. Darüber hinaus enthält sie weitere zu t relevante Begriffe, die mithilfe eines von Felser u. a. [258] beschriebenen automatischen Verfahrens zur Begriffsempfehlung basierend auf syntagmatischen Relationen aus dem Datensatz extrahiert wurden [2]. Das entsprechende Vorschlagsystem ist Bestandteil der Software MoNA [2]. p_i bezeichnet eine Menge an regulären Ausdrücken [2]. Bei einem Syntagma syn handelt es sich somit um eine Konjunktion von Termen t_i in einer Nachricht m_j, was formal als $syn = t_0 \wedge t_1 \wedge ...$ beschrieben werden kann.

Jedes Syntagma diente als Seed-Wortmenge für ein Thema. Der in dieser Arbeit verwendete Begriffsbaum war ursprünglich von dem zuständigen Staatsanwalt für den Fall erstellt worden und wurde mithilfe des Vorschlagsystems von MoNA [2, 258] um weitere fallrelevante Begriffe ergänzt. Auf diese Weise wurden insgesamt acht Seed-Wortmengen gebildet. Beide verwendeten halbüberwachte Algorithmen, keyATM und die Seeded LDA setzen voraus, dass die Seed Wörter in dem Vokabular des Trainingsdatensatzes enthalten sind [14, 15]. Daher wurden alle Begriffe der Seed-Wortmengen, die nicht im Datensatz vorkamen entfernt. Sowohl keyATM [14] als auch die Seeded LDA [15] ermöglichen es, vor der Anpassung des Modells einen Begriff aus jeder Seed Wortmenge als Label für das gewünschte Thema auszuwählen. Hierbei wurde jeweils der Startbegriff eines Syntagmas, der in Abbildung 3.2 fett hervorgehoben wurde, als Themenlabel ausgewählt, da dieser als Art Oberbegriff betrachtet werden konnte. Exemplarisch sind die Themenlabel der acht Seed Wortmengen, fünf repräsentative Seed Wörter und die Anzahl der Seed Wörter jeder Wortmenge (einschließlich des Topic Labels) in Tabelle 3.3 dargestellt.

Es ist darauf hinzuweisen, dass Eshima u. a. [14] betont, dass die Seed Wortmengen nicht disjunkt sein müssen. Hier teilte sich die Seed Wortmenge für das Thema „Terror“ einen identischen Begriff mit dem Thema „Geld“, nämlich „PayPal“ sowie das Wort „vereinen“ mit dem Thema „Verein“.

3.3.2 Bestimmung der Themenanzahl

Sowohl bei keyATM [14] als auch bei der Seeded LDA [15] entsprach die Anzahl der „‘Seeded Topics“ $\tilde{K}$ der Anzahl der erstellten Seed Wortmengen. Hingegen war es erforderlich, die Anzahl der zusätzlichen „Unseeded Topics“‘, hier bezeichnet mit K' anzugeben [14, 15]. Die Vorgehensweise zur Bestimmung der Anzahl an „Unseeded Topics“ K' ist in Abbildung 3.3 skizziert.

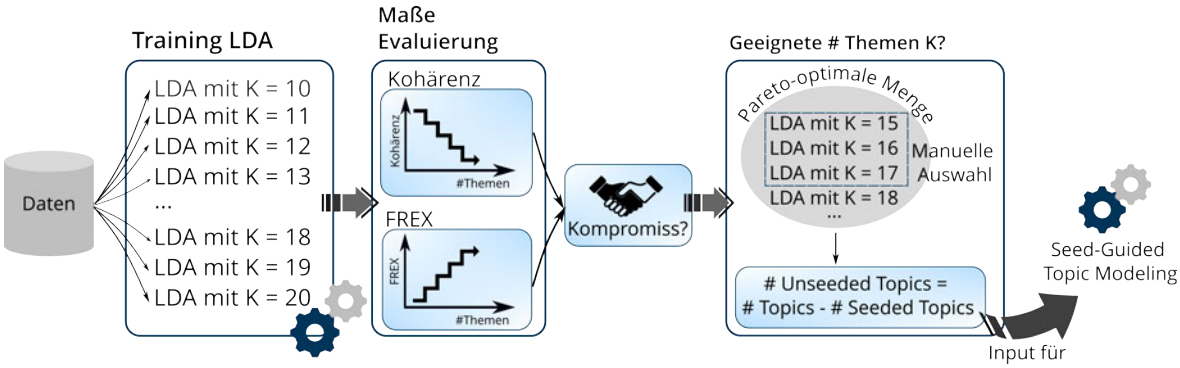

Abbildung 3.3: Vorgehensweise zur Ermittlung einer geeigneten Anzahl an „Unseeded Topics“ K'. Es wurden insgesamt elf Themenmodelle mit einer Themenanzahl von $K \in [10, 20]$ unter Verwendung der Standard-LDA trainiert. Die Themenmodelle, die einen guten Kompromiss zwischen zwei quantitativen Evaluierungsmaßen - der semantischen Kohärenz [19] und FREX [310, 311] - darstellen, wurden ermittelt. Ausgehend von ihrer Themenanzahl wurde eine geeignete Anzahl von „Unseeded Topics“ K' für die Algorithmen des Seed-Guided Topic Modellings bestimmt.

Die grundlegende Idee bestand darin, einen Überblick über die Gesamtanzahl K an besprochenen Themen in dem Kommunikationsdatensatz zu erhalten, indem eine geeignete Themenanzahl für das Baseline-System, die unüberwachte LDA [4], ermittelt wurde. Hierzu wurden eine Reihe von Kandidatenmodellen unter Verwendung der LDA trainiert und mithilfe von zwei quantitativen Evaluierungsmaßen eine geringe Anzahl an geeigneten Themenmodellen ausgewählt, was der von Weston u. a. [332] empfohlenen Vorgehensweise entsprach. Genauer gesagt wurden Themenmodelle bestimmt, die einen möglichst guten Kompromiss zwischen der semantischen Kohärenz [19] und der FREX [310, 311] darstellten. Hierzu wurde die pareto-optimale Menge von Themenmodellen in Bezug auf die beiden Evaluierungsmaße bestimmt.

Ausgehend von der Themenanzahl K dieser Modelle wurde die Anzahl an „Unseeded Topics“ als Eingabe für die Algorithmen des Seed-Guided Topic Modellings gewählt. Hierzu wurde die bekannte Anzahl der „Seeded Topics“ $\tilde{K}$ von dem geschätzten K subtrahiert, d.h. $\#\textit{Unseeded Topics} = K' = K - \tilde{K} = K - 8$, da acht Seed Wortmengen gebildet wurden. Von Eshima u. a. [14] und Watanabe und Baturo [15] wurde betont, dass, selbst wenn die Exploration des Datensatzes nicht von Interesse ist, zumindest eine geringe Anzahl an „Unseeded

Topics“ benötigt wird, falls die durch die Seed Wörter beschriebenen Themen den Inhalt des Datensatzes nicht vollständig beschreiben. Hierbei wird bezweckt, dass es sich bei den „Unseeded Topics“ um die dominanten Themen des Datensatzes handelt und die interessanten „Seeded Topics“ nicht mit für sie irrelevanten, allgemeinen Wörtern vermischt werden [14, 15]. Da in dieser Arbeit davon auszugehen war, dass die interessanten, fallrelevanten Themen nur einen geringen Anteil der Konversationen ausmachen, sollten mindestens zwei „Unseeded Topics“ erstellt werden. Dementsprechend wurden insgesamt elf Themenmodelle mit einer Themenanzahl von $K \in [10, 20]$ mittels der LDA trainiert, was somit einer möglichen Anzahl von zwei bis zwölf „Unseeded Topics“ entsprach.

Eine alternative Vorgehensweise zur Bestimmung der Anzahl der „Unseeded Topics“ würde darin bestehen, wie beispielsweise von Nikolenko u. a. [12] vorgeschlagen wurde, die einzelnen Algorithmen des Seed-Guided Topic Modellings mit unterschiedlichen K' zu lernen und von ihnen jeweils die besten Modelle zu bestimmen. Jedoch könnte dies in unterschiedlich vielen „Unseeded Topics“ der verschiedenen Themenmodelle resultieren, was die Vergleichbarkeit der Ergebnisse und die Evaluierung erschweren würde. Daher wurde die Schätzung der Themenanzahl basierend auf der Standard-LDA bevorzugt. Im Folgenden sollen die Evaluierungsmaße und die Bestimmung des Kompromisses detaillierter beschrieben werden.

3.3.2.1 Semantische Kohärenz

Als eines der beiden quantitativen Evaluierungsmaße wurde die von Mimno u. a. [19] beschriebene semantische Kohärenz gewählt, die auch unter der Bezeichnung C_{UMass} bekannt ist [308]. Sie wurde ebenfalls beispielsweise von Wang u. a. [51] und Chen u. a. [225] zur Wahl der Themenanzahl angewendet. Die semantische Kohärenz wurde gegenüber den anderen in Abschnitt 2.11 genannten Kohärenzmaßen bevorzugt, da es sich um ein intrinsisches Kohärenzmaß handelt. Extrinsische Kohärenzmaße wie C_{NPMI} [309] und C_{UCI} [150, 309] wurden hingegen als ungeeignet betrachtet, da, wie in Abschnitt 2.11 erläutert wurde, die Informationen darüber, inwiefern die wahrscheinlichsten Wörter eines Themas miteinander assoziiert werden nicht zwangsläufig auf die speziellen Charakteristika von forensischen Kommunikationsdaten zutreffen. Dies kann damit begründet werden, dass in forensischen Kurznachrichten die Wörter in ungewöhnlicher Bedeutung verwendet werden können [133]. Hinzu kommt, dass gerade in externen Datensätzen, die aus Dokumenten wie Wikipedia-Artikeln bestehen, die beispielsweise in [123, 141, 229] verwendet wurden, viele der umgangssprachlichen Wörter aus den Kurznachrichten nicht vorkommen und somit für die Berechnung der Kohärenz auf einem externen Datensatz nicht berücksichtigt werden könnten.

Die semantische Kohärenz basiert auf der Annahme, dass die wahrscheinlichsten Wörter eines Themas semantisch zusammenhängen, falls diese häufig gemeinsam in den Dokumenten des Trainingsdatensatzes des Themenmodells auftreten [19]. Zur Berechnung der semantischen Kohärenz eines ausgewählten Themas t werden die N wahrscheinlichsten Wörter von t betrachtet, also $V^{(t)} = (v_1^{(t)}, ..., v_N^{(t))}$ [19], wobei in dieser Arbeit $N = 10$ gewählt wurde. Wie in Abschnitt 2.11 erläutert wurde, wird zur Ermittlung der Kohärenz C zwischen diesen N Wörtern ein Bestätigungsmaß berechnet [308]. Im Falle der semantischen Kohärenz nach Mimno u. a. [19] dient als Bestätigungsmaß eine Adaption der PMI [191], die im Gegensatz

zu der ursprünglichen Definition des PMI asymmetrisch ist und somit die Reihenfolge der Wörter in der Rangliste der wahrscheinlichsten Begriffe berücksichtigt [19]. Genauer gesagt kann die semantische Kohärenz für ein Thema t mit Gleichung (3.2) berechnet werden.

$$C(t; V^{(t)}) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} \log \frac{D(v_n^t, v_l^t) + \epsilon}{D(v_l^t)} \tag{3.2}$$

Hierbei gibt $D(v)$ die Dokumentenfrequenz des Wortes v an und $D(v, v')$ steht für die Anzahl der Dokumente, in denen die Wörter v und v' gemeinsam auftreten [19]. Hierzu wurde das Vorkommen von Wörtern innerhalb der gebildeten Konversationsdokumente betrachtet anstatt, wie bei Mehrotra u. a. [153], in den Originaldokumenten beziehungsweise den Nachrichten. Auf diese Weise sollte dem von Quan u. a. [157] beschriebenen Problem entgegengewirkt werden, dass Kohärenzmaße basierend auf dem gemeinsamen Auftreten von Wörtern bei kurzen Texten teilweise unzuverlässige Aussagen über die Qualität von Themen liefern. Der Glättungsparameter ϵ, der verhindert, dass der Logarithmus von null für Wörter genommen wird, die nie gemeinsam vorkommen, wurde wie bei Mimno u. a. [19] auf $\epsilon = 1$ gesetzt.

Eine höhere semantische Kohärenz indiziert Themen mit höherer Qualität [19]. Um einen Gesamtkohärenzwert für ein Themenmodell mit einer bestimmten Themenanzahl K zu erhalten, wurde, wie von Mimno u. a. [19] vorgeschlagen wurde, das arithmetische Mittel der Kohärenzwerte aller Themen des Modells berechnet.

3.3.2.2 FREX

Ein Problem bei der Schätzung der optimalen Themenanzahl basierend auf der semantischen Kohärenz besteht nach Roberts u. a. [11] darin, dass eine hohe semantische Kohärenz durch Themenmodelle erreicht werden kann, die aus wenigen, von allgemeinen Begriffen dominierten Themen bestehen. Mit anderen Worten tendiert somit die semantische Kohärenz dazu, Modelle mit wenigen, überlappenden Themen zu bevorzugen. Daher wurde von Roberts u. a. [11] vorgeschlagen, neben der semantischen Kohärenz ebenfalls die von Airoldi und Bischof [310] und Bischof und Airoldi [311] beschriebene FREX bei der Wahl der Themenanzahl zu berücksichtigen. Diese wurde ursprünglich zur Bestimmung von Topic Labels entwickelt und auf der Ebene einzelner Wörter definiert, wobei ein hoher Wert für ein Wort v in einem Thema k indiziert, dass das Wort eine hohe Wahrscheinlichkeit in k aufweist, jedoch in allen anderen Themen mit einer geringen Wahrscheinlichkeit auftritt [310, 311]. Dementsprechend gelten Wörter mit einem hohen FREX-Wert zugleich als wichtig und spezifisch für das entsprechende Thema [310, 311]. Der FREX-Wert eines Wortes v in einem Thema k kann durch Gleichung (3.3) [310, 311] berechnet werden, wobei $ECDF_{x_{,.k}}$ für empirical cumulative distribution function (eCDF), die empirische Verteilungsfunktion, steht.

$$FREX_{v,k} = \left(\frac{w}{ECDF(\mu_{v,k})} + \frac{1-w}{ECDF(\phi_{v,k})} \right)^{-1} \tag{3.3}$$

Somit wird der FREX-Wert von v in k als das gewichtete harmonische Mittel aus dem Rang des Wortes v in den beiden Verteilungen $\mu_{.,k}$ und $\phi_{.,k}$ des Themas k definiert [310, 311]. Hierbei handelt es sich bei $\phi_{.,k}$ um die Wortverteilung eines Themas k, die direkt als Ausgabe von probabilistischen Themenmodellen geliefert wird [310, 311]. $\mu_{.,k}$ steht für die Verteilung der Exklusivität der Wörter in dem Thema k, wobei die Exklusivität eines Wortes v in k mit Gleichung (3.4) berechnet werden kann [310, 311].

$$\mu_{v,k} = \phi_{v,k} / \sum_{j \in S} \phi_{v,j} \tag{3.4}$$

Dementsprechend wird zur Ermittlung von $\mu_{v,k}$ die Wahrscheinlichkeit $\phi_{v,k}$ des Wortes v in dem Thema k mit seiner Wahrscheinlichkeit in allen anderen Themen S des Modells verglichen [310, 311]. Der Parameter w in Gleichung (3.3) kontrolliert, ob der Wahrscheinlichkeit des Wortes oder seiner Exklusivität in dem Thema eine höhere Bedeutung zugemessen wird [310, 311]. In dieser Arbeit wurde $w = 0.7$ gewählt, wie beispielsweise von Roberts u. a. [333] vorgeschlagen wurde, wodurch die Exklusivität höher gewichtet wurde als die Häufigkeit [310].

Um einen FREX-Wert für ein Thema zu erhalten, wurden die einzelnen FREX-Werte der N Wörter mit der höchsten Wahrscheinlichkeit in diesem Thema summiert, was der Vorgehensweise von Roberts u. a. [333] entsprach. Analog zu der semantischen Kohärenz wurde $N = 10$ festgelegt. Der FREX-Wert für ein Thema kann Werte im Bereich von $[0, +\infty]$ annehmen, wobei höhere Werte darauf hinweisen, dass ein Thema einzigartig ist und seine wahrscheinlichsten Wörter nicht ebenfalls in den anderen Themen über hohe Wahrscheinlichkeiten verfügen [334]. Für jedes gelernte Themenmodell wurde wiederum das arithmetische Mittel der FREX Werte jedes Themas ermittelt.

3.3.2.3 Kompromiss aus semantischer Kohärenz und FREX

Nach Roberts u. a. [11] sollten bedeutungsvolle Themen sowohl über eine hohe semantische Kohärenz als auch eine hohe FREX verfügen. Hinsichtlich der Bestimmung der Themenanzahl ist jedoch ein Problem darin zu sehen, dass ein Trade-off zwischen der semantischen Kohärenz und der FREX besteht [332]. Üblicherweise wird die FREX mit einer steigenden Anzahl an Themen verbessert, während die semantische Kohärenz meist bei einer geringen Anzahl an Themen ihr Maximum erreicht [13]. Um geeignete Themenanzahlen zu ermitteln, wurden daher Modelle bestimmt, bei denen ein möglichst guter Kompromiss zwischen der semantischen Kohärenz und der FREX besteht. Hierzu wurde der von Lee und Kolodge [335] und

Rutkowski u. a. [336] vorgeschlagene Ansatz angewendet, Strategien der Pareto-Optimierung einzusetzen. Diese können grundsätzlich dazu eingesetzt werden, zwei komplementäre Ziele simultan zu maximieren [337, 338].

In diesem Kontext enthält die pareto-optimale Menge $\max_{<_\otimes} M$ alle Themenmodelle m, die nicht durch ein anderes Themenmodell m' pareto-dominiert werden, was in Gleichung (3.5) resultiert.

$$\max_{<_\otimes} M = \{m \in M | \nexists m' \in M : m <_\otimes m'\} \tag{3.5}$$

Hierbei wird ein Themenmodell m von einem anderen Themenmodell m' hinsichtlich der semantischen Kohärenz C und der FREX pareto-dominiert, wenn m' in Bezug auf die semantischen Kohärenz oder FREX besser als m ist und in dem anderen Evaluierungsmaß nicht schlechter ist als m. Dies kann formal durch Gleichung (3.6) beschrieben werden.

$$\begin{aligned} m <_{C \otimes FREX} m' :\Leftrightarrow & (m \leq_C m' \wedge m <_{FREX} m') \vee \\ & (m <_C m' \wedge m \leq_{FREX} m') \end{aligned} \tag{3.6}$$

Zur Berechnung dieser pareto-optimalen Menge wurde der von Endres u. a. [337] entwickelte Algorithmus Scalagon eingesetzt. Dieser vergleicht die Werte der Themenmodelle in Bezug auf die beiden Evaluierungsmaße, wobei er zudem zur Effizienzsteigerung eine Vorfilterung basierend auf einem gitterbasierten Algorithmus [339] vornimmt.

Ausgehend von der Themenanzahl aller Themenmodelle von $\max_{<_\otimes} M$ wurden geeignete Anzahlen an „Unseeded Topics“ K' bestimmt. Mit diesen K' als Eingabe wurden für jeden untersuchten Algorithmus des Seed-Guided Topic Modellings mehrere Themenmodelle gelernt.

3.3.3 Verwendete Algorithmen

Zwei grundsätzliche Ansätze zum Seed-Guided Topic Modelling wurden miteinander verglichen: das von Eshima u. a. [14] vorgeschlagene keyATM, das auf einem von Jagarlamudi u. a. [27] entwickelten Algorithmus aufbaut und die Seeded LDA, die ursprünglich von Lu u. a. [232] vorgestellt und von Watanabe und Baturo [15] verbessert wurde. Die beiden Algorithmen unterscheiden sich insofern, welche der in Abschnitt 2.7 vorgestellten Strategien sie anwenden, um das Vorwissen in Form von Seed Wortmengen in die LDA [4] zu integrieren. Beide zielen darauf ab, dass den Seed Wörtern eine höhere Bedeutung in dem entsprechenden „Seeded Topic“ zugemessen wird und zudem mit ihnen assoziierte Wörter, basierend auf dem gemeinsamen Vorkommen in Dokumenten, eine hohe Wahrscheinlichkeit in dem entsprechenden Thema aufweisen [14, 15, 27].

3.3.3.1 keyATM

Wie bereits in Abschnitt 2.7 kurz beschrieben wurde, besteht die grundlegende Idee von keyATM darin, jedes „Seeded Topic“ als eine Zusammensetzung aus der gewöhnlichen Wahrscheinlichkeitsverteilung über alle Wörter des Vokabulars ϕ_k einschließlich der Seed Wörter und einer Wahrscheinlichkeitsverteilung $\tilde{\phi}_k$, die sich nur über Seed Wörter erstreckt, zu definieren [14]. Es sollte darauf hingewiesen werden, dass nur die Seed Wörter vorgegeben werden müssen und die Wahrscheinlichkeitsverteilung über diese Begriffe $\tilde{\phi}_k$ wie die gewöhnliche Themen-Wort-Verteilung durch das Modell abgeleitet wird [27]. Ein „Unseeded Topic“ wird weiterhin durch eine einzige Wahrscheinlichkeitsverteilung über alle Wörter wie bei der Standard-LDA [4] beschrieben [14].

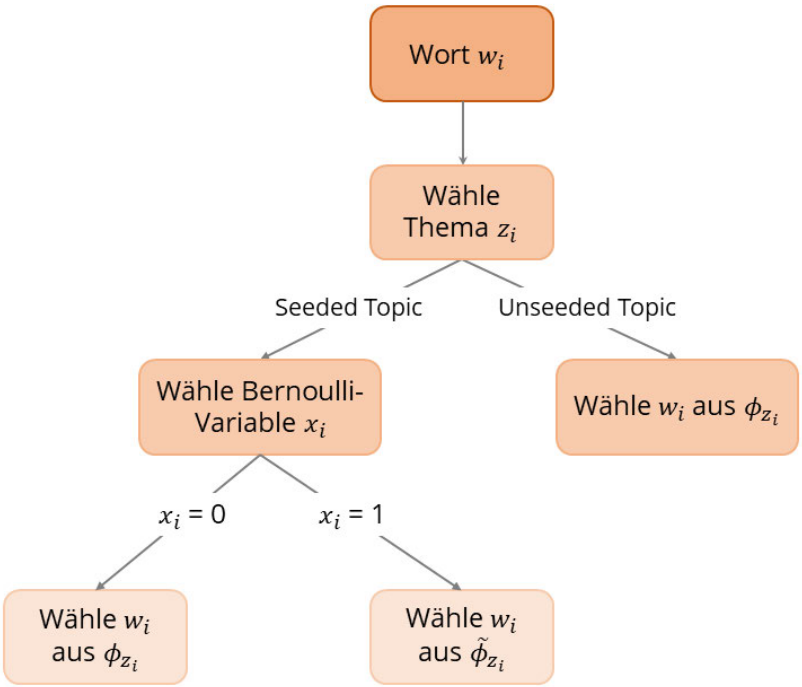

Abbildung 3.4: Skizze des generativen Prozesses von keyATM. Bei einem gewählten Thema z_i für ein Wort w_i kann es sich entweder um ein „Seeded Topic“ oder ein „Unseeded Topic“ handeln. Während bei einem „Unseeded Topic“ w_i aus der gewöhnlichen Themen-Wortverteilung ϕ_{z_i} gewählt wird, kann w_i bei einem „Seeded Topic“ entweder mit ϕ_{z_i} oder $\tilde{\phi}_{z_i}$ generiert werden.

Eshima u. a. [14] adaptierten für das keyATM den generativen Prozess der Standard-LDA. Der grundsätzliche Ablauf, wie ein Wort w_i eines Dokuments d generiert wird, wird in Abbildung 3.4 dargestellt. Wie bei der ursprünglichen LDA wird für w_i eine Themenzuweisung z_i aus der Themenverteilung θ_d des Dokuments d gewählt [14]. Bei θ_d handelt es sich weiterhin um eine Wahrscheinlichkeitsverteilung über die insgesamt K Themen des Modells, zu denen $\tilde{K}$ „Seeded Topics“ zählen. Handelt es sich bei z_i um ein „Unseeded Topic“, wird das Wort w_i aus der Wortverteilung des entsprechenden Themas ϕ_{z_i} gewählt [14]. Ist hingegen z_i ein „Seeded Topic“, bestehen im Gegensatz zu der Standard-LDA zwei Möglichkeiten für Wortverteilungen, mit denen w_i generiert werden kann: entweder mit der gewöhnlichen

Wortverteilung ϕ_{z_i} des Themas oder mit der speziellen Wortverteilung $\tilde{\phi}_{z_i}$, die ausschließlich Seed Wörter umfasst [14]. Von welchen der beiden Verteilungen ein Wort gewählt wird, wird, wie Abbildung 3.4 entnommen werden kann, durch die Indikatorvariable x_i angezeigt.

Der Wert dieser Indikatorvariablen wird durch eine Bernoulli-Verteilung für jedes „Seeded-Topic“ k mit der Erfolgswahrscheinlichkeit π_k festgelegt [14]. Diese Erfolgswahrscheinlichkeit π_k steuert somit den Einfluss der Seed-Wörter auf die Themen und wird wie die anderen Modellparameter gelernt, wobei für diese eine Beta-Verteilung $Beta(\gamma_1, \gamma_2)$ als a-priori-Verteilung eingeführt wird.

Um ein „Seeded Topic“ als eine einzelne Wahrscheinlichkeitsverteilung über Wörter darzustellen, werden schließlich die beiden Verteilungen ϕ_k und $\tilde{\phi}_k$ durch Gleichung (3.7) in Bezug auf ein Wort w und ein Thema k kombiniert [14].

$$\phi*_{kw} = (1 - \pi_k)\phi_{kw} + \pi_k\tilde{\phi}_{kw} \tag{3.7}$$

Hinsichtlich des verwendeten Inferenz-Verfahrens schlugen die Autoren zwei Varianten vor: Zum einen wurde der ebenfalls bei der Standard-LDA häufig angewendete Algorithmus Collapsed Gibbs Sampling [47] gewählt. Zum anderen beschrieben Eshima u. a. [14] eine Version von keyATM, bei der eine Adaption von Collapsed Gibbs Sampling zum Einsatz kommt, die ursprünglich von Wilson und Chew [189] für die in Abschnitt 2.5 beschriebene wLDA entwickelt wurde. Diese zielt darauf ab, die bedeutungslosen, allgemeinen Begriffe durch die Einführung eines Termgewichtungsschemas in den Inferenzalgorithmus abzuwerten, wobei die Gewichtung der Terme je nach Dokument varriert und das Termgewicht eines Wortes $m(w_i)$ in einem Dokument d durch Gleichung (3.8) angegeben wird, die auf dem PMI basiert.

$$m(w_i) = -\log_2 \frac{p(w_i|d)}{p(w_i)} \tag{3.8}$$

Die Version von keyATM unter Einbeziehung dieser Termgewichtung ist für die Themenextraktion aus forensischen Kommunikationsdaten insofern vielversprechend, da diese zusätzlich zu der in Abschnitt 3.2.1 beschriebenen umfangreichen Entfernung von irrelevanten Wörtern das Problem der hohen Anzahl an Noise Words (siehe Abschnitt 2.5) abmildert. keyATM basierend auf dem gewöhnlichen Collapsed Gibbs Sampling wurde ebenfalls angewendet, um eine bessere Vergleichbarkeit mit der nachfolgend erklärten Seeded LDA [15] zu ermöglichen, da diese ebenfalls Collapsed Gibbs Sampling zur Parameterschätzung einsetzt.

Für beide Varianten wurde der Hyperparameter $\tilde{\beta}$ als Startwert für die a-priori-Verteilung über die spezielle Themen-Wort-Verteilung $\tilde{\phi}$ auf den von Eshima u. a. [14] empfohlenen Standardwert von $\tilde{\beta} = 0,1$ gesetzt. Für die Verteilung $Beta(\gamma_1, \gamma_2)$ als a-priori-Wahrscheinlichkeitsverteilung für die Indikatorvariable wurde eine stetige Gleichverteilung gewählt, wozu $\gamma_1 = \gamma_2 = 1$ festgelegt wurde, da diese bei Eshima u. a. [14] gute Ergebnisse erzielte.

3.3.3.2 Seeded LDA

Die Seeded LDA, die ursprünglich von Lu u. a. [232] vorgeschlagen wurde und von Watanabe und Baturo [15] verbessert wurde, adaptiert im Gegensatz zu keyATM nicht den generativen Prozess der LDA. Stattdessen besteht ihr einziger Unterschied zur Standard-LDA darin, dass diese üblicherweise eine symmetrische Dirichlet-Verteilung $Dir(\vec{\beta})$ für die Themen-Wort-Verteilung einsetzt [27], während für die Seeded LDA eine asymmetrische Dirichlet-Verteilung $Dir(\vec{\beta} + \omega_{kv})$ verwendet wurde [15, 232]. Die einzelnen β_i des Vektors $\vec{\beta}$ können als Pseudoanzahl eines Wortes v_i in einem Thema k gemäß des Priors interpretiert werden [5]. ω_{kv} kann somit als zusätzliche Pseudofrequenz aufgefasst werden, die für ein Seed Wort des entsprechenden „Seeded Topics“ vor der Anpassung des Modells hinzugefügt wird, wodurch angegeben wird, dass es eine Präferenz für dieses Wort in der Wortverteilung des Themas gibt [15, 232]. Für ein Wort v des Vokabulars und ein Thema k wird die zusätzliche Pseudofrequenz ω_{kv} durch Gleichung (3.9) angegeben, wobei S_k die Menge von Seed Wörtern des entsprechenden Themas k bezeichnet und f_v für die Häufigkeit von v im Datensatz steht [15].

$$\omega_{kv} = \begin{cases} 100\mu\frac{f_v}{\sum f} & v \in S_k \\ 0 & v \notin S_k \end{cases} \tag{3.9}$$

Wie in Abschnitt 3.3 bereits kurz erwähnt wurde, erlaubt die Seeded LDA ebenfalls das Finden von „Unseeded Topics“ [15], bei denen somit für alle Wörter v $\omega_{kv} = 0$ gilt. Für die Seed Wörter $v \in S_k$ wird ω_{kv}, wie in Gleichung (3.9) beschrieben wird, in Abhängigkeit von ihrer Frequenz im Datensatz bestimmt, wobei ω_{kv} für seltenere Seed Wörter niedriger ist [15]. Durch den Hyperparameter $\mu > 0$ wird gesteuert, welche Bedeutung grundsätzlich den Seed Wörtern zugemessen werden soll [15]. Da einige der Seed Wörter wie beispielsweise „Nachforschungsauftrag“, „ablästern“ und „Belästigung“ nur einmal im Datensatz vorkamen, wurde ein hoher Wert von $\mu = 0,05$ gewählt, um zu gewährleisten, dass diese trotz ihrer geringen Frequenz einen Einfluss auf die Themen haben. Zudem wurde von Watanabe und Baturo [15] betont, dass sich der Algorithmus robust gegenüber unterschiedlichen Werten von μ verhält und hohe Werte seine Leistung nicht beeinträchtigen.

3.4 Erweiterung basierend auf CluWords

Wie bereits in Abschnitt 3.3.3 kurz angesprochen wurde, bezwecken die beiden verwendeten halbüberwachten Algorithmen zwar, dass mit den Seed Wörtern in Verbindung stehende Begriffe in dem entsprechenden „Seeded Topic“ mit einer hohen Wahrscheinlichkeit auftreten [14, 15]. Sie setzen jedoch hierfür voraus, dass diese Begriffe mit den Seed Wörtern gemeinsam in Dokumenten vorkommen [14, 15]. Um zu garantieren, dass die wahrscheinlichsten Wörter der „Seeded Topics“ Begriffe sind, die möglichst stark mit den Seed Wörtern zusammenhängen und relevant für das gewünschte Thema sind, kann es förderlich sein, weitere Informationen über die Ähnlichkeit zwischen den Begriffen des Datensatzes und Seed Wörtern neben den Kookkurrenzen auf Dokumentenebene einzubeziehen. Gerade für Texte

geringer Länge kann dies von hoher Bedeutung sein, da, wie bereits in Abschnitt 2.4 erläutert wurde, ein Problem darin zu sehen ist, dass Wörter innerhalb der kurzen Texte nur selten gemeinsam vorkommen [97]. Dieses Problem war trotz der Aggregation der Nachrichten zu Pseudodokumenten noch vorhanden, da diese dennoch durchschnittlich nur 13 Wörter umfassten (siehe Abschnitt 3.1). Eine Idee bestand deshalb darin, mithilfe einer Adaption des von Viegas u. a. [16] beschriebenen Ansatzes CluWords zu bezwecken, dass zu den Seed Wörtern ähnliche Begriffe unabhängig von ihrer Kookkurrenzhäufigkeit eine hohe Wahrscheinlichkeit in dem entsprechenden Seed Topic aufweisen.

Unter einem CluWord wird ein Cluster von semantisch ähnlichen Wörtern verstanden [16]. Genauer gesagt definiert Viegas u. a. [16] das CluWord $C(w_i)$ eines Terms w_i als die Menge aller Wörter des Vokabulars $w \in V$, die mindestens eine semantische Ähnlichkeit $sim(w_i, w)$ von α zu w_i aufweisen, wobei die Autoren die semantische Ähnlichkeit von Wörtern als die Ähnlichkeit ihrer Word Embeddings festlegten. Formal kann ein CluWord in Anlehnung an die Definiton von Viegas u. a. [16] durch Gleichung (3.10) beschrieben werden.

$$C(w_i) = \{w | sim(w_i, w) \geq \alpha\} \tag{3.10}$$

Die ursprüngliche Idee des Ansatzes der Autoren bestand darin, in jedem Dokument die Wörter durch ihr CluWord zu ersetzen und anschließend die neu entstandenen Pseudodokumente als Eingabe für den Algorithmus NMF zur Themenmodellierung [32, 34] zu verwenden [16]. Damit verfolgten Viegas u. a. [16] die in Abschnitt 2.4 beschriebene Strategie, die Resultate der Themenmodellierung für kurze Texte durch die Einbeziehung der Word Embedding Ähnlichkeit zu verbessern. Da sie dies durch die Veränderung der Dokumentenrepräsentation anstatt durch die Adaption des Algorithmus erreichten, bietet sich der Ansatz an, um ihn mit anderen Algorithmen als NMF wie den beiden Algorithmen des Seed-Guided Topic Modellings keyATM [14] und Seeded LDA [15] zu kombinieren.

Die prinzipielle Vorgehensweise zur Einbeziehung von CluWords in das Seed-Guided Topic Modelling ist in Abbildung 3.5 skizziert. Der wesentliche Unterschied zu dem ursprünglichen Ansatz von Viegas u. a. [16] bestand darin, dass nicht zu allen Wörtern des Vokabulars des Datensatzes die CluWords bestimmt wurden, sondern nur zu ausgewählten Seed Wörtern. Konkret wurden zu den Topic Labeln jeder Seed Wortmenge ähnliche Wörter des Datensatzes ermittelt. Die Entscheidung nur die Wörter mit einer hohen Ähnlichkeit zu den Topic Labeln anstatt zu allen Seed Wörtern einzubeziehen beruhte darauf, dass die tatsächliche Relevanz von einigen Seed Wörtern für den Fall unklar war, da diese, wie in Abschnitt 3.3.1 beschrieben wurde, teilweise basierend auf automatisierten Verfahren extrahiert wurden. Die Topic Label wurden hingegen als besonders repräsentativ für die Seed Wortmengen betrachtet. In jedem Konversationsdokument wurde ein Topic Label durch sein CluWord ersetzt. Es muss an dieser Stelle darauf hingewiesen werden, dass ein CluWord C_{w_i} aufgrund der Definition der semantischen Ähnlichkeit stets auch das Wort w_i selbst beinhaltete und zudem nur Wörter w aus dem Trainingsdatensatz des Themenmodells enthielt.

Die auf diese Weise gebildeten Pseudodokumente dienten als Eingabe für das Training von keyATM und Seeded LDA. Hinsichtlich keyATM wurde ausschließlich die Version basierend auf dem gewöhnlichem Collapsed Gibbs Sampling herangezogen. Die Version von keyATM unter Einbeziehung des Termgewichtungsschemas basierend auf der PMI wurde nicht in Kombination mit den CluWords verwendet, da diese Wörter abgewichtet, die in vielen Dokumenten auftreten [14, 189]. Hierdurch bestand die Gefahr, dass gerade die Begriffe der CluWords ein niedrigeres Gewicht erhalten und damit ihre Auswirkung abgeschwächt wird, da diese in mehrere Konversationsdokumente künstlich eingefügt wurden.

Die CluWords wurden zunächst wie bei Viegas u. a. [16] basierend auf der Ähnlichkeit von vortrainierten Word Embeddings ermittelt. Darüber hinaus wurden diese Ergebnisse mit einer Variante verglichen, die anstatt Word Embeddings CluWords mithilfe von paradigmatischen Relationen bildet. Im Folgenden werden diese beiden Methoden zur Erstellung der CluWords detaillierter beschrieben. In Abbildung 3.5 ist zudem die grundlegende Vorgehensweise zum Anreichern der Konversationsdokumente mit den CluWords der Topic Labels dargestellt.

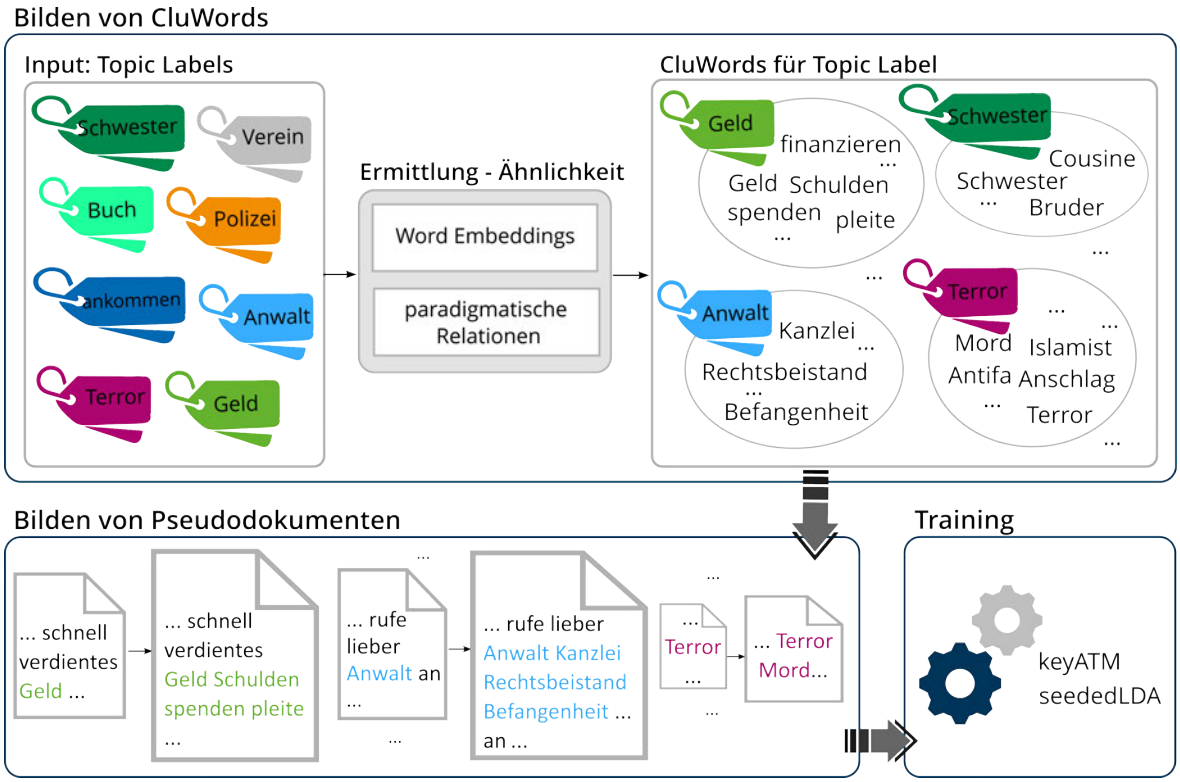

Abbildung 3.5: Vorgehensweise zur Anreicherung von Dokumenten mit den CluWords der Topic Labels. Für jedes Topic Label setzte sich das CluWord aus den Begriffen zusammen, deren semantische Ähnlichkeit, ermittelt mithilfe von Word Embeddings oder paradigmatischen Relationen, zu dem Topic Label einen Schwellenwert überstieg. Die Topic Labels wurden anschließend in jedem Konversationsdokument durch ihr CluWord ersetzt. Die gebildeten Pseudodokumenten dienten als Eingabe für das Seed-Guided Topic Modelling.

3.4.1 Bestimmung der CluWords unter Verwendung von fastText

Zunächst wurden, wie bei Viegas u. a. [16], auf einem externen Datensatz trainierte Word Embeddings zur Berechnung der Wortähnlichkeit verwendet. Als Verfahren zum Lernen von Word Embeddings wurde ein unüberwachtes fastText Modell [179] gewählt, da fastText ebenfalls Vektorrepräsentationen von sogenannten Out-of-Vocabulary (OOV) Wörtern bilden kann, die nicht im Trainingsdatensatz erhalten waren. Das fastText Modell wurde mit dem Skip-Gram-Verfahren trainiert, wobei die gewählten Werte für die Hyperparameter Tabelle 3.4 entnommen werden können. Wie von Lai u. a. [59] betont wird, sollte die Datengrundlage, auf der die Word Embeddings trainiert werden, dem Zieldatensatz, d.h. hier dem Trainingsdatensatz für die Themenmodelle, möglichst ähnlich sein beziehungsweise aus dem gleichen Bereich stammen. Daher wurde das fastText Modell auf einer großen Menge deutscher Tweets trainiert, die zumindest der sprachlichen Qualität der Falldaten ähnelten und ebenfalls umgangssprachlich waren. Der hierfür verwendete Trainingsdatensatz umfasste insgesamt 24 Millionen Tweets zu verschiedensten Themen, wobei neben den von Kratzke [340] bereitgestellten 20 Millionen Tweets weitere vier Millionen Tweets mit dem Twitter Intelligence Tool (TWINT) [341] im Februar und März 2022 gecrawlt wurden.

Tabelle 3.4: Verwendete Werte für die Hyperparameter für das Training eines unüberwachten fastText Modells. Das trainierte fastText Modell wurde für die Ermittlung der ähnlichsten Begriffe zu den Topic Labels der Seed Wortmengen eingesetzt.

Hyperparameter	**Wert**
Anzahl der Dimensionen	300
Anzahl der Epochen	50
Fenstergröße	5
Minimale Länge der Character N-Gramme	2
Maximale Länge der Character N-Gramme	6
Lernrate	0,05

Für jedes Topic Label wurden die ähnlichsten Wörter aus dem Datensatz ermittelt, wozu die Kosinusähnlichkeit zwischen der Word Embedding Repräsentation des Topic Labels und jedes Begriffs des Datensatzes berechnet wurde. Um zu bestimmen, welche Wörter zu dem CluWord des Topic Labels gehörten, war es erforderlich den Schwellenwert α festzulegen. Von Viegas u. a. [16] wurde empfohlen, einen hohen Schwellenwert α zu wählen, um verrauschte Wortpaare herauszufiltern. Dementsprechend wurde ein Schwellenwert von $\alpha = 0.45$ gewählt. Dies resultierte in zwölf bis 83 Begriffen je CluWord für die Topic Labels. In Tabelle 3.5 ist die Anzahl der Wörter des CluWord für jedes Topic Label sowie Wörter dieses CluWord aufgeführt. Aus Lesbarkeitsgründen wurden exemplarisch je CluWord fünf repräsentative Wörter ausgewählt.

Tabelle 3.5: Übersicht über die gebildeten CluWords zu den Topic Labels der Seed-Wortmengen basierend auf fastText Embedding Ähnlichkeit. Aufgeführt sind das Topic Label, zu dem das CluWord gebildet wurde und die Anzahl der Wörter, die dieses CluWord enthielt. Ebenfalls sind exemplarisch fünf repräsentative Begriffe jedes CluWord dargestellt.

Topic Label	**#Wörter**	**Beispielhafte Wörter des CluWords**
Geld	83	finanzieren, erwirtschaften, überweisen, Schulden, spenden, pleite
Schwester	75	Mutter, Bruder, Cousine, Grundschulfreundin, schwanger
Buch	43	Rezension, Verlag, Autoren, Leseratte, verfilmen
Polizei	40	Blaulicht, kripo, Durchsuchungsbefehl, Justiz, Sondereinsatz
Anwalt	36	Kanzlei, Rechtsbeistand, Gerichtstermin, Befangenheit, Staatsanwaltschaft
ankommen	19	verschicken, Paket, Zustellbasis, Ankunft, eintrudeln
Terror	16	Islamist, Anschlag, Antifa, Mord, Moscheen
Verein	12	Zusammengehörigkeit, Gemeinsamkeiten, gründen, gemeinnützig, miteinander

3.4.2 Bestimmung der CluWords basierend auf paradigmatischen Relationen

Wie in Absatz 2.4 erläutert wurde, kann es sich als problematisch erweisen, dass die Informationen über die semantische Ähnlichkeit, die auf einem externen Trainingsdatensatz gewonnen wurde, nicht zwangsläufig für die forensischen Kommunikationsdaten relevant sind. Eine alternative Möglichkeit wäre darin zu sehen, die Word Embeddings direkt auf dem forensischen Kommunikationsdatensatz zu trainieren. Jedoch wurde dieser nicht als umfassend genug betrachtet, um zuverlässige Word Embeddings erlernen zu können. Daher wurde, wie von Liu u. a. [120] vorgeschlagen wurde, stattdessen die Ähnlichkeit basierend auf paradigmatischen Relationen [342] auf dem Datensatz berechnet und bei der Themenmodellierung berücksichtigt.

Grundsätzlich besteht zwischen zwei Wörtern eine paradigmatische Relation, wenn diese in ähnlichen Kontexten auftreten [343], wobei von Miller und Charles [344] davon ausgegangen wurde, dass diese über eine ähnliche Bedeutung verfügen. Formal stehen zwei Wörter in paradigmatischer Relation $PARA(w_i, w_j)$, wenn die Ähnlichkeit ihrer globalen Kontexte $K_G(w_i)$ und $K_G(w_j)$ bezüglich eines bestimmten Ähnlichkeitsmaßes SIM einen Schwellenwert t übersteigt [345]. Diese Beziehung kann mit Gleichung (3.11) beschrieben werden [345].

$$PARA(w_i, w_j) \leftrightarrow SIM_t(K_G(w_i), K_G(w_j)) \tag{3.11}$$

In dieser Arbeit wird für den globalen Kontext die Definition von Biemann und Riedl [346] und Bordag [347] verwendet, nach der der globale Kontext $K_G(w_i)$ die N signifikantesten Kookkurrenzen von w_i umfasst, da dieser Ansatz als rechnerisch effizient gilt. Unter den signifikanten Kookkurrenzen eines Wortes w_i werden Wörter verstanden, die statistisch auffällig häufig gemeinsam mit w_i in einem festgelegten Textfenster auftreten [343]. Als Textfenster könnte beispielsweise eine Nachricht definiert werden. Jedoch verfügten diese nach der Vorverarbeitung durchschnittlich über weniger als drei Wörter. Die Länge der Konversationsdokumente variierte jedoch sehr stark und umfasste bei 15 Dokumenten mehr als 250 Wörter. Daher wurde als Kompromiss die signifikanten Kookkurrenzen basierend auf dem gemeinsamen Vorkommen von Wörtern in einem gleitenden Wortfenster innerhalb eines Konversationsdokuments definiert, das jeweils sechs aufeinanderfolgende Wörter umfasste. Das gemeinsame Auftreten der Wörter in diesen Wortfenstern wurde mit dem von Dunning [348] definierten Log-Likelihood-Maß als Signifikanzmaß beurteilt. Analog zu dem Vorgehen von Bordag [347] bildeten für jedes Wort w_i im Datensatz die N Wörter mit dem höchsten Log-Likelihood-Wert seinen globalen Kontext $K_G(w_i)$, wobei N auf 50 festgelegt wurde.

Diese globalen Kontexte wurden, wie beispielsweise von Bordag [347] und Otero [349] empfohlen wurde, als binäre Vektoren dargestellt, deren Elemente angaben, ob ein Wort in dem globalen Kontext von w_i enthalten war. Anschließend wurden diese Vektoren mit dem binären Kosinus als Ähnlichkeitsmaß verglichen, da dieses Maß nach Lapesa u. a. [350] und Gamallo und Bordag [351] zu besonders bedeutungsvollen paradigmatischen Relationen führt.

Zu jedem der acht Topic Labels wurden die Wörter, die zu diesem in starker paradigmatischer Relation standen beziehungsweise eine hohe Kontextähnlichkeit aufwiesen zur Bildung der CluWord bestimmt. Hierfür war es erneut erforderlich einen Schwellenwert festzulegen. Dafür wurden die Ähnlichkeitsverteilungen der Wörter zu jedem Topic Label mithilfe von Histogramm-Analysen untersucht. Hierbei konnte festgestellt werden, dass die Topic Label zu den meisten Wörtern eine Kontextähnlichkeit von null aufwiesen. Um erneut gewährleisten zu können, dass nur die Wörter mit einer starken Kontextähnlichkeit dem CluWord zugeordnet wurden, wurden nur die 0,25 % ähnlichsten Wörter zu jedem Topic Label berücksichtigt. Jedes entstandene CluWord bestand aus elf bis 19 Wörtern. In Tabelle 3.6 sind fünf repräsentative CluWords jedes Topic Labels mit der Anzahl der diesem CluWord zugeordneten Wörter aufgeführt.

Tabelle 3.6: Übersicht über die gebildeten CluWords zu den Topic Labels der Seed-Wortmengen basierend auf paradigmatischen Relationen. Das Topic Label, zu dem das CluWord erstellt wurde und die Anzahl der Wörter, aus denen dieses CluWord bestand, sind dargestellt. Ebenfalls sind exemplarisch fünf repräsentative Wörter jedes CluWords aufgeführt.

Topic Label	# Wörter	Beispielhafte Wörter des cluWords
Geld	15	boahhh, Rückfuhr, Justizkasse, Euro, Schuldfrage
Schwester	13	Bruder, beisammen, Frauenraum, Schöpfer, Familienmitglied
Buch	16	geisteswissenschaftlich, Ausgabe, Tolkien, Fantasy, Mythos
Polizei	14	Blaulicht, schlimmstenfalls, Verwarnung, Notruf, durxh
Anwalt	19	dokumnetieren, Fachgebiet, Vereinsgründer, Mehrheitsbeschluss, belangen
ankommen	11	Briefes, hättendann, verantwortlichen, Watte
Terror	22	Emotionsachterbahn, Unicef, Anklagepunkt, beglaubigen, Chronic
Verein	11	rechtsfähig, Rechtspersönlichkeit, Vereinsregister, eingetragen, Haftung

3.5 Evaluierung

Der folgende Abschnitt beschreibt das Vorgehen bezüglich der Evaluierung der Themenmodelle.

3.5.1 Qualitative und automatische, quantitative Evaluierung

Zunächst erfolgte eine qualitative Evaluierung der mit den verschiedenen Algorithmen extrahierten Themen. Diese beruhte, wie beispielsweise in [51, 73, 139], auf der Analyse und Interpretation der wahrscheinlichsten Wörter ausgewählter „Seeded Topics“. Darüber hinaus wurden die Themenmodelle ebenfalls quantitativ evaluiert. Als Evaluierungsmaß diente die in Abschnitt 3.3.2.1 beschriebene semantische Kohärenz [19]. Um die verschiedenen halbüberwachten Ansätze mit der unüberwachten LDA als Baseline-System vergleichen zu können, wurde die durchschnittliche semantische Kohärenz über alle erstellten Themen berechnet. Zudem wurde für alle Themenmodelle basierend auf den beiden untersuchten Verfahren des Seed-Guided Topic Modellings die durchschnittliche Kohärenz über die „Seeded Topics“ ermittelt, um beurteilen zu können, welcher der beiden unterschiedlichen Strategien zum Integrieren des Vorwissens in die LDA besser geeignet ist.

3.5.2 Durchführung und Auswertung der Nutzerstudie

Wie von Doogan und Buntine [17] gezeigt wurde, hängt die Übereinstimmung von automatischen Kohärenzmaßen mit der menschlichen Interpretierbarkeit von Themen stark von dem zugrundeliegenden Datensatz ab (siehe auch Abschnitt 2.11). Daher wurde die quantitativ ermittelte Qualität der Themen mit den Ergebnissen einer Nutzerstudie verglichen.

Als Nutzerstudie wurde der von Chang u. a. [315] vorgestellte Word Intrusion Test gewählt, der in bisherigen Arbeiten ebenfalls zur Evaluierung von Ansätzen des Seed-Guided Topic Modellings [12, 233] sowie zur Beurteilung von quantitativen Evaluierungsmaßen eingesetzt wurde [z.B. 18]. Der Word Intrusion Test wurde gegenüber dem ebenfalls von Chang u. a. [315] vorgeschlagenen und in Abschnitt 2.11 beschriebenen Topic Intrusion Test bevorzugt, da dieser nach Eshima u. a. [14] und Ying u. a. [116] als zu schwierig für die Annotatoren empfunden wurde, weshalb die Ergebnisse wenig informativ waren. Darüber hinaus ist dieser Test, wie von Eshima u. a. [14] und Ying u. a. [116] gezeigt wurde, stärker als beispielsweise der Word Intrusion Test von den ausgewählten, präsentierten Wörtern abhängig.

Die grundlegende Aufgabe der Annotatoren im Rahmen des Word Intrusion Tests besteht darin, einen Eindringling unter den präsentierten Begriffen zu identifizieren [315]. Die Umsetzung des Word Intrusion Tests entsprach der von Chang u. a. [315] vorgeschlagenen Vorgehensweise. Hierbei wurde den Annotatoren für jedes zu bewertende Thema sechs Wörter angezeigt. Bei den fünf Begriffen, die tatsächlich zu dem Thema gehörten, handelte es sich um die fünf Wörter mit der höchsten Wahrscheinlichkeit in diesem Thema [315, 352]. Als Eindringling wurde, wie von Chang u. a. [315] empfohlen wurde, ein Wort bestimmt, das eine niedrige Wahrscheinlichkeit in dem betrachteten Thema aufwies, sich jedoch unter den wahrscheinlichsten fünf Wörtern eines anderen Themas befand. Hierdurch soll nach Chang u. a. [315] sichergestellt werden, dass der Eindringling nicht dadurch auffällt, dass es sich um ein besonders ungewöhnliches beziehungsweise seltenes Wort handelt.

Hinsichtlich der extrahierten Themen aller Ansätze des Seed-Guided Topic Modellings - keyATM, Seeded LDA und die Kombinationen dieser Ansätze mit der Adaption des CluWord-Ansatzes basierend auf Word Embedding Ähnlichkeit und paradigmatischen Relationen - wurden nur die acht „Seeded Topics“ bewertet. Zum einen sollte auf diese Weise der Aufwand für die Annotatoren möglichst gering gehalten werden. Zum anderen entsprach bei beiden Algorithmen für die Extraktion von unüberwachten Themen der Standard-LDA, weshalb die unterschiedlichen Verfahren zur Integration von Seed Wörtern vor allem Auswirkungen auf die „Seeded Topics“ haben [12]. Jedes Thema wurde durch drei Annotatoren bewertet. Diesen wurde neben den Arbeitsanweisungen ebenfalls eine Tabelle mit kurzen Erklärungen zu teilweise ungewöhnlichen Begriffen der Word Intrusion Aufgaben gegeben. Darüber hinaus wurden die Teilnehmer, wie von Piccardi und West [290] vorgeschlagen wurde, aufgefordert, nach Begriffen, deren Bedeutung sie nicht kannten, im Internet zu recherchieren oder diese in einem Wörterbuch nachzuschlagen. Ein Beispiel für eine Aufgabe des Word Intrusion Tests, mit dem ein ausgewähltes Thema bewertet werden sollte, ist in Abbildung 3.6 skizziert.

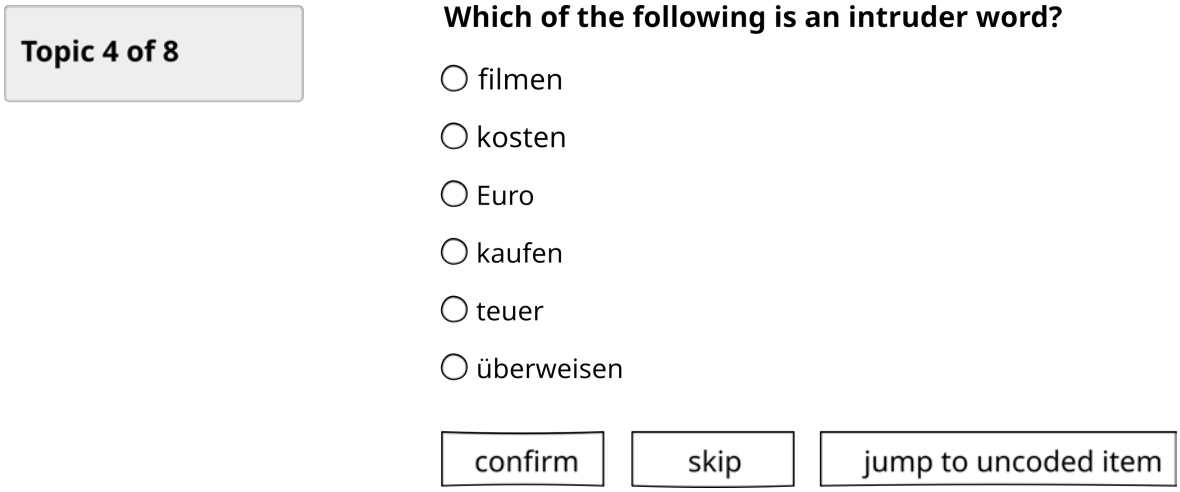

Abbildung 3.6: Beispiel für eine Aufgabe des Word Intrusion Tests zur Evaluierung der trainierten Themenmodelle. Die Teilnehmer wurden aufgefordert, den Eindringling unter den präsentieren Wörtern auszuwählen.

Der Erfolg der Annotatoren bei der Auswahl des richtigen Eindringlings zeigte, inwieweit die wahrscheinlichsten Wörter eines Themas tatsächlich miteinander verbunden waren und wie kohärent das Thema wahrgenommen wurde [315]. Als quantitatives Maß zur Beurteilung der Kohärenz eines ausgewählten Themas wurde, wie von Harandizadeh u. a. [233] vorgeschlagen wurde, der Prozentsatz der Annotatoren, die den Eindringling des Themas korrekt identifizieren konnten, verwendet.

Für ein ausgewähltes Themenmodell m wurde die Kohärenz seiner Themen mit der von Chang u. a. [315] und Chan und Sältzer [352] beschriebenen Modell Precision (MP) berechnet, die für einen einzelnen Annotator als der Prozentanteil der bewerteten Themen K definiert ist, bei denen es ihm gelang, den korrekten Eindringling zu bestimmen. Formal ist die MP für einen bestimmten Annotator wie in Gleichung (3.12) definiert, wobei $i^m_{k,s}$ der Index des von dem Annotator ausgewählten Wortes und ω^m_k den Index des tatsächlichen Eindringlings kennzeichnet.

$$MP^m_k = \mathbb{1} \sum_{k=1}^{K} (i^m_{k,s} = \omega^m_k)/K \tag{3.12}$$

Für die drei Annotatoren wurde der arithmetische Mittelwert der einzelnen Werte für die MP bestimmt.

Darüber hinaus wurde mit einem Signifikanztest beurteilt, ob die Ergebnisse der Annotatoren besser waren, als wenn sie den Eindringling zufällig geraten hätten. Hierfür wurde die von Chan und Sältzer [352] vorgeschlagene Vorgehensweise herangezogen. Zur Beurteilung der Ergebnisse eines einzelnen Annotators bei dem Word Intrusion Test für ein zu bewertendes Themenmodell wurde ein exakter Binomialtest durchgeführt, wobei das Signfi-

ikanzniveau auf $\alpha = 0.05$ gesetzt wurde. Die p-Werte von allen Annotatoren wurden mit der Fisher-Methode [353] kombiniert [352]. Die Nullhypothese H_0 und die Alternativhypothese H_1 können Gleichung (3.13) entnommen werden [352]:

$$\begin{aligned} H_0 &: MP \leq 1/(n+1) \\ H_1 &: MP > 1/(n+1) \end{aligned} \tag{3.13}$$

Hierbei steht n für die Anzahl der Wörter, die tatsächlich zu dem Thema gehören und zu denen der Eindringling hinzugefügt wurde [352]. In diesem Fall galt $n = 5$.

Hinsichtlich der Bedeutung des Signifikanztests wurde von Chan und Sältzer [352] hervorgehoben, dass die Ablehnung der Nullhypothese nicht darauf hindeutet, dass das Themenmodell über eine hohe Qualität und Interpretierbarkeit verfügt. Stattdessen stellt die Tatsache, dass der Word Intrusion Test für ein Themenmodell besser als eine zufällige Schätzung war, ein Minimum dar, das ein Themenmodell erfüllen sollte, um für menschliche Bewerter nützlich sein zu können. Für den Vergleich der Themenmodelle ist hingegen nach Chan und Sältzer [352] die MP ausschlaggebend.

3.5.3 Vergleich der automatischen Evaluierung und der Nutzerstudie

Sowohl die semantische Kohärenz [19] als auch der Word Intrusion Test [315] bewerten die Kohärenz der Themen eines Modells. Eine weitere untersuchte Fragestellung bestand darin, inwiefern die automatisch ermittelte Kohärenz mit den Erkenntnissen aus der Nutzerstudie übereinstimmte. Hierfür wurde der Kendall-Tau-Korrelationskoeffizient [354] zwischen den automatisch ermittelten Kohärenzwerten und den Ergebnissen des Word Intrusion Tests berechnet. Genauer gesagt wurde die Variante Kendall-Tau-b gewählt [355]. Zudem wurde die Signifikanz der Korrelation beurteilt, wofür die folgenden Hypothesen aufgestellt wurden:

- H_0: Es besteht keine Korrelation zwischen den Ergebnissen der automatischen Evaluierung und der Nutzerstudie.
- H_1: Es besteht ein korrelativer Zusammenhang (positive oder negative Korrelation) zwischen den Resultaten der automatischen Evaluierung und der Nutzerstudie.

Das Signifikanzniveau wurde auf $\alpha = 0.05$ gesetzt.

Die Beziehung zwischen der automatischen und der menschlichen Bewertung wurde analog zu der Vorgehensweise von Lau u. a. [318] sowohl auf der Ebene von Themenmodellen als auch von einzelnen Themen eines Modells untersucht. Bezüglich der Evaluierung auf Modellebene wurde der Korrelationskoeffizient zwischen der durchschnittlichen semantischen Kohärenz und der MP der Themenmodelle berechnet [318]. Der Korrelationskoeffizient wurde somit über alle untersuchten Algorithmen und Einstellungen für die Gesamtthemenanzahl K ermittelt. Auf diese Weise sollte beurteilt werden, inwiefern die Rangfolge der Themenmodelle nach der semantischen Kohärenz mit ihrer Rangfolge, die sich aus der Nutzerstudie

ergab, im Einklang stand [19]. Hinsichtlich der Evaluierung auf Themenebene wurde die semantische Kohärenz und der Prozentanteil der Annotatoren, die den richtigen Eindringling des Themas auswählten, verglichen. Es muss hierbei betont werden, dass, wie bei Hoyle u. a. [18] keine Unterscheidung getroffen wurde, durch welches Modell die Themen erstellt wurden, sondern ein einziger Korrelationskoeffizient für alle Themen gebildet wurde.

4 Ergebnisse und Diskussion

Insgesamt wurden sieben verschiedene Themenmodelle des Seed-Guided Topic Modellings untersucht, die sich in den folgenden Aspekten voneinander unterschieden: dem verwendeten Algorithmus, der Einbeziehung von CluWords und gegebenenfalls der Methode, die zur Bildung der CluWords angewendet wurde. Im Falle von keyATM wurde zudem unterschieden, ob das gewöhnliche Collapsed Gibbs Sampling oder die in Abschnitt 3.3.3.1 beschriebene Termgewichtung gewählt wurde. Darüber hinaus wurden die Modelle mit unterschiedlichen Anzahlen an „Unseeded Topics“ K' trainiert, die nach dem in Abschnitt 3.3.2 beschriebenen Vorgehen alle als geeignet betrachtet wurden. Die Abkürzungen, die für die Modelle verwendet wurden, sind in Tabelle 4.1 aufgeführt. Zusätzlich zu der Zusammensetzung der Abkürzungen für die verschiedenen Methoden wird jeweils angegeben, mit welcher Anzahl an „Unseeded Topics“ K' das Modell trainiert wurde. Beispielsweise steht „KaCF-7“ für ein Themenmodell, das mit dem Algorithmus keyATM auf Pseudodokumenten trainiert wurde, bei denen die Topic Labels durch CluWords basierend auf fastText-Embeddings ersetzt wurden, und das sieben „Unseeded Topics“ umfasst.

Tabelle 4.1: Erläuterung der Abkürzungen für die Bezeichnungen der einzelnen Themenmodelle des Seed-Guided Topic Modellings. Die Bezeichnung der Themenmodelle besteht aus der Abkürzung des verwendeten Algorithmus (Ka oder S), der Information, ob CluWords einbezogen wurden (C oder nichts), mit welcher Methode diese gebildet wurden (F oder P) und ob für keyATM die Adaption von Collapsed Gibbs Sampling zur Einbeziehung eines Termgewichtungsschemas (T oder nichts) verwendet wurde.

Methode	Wert	Abkürzung
Algorithmus	keyATM/ Seeded LDA	Ka/ S
Einbeziehung von CluWords	ja/ nein	C
ggf. Methode für CluWords	fastText/ paradigmatische Relationen	F/ P
Einsatz der Termgewichtung bei keyATM	ja/ nein	T

Die Ergebnisse der untersuchten Ansätze zur Themenmodellierung werden in dem folgenden Kapitel dargestellt. Hierbei wird zunächst auf die ermittelten Themenanzahlen eingegangen, die mithilfe des in Abschnitt 3.3.2 beschriebenen Verfahrens als geeignet anzusehen waren und mit denen alle Experimente durchgeführt wurden. Anschließend werden die Ergebnisse der qualitativen und automatischen, quantitativen Evaluierung sowie der Nutzerstudie beleuchtet und es wird diskutiert, inwiefern die Resultate der automatisch ermittelten Kohärenz mit der Nutzerstudie übereinstimmten.

4.1 Ermittelte optimale Themenanzahl

In Abbildung 4.1 sind die durchschnittliche semantische Kohärenz und die FREX für die Themenmodelle abgebildet, die, wie in Abschnitt 3.3.2 beschrieben, mit der Standard-LDA mit einer Gesamtanzahl von $K \in [10, 20]$ Themen trainiert wurden. Wie aus der Abbildung hervor-

geht, erreichte kein Themenmodell sowohl die höchste semantische Kohärenz als auch die höchste FREX. Stattdessen verfügte das Themenmodell mit $K = 15$ Themen über die höchste durchschnittliche semantische Kohärenz der untersuchten Modelle, jedoch auf Kosten der FREX. Hohe Werte für die FREX wurden vor allem durch eine höhere Anzahl an Themen erreicht, was mit den Erkenntnissen bisheriger Arbeiten [z.B. 13, 70, 332] übereinstimmte. Konkret wies das Themenmodell mit $K = 19$ Themen die höchste FREX auf, gefolgt von $K = 20, 18, 17$. Die Themenmodelle bestehend aus elf bis 14 Themen verfügten dagegen über geringere Werte für die FREX, jedoch über eine höhere semantische Kohärenz als beispielsweise die Themenmodelle mit 18 oder 19 Themen, was den Trade-off zwischen semantischer Kohärenz und FREX verdeutlicht.

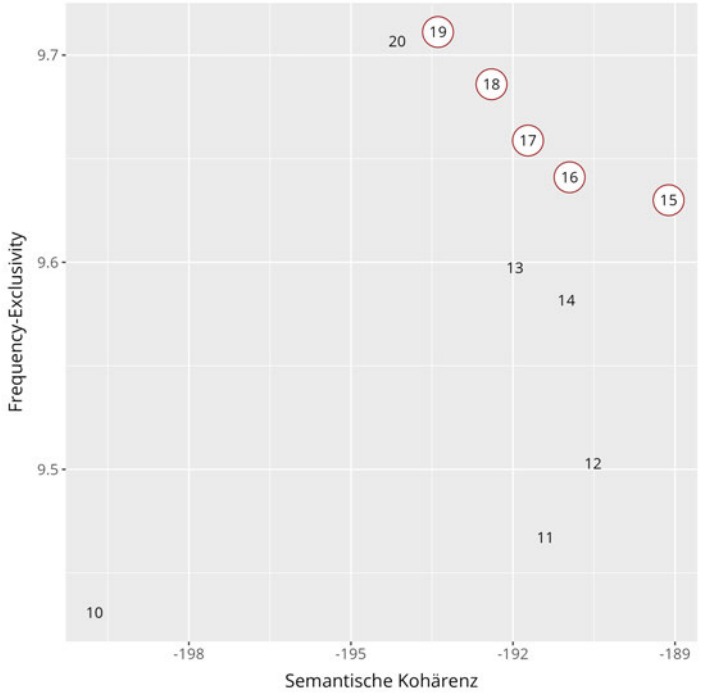

Abbildung 4.1: Semantische Kohärenz und FREX der Themenmodelle unter Verwendung der Standard-LDA mit $K = [10, 20]$ Themen. Die dargestellten Zahlen stehen für die Gesamtanzahl K der Themen, mit denen das Modell trainiert wurde. Die Anzahl der Themen in der pareto-optimalen Menge an Themenmodellen sind rot umkreist.

Das Themenmodell mit $K = 10$ Themen fiel dadurch auf, dass es sowohl hinsichtlich der semantischen Kohärenz als auch der FREX von allen anderen untersuchten Themenmodellen übertroffen wurde. Dies überraschte insofern, dass $K = 10$ die geringste untersuchte Themenanzahl darstellte und beispielsweise nach Kim u. a. [13], Srivastava u. a. [70] und Weston u. a. [332] die semantische Kohärenz üblicherweise mit einer zunehmenden Anzahl an Themen sinkt. Jedoch ist anzumerken, dass weitere Analysen ergaben, dass tatsächlich bei nur zwei Themen die durchschnittliche semantische Kohärenz deutlich über dem Wert von $K = 15$ lag und nur bei $K = 10$ einen Tiefwert hatte. Jedoch wurden Werte von $K < 10$ für die Gesamtthemenanzahl von vorneherein ausgeschlossen, da, wie in Abschnitt 3.3.2 beschrieben wurde, das Ziel vor allem darin bestand, aus der ermittelten Gesamtanzahl an Themen für die Ansätze des Seed-Guided Topic Modellings eine geeignete Anzahl an zusätzlichen „Unseeded Topics“ neben den bekannten acht „Seeded Topics“ zu finden.

Die Modelle im rechten oberen Quadranten in Abbildung 4.1, die rot umkreist wurden, befanden sich in der pareto-optimalen Menge von Themenmodellen und stellten dementsprechend einen geeigneten Kompromiss zwischen semantischer Kohärenz und FREX dar. Die verschiedenen Ansätze des Seed-Guided Topic Modellings wurden daher mit $K = 15, 16, 17$ Themen beziehungsweise $K' = 7, 8, 9$ „Unseeded Topics" trainiert und evaluiert. Themenmodelle mit insgesamt 18 und 19 Themen beziehungsweise zehn und elf „Unseeded Topics" wurden nicht gelernt, obwohl diese ebenfalls zur pareto-optimalen Menge der Themenmodelle zählten, da im Rahmen der Nutzerstudie die Evaluation von fünf Modellen pro Algorithmus als zu aufwendig für die Annotatoren erachtet wurde. Die Anzahl der Themen von $K = 15, 16, 17$ wurde anstelle von 18 oder 19 Themen gewählt, da mehr Themen auch einen höheren Aufwand für die Annotatoren beziehungsweise die Nutzer bedeuten, um die einzelnen Themen zu betrachten.

4.2 Qualitative Evaluierung

In diesem Abschnitt werden die Ergebnisse der verschiedenen Ansätze zur Themenmodellierung qualitativ anhand der zehn wahrscheinlichsten Wörter der folgenden ausgewählten „Seeded Topics" präsentiert: „ankommen", „Anwalt", „Buch", „Geld" und „Verein". Themen hoher Qualität sollten möglichst gut mit dem Topic Label beziehungsweise den Seed Wörtern assoziiert werden können und zudem Informationen über den Fall der finanziellen Unterstützung einer terroristischen Vereinigung liefern. Zur besseren Vergleichbarkeit wurden alle in Abschnitt 4.2.1, Abschnitt 4.2.2 und Abschnitt 4.2.3 dargestellten Themen mit Modellen gelernt, die acht „Unseeded Topics" beziehungsweise insgesamt 16 Themen umfassten.

4.2.1 keyATM

In Tabelle 4.2 sind die Wörter mit der höchsten Wahrscheinlichkeit in den entsprechenden „Seeded Topics" abgebildet, die mit dem Modell „Ka-8", d.h. mit dem in Abschnitt 3.3.3.1 beschriebenen Algorithmus keyATM unter Verwendung von gewöhnlichem Collapsed Gibbs Sampling extrahiert wurden. Im Hinblick auf die Themen „Anwalt" und „Verein" lässt sich feststellen, dass die wahrscheinlichsten Wörter nicht das gewünschte Thema beschreiben, sondern stattdessen eher mit Themen wie „Musik" beziehungsweise „Gesundheit" in Verbindung gebracht werden können. Eine mögliche Ursache könnte darin liegen, dass die Themen tatsächlich nicht in den Nachrichten besprochen wurden und keyATM daher, wie in Abschnitt 3.3.3 beschrieben wurde, das Vorwissen des Nutzers ignoriert hat. Wie jedoch in den folgenden Unterabschnitten gezeigt wird, konnte insbesondere das Thema „Verein" durch andere Algorithmen detektiert werden. Daher ist es naheliegend, dass die Themen zwar im Datensatz vorhanden waren, aber die Auswahl der Seed Wörter dazu führte, dass diese dennoch nicht gefunden werden konnten.

Gemäß Eshima u. a. [14] beeinflussen diese die Qualität der „Seeded Topics" wesentlich. Dabei sollten Seed Wörter möglichst diskriminativ für das Thema sein und zudem eine hohe Frequenz im Datensatz aufweisen. Bezüglich des Themas „Anwalt" ist vor allem die geringe Frequenz der Seed Wörter als problematisch zu betrachten. Selbst das häufigste Seed Wort „prüfen" hatte nur eine Frequenz von 29, während Seed Wörter wie „beschuldigen" und

„Klage“ nur dreimal beziehungsweise zweimal im gesamten Datensatz vorkamen. Bezüglich des Themas „Verein“ sind die Seed Wörter vor allem als wenig spezifisch anzusehen, da zu diesen beispielsweise Begriffe wie „privat", „Screenshot“ und “besitzen“ zählen, die in vielen Kontexten verwendet werden können.

Ein weiterer Grund dafür, dass die beiden gewünschten „Seeded Topics“ „Anwalt“ und „Verein“ nicht gefunden werden konnten, könnte in einer zu geringen Anzahl an „Unseeded Topics“ liegen. Wie bereits in Abschnitt 4.1 erläutert wurde, erfassen diese vor allem häufige Themen des Datensatzes, die nicht durch die Seed Wörter beschrieben werden. Wenn zu wenige „Unseeded Topics“ extrahiert werden, könnte dies dazu führen, dass Themen mit einer starken Dominanz in dem Datensatz die selteneren „Seeded Topics“ verdrängen.

Wie in Tabelle 4.2 zu sehen ist, sind die wahrscheinlichsten Wörter in dem Thema „ankommen“ hauptsächlich mit der Ankunft an einem Urlaubsziel oder zu Hause verbunden. Sie stehen somit zwar mit einem Seed Wort beziehungsweise dem Topic Label in Verbindung, jedoch scheinen sie weder für den Fall über die finanzielle Unterstützung einer terroristischen Vereinigung relevant zu sein, noch mit den anderen Seed Wörtern des Themas wie beispielsweise „Nachforschungsauftrag“ zusammenzuhängen. Eine mögliche Ursache ist darin zu sehen, dass nach Eshima u. a. [14] bestimmte Begriffe der Seed Wortmenge eine wesentlich höhere Auswirkung auf das entsprechende Thema haben können als die anderen Wörter. Es liegt nahe anzunehmen, dass das entsprechende „Seeded Topic“ vor allem durch den allgemeinen und wenig speziellen Begriff „ankommen“ beeinflusst wurde. Dieser trat 54-mal in dem Datensatz auf, während spezifischere Begriffe wie „Nachforschungsauftrag“, der für den Fall relevant ist, nur einmal in den Nachrichten verwendet wurde.

Das Thema „Buch“ war von allgemeinen Begriffen dominiert. Dies kann auf das von Churchill u. a. [130] beschriebene Problem zurückgeführt werden, dass auch bei Ansätzen des Seed Guided Topic Modellings weiterhin hochfrequente Noise Words die Themen dominieren können. Hierbei wiesen Wörter wie „Bild", „überlegen“ und „Idee“ unter den wahrscheinlichsten Wörtern des Themas „Buch“ Frequenzen von über 400 auf.

Das Thema „Geld“ unterschied sich von den anderen „Seeded Topics“ dahingehend, dass unter den wahrscheinlichsten Wörtern tatsächlich Begriffe auftauchten, die intuitiv mit dem gewünschten Thema assoziiert werden können. Dazu gehören neben den Seed Wörter wie „Geld“, „Euro“ und „€“ auch semantisch verwandte Begriffe wie „kaufen“, „bestellen“ und „kosten“. Anders als die anderen in Tabelle 4.2 dargestellten Themen wiesen mehrere Seed Wörter im Thema „Geld“, darunter „Euro", „€“ und „schicken“, eine Frequenz von über 200 in dem Datensatz auf. Dies verdeutlicht erneut, dass die Wahl der Seed Wörter und deren Häufigkeit von entscheidender Bedeutung für die resultierenden „Seeded Topics“ sind. Allerdings muss ebenfalls darauf hingewiesen werden, dass die wahrscheinlichsten Wörter des „Seeded Topics“ „Geld“, wie „zahlen“ und „Preis“, sehr allgemeine Begriffe sind, die nicht unbedingt mit der Finanzierung einer terroristischen Vereinigung in Verbindung stehen müssen. Stattdessen handelt es sich hierbei um Wörter, die bei diesem Thema zu erwarten sind und somit keine zusätzlichen, tieferen Erkenntnisse über den Datensatz liefern.

Tabelle 4.2: Zehn wahrscheinlichste Wörter von ausgewählten „Seeded Topics“ des Modells „Ka“ (keyATM bei Einsatz von Standard Collapsed Gibbs Sampling). Die Seed Wörter des entsprechenden „Seeded Topics“ wurden fett hervorgehoben.

ankommen	**Anwalt**	**Buch**	**Geld**	**Verein**
Canan	Song	Bild	**Geld**	Arzt
Villa	Lied	Hammer	**Euro**	**Screenshot**
Wohnung	Video	Foto	**€**	Hand
Flug	Musik	Schuh	Mail	weh
buchen	Mukke	aussehen	**schicken**	Schmerzen
Hotel	alt	schwarz	kaufen	Glückwunsch
kosten	happy	überlegen	bestellen	Bauch
Auto	Spotify	laufen	kosten	herzlich
zimmern	Playlist	Idee	Preis	Körper
Antalya	listen	Inga	Karte	Honig

In Tabelle 4.3 sind die Wörter mit der höchsten Wahrscheinlichkeit in den ausgewählten „Seeded Topics“ dargestellt, die mit dem „KaT“, d.h. dem Algorithmus keyATM unter Einbeziehung der zusätzlichen Termgewichtung, detektiert wurden. Im Vergleich zu den zuvor aufgezeigten Themen von keyATM mit gewöhnlichem Collapsed Gibbs Sampling fällt vor allem auf, dass die wahrscheinlichsten Wörter des Themas „Verein“ nicht mehr das Thema „Gesundheit“ repräsentieren. Stattdessen konnte bei näherer Betrachtung des Kontextes der Begriffe im Datensatz erkannt werden, dass sich Wörter wie das Live-Streaming-Videoportal „Twitch“, „Messen“, wobei es sich um das Lemma von „Messe“ handelte, und „Gamevention“ auf den Verein Streaming with Heart (SWH) bezogen und somit tatsächlich zum gewünschten Thema passten. Das Thema „Geld“ konnte erneut hervorgebracht werden, enthielt aber im Gegensatz zu dem in Tabelle 4.3 abgebildeten Thema Wörter wie „Stream“ und „spenden“ sowie den Vornamen „Marvin“, die im Gegensatz zu Wörtern wie „kosten“ oder „Preise“ nicht unbedingt im Zusammenhang mit diesem Thema erwartet werden bzw. präziser sind. Dies deutet darauf hin, dass durch die in Abschnitt 3.3.3.1 beschriebene Termgewichtung tatsächlich informativere Wörter stärker gewichtet und allgemeinere Begriffe abgewertet werden.

Tabelle 4.3: Zehn wahrscheinlichste Wörter von ausgewählten „Seeded Topics“ des Modells „KaT“ (keyATM unter Einbeziehung eines Termgewichtungsschemas). Die Seed Wörter des entsprechenden „Seeded Topics“ wurden fett hervorgehoben und Seed Wörter für ein anderes „Seeded Topic“ mit einem Stern markiert.

ankommen	**Anwalt**	**Buch**	**Geld**	**Verein**
ankommen	Song	Bild	**€**	Hamburg
Samstag	Menschen	Kopf	**Euro**	Fuba
planen	Text	**Seiten**	Mail	Event
Abend	deutsch	benutzen	**Geld**	Woche
Party	dick	aussehen	zahlen	Gaming
Freitag	**Anwalt**	zufrieden	Stream	Twitch
Jan	schlimm	winden	Marvin	Messen
Hotel	schwarz	bauen	kosten	melden
Lee	jung*	Gesicht	spenden	Gamevention
jung*	scheinbar	mausen	Statement	Termin

Bei den Themen „ankommen“, „Anwalt“ und „Buch“ waren es jedoch vor allem allgemeine, wenig informative Begriffe wie „Abend“ und „Freitag“ beim Thema „ankommen“, „scheinbar“ beim Thema „Anwalt“ und „benutzen“ und „zufrieden“ beim Thema „Buch“, die eine hohe Wahrscheinlichkeit aufwiesen. Dies war insofern überraschend, als „Abend“ eine Dokumentenfrequenz von 432 aufwies und „Freitag“ in insgesamt 194 bereinigten Konversationsdokumenten vorkam. Zum Vergleich sei angemerkt, dass speziellere Begriffe wie „Marvin“ nur eine Dokumentenfrequenz von 47 aufwiesen. Gerade Begriffe, die in vielen Dokumenten vorkommen, sollten durch die in keyATM integrierte Termgewichtung eine geringere Wahrscheinlichkeit in den Themen aufweisen. Es bedarf weiterer Untersuchungen, um zu klären, warum das Thema „ankommen“ dennoch von diesen unspezifischen Wörtern mit hoher Dokumentenfrequenz beherrscht wird. Generell ist zu beachten, dass gerade von Li u. a. [127] betont wurde, dass das PMI auf dem das verwendete Termgewichtungsschema basiert, für kurze Texte teilweise unzuverlässige Ergebnisse liefern kann, die sich nach Ansicht der Autoren auch negativ auf die erzeugten Themen auswirken können.

Jedoch ist darauf hinzuweisen, dass nicht alle unspezifischen Wörter in den Themen „ankommen“, „Anwalt“ und „Buch“ eine hohe Dokumentfrequenz aufwiesen. Zum Beispiel kam das Wort „zufrieden“ lediglich 75-mal und „benutzen“ sogar nur 45-mal in den Konversationsdokumenten vor. Demnach könnte eine weitere Möglichkeit der Verbesserung darin bestehen, dass anstelle des hier angewendeten Termgewichtungsschemas eine der anderen Gewichtungen aus Abschnitt 2.5 angewendet wird, da laut Yang u. a. [190] die Dokumentenfrequenz nicht unbedingt das entscheidende Kriterium für die Bedeutung eines Begriffs für die Themen ist.

4.2.2 Seeded LDA

Die wahrscheinlichsten Wörter der „Seeded Topics“, die mit dem Modell „S-8“, der Seeded LDA detektiert wurden, können Tabelle 4.4 entnommen werden. Wie in dieser zu sehen ist, befanden sich im Gegensatz zu den durch keyATM extrahierten „Seeded Topics“ unter den

zehn wahrscheinlichsten Wörtern jeweils mindestens zwei Seed Wörter. Jedoch ist ein Problem darin zu sehen, dass insbesondere bei den Themen „ankommen“, „Buch“ und „Verein“ nicht ersichtlich war, inwiefern die Seed Wörter, die eine hohe Wahrscheinlichkeit in dem Thema aufwiesen, mit den anderen wahrscheinlichen Begriffen in Beziehung standen. Zu den zehn wahrscheinlichsten Wörtern des Themas „Buch“ zählen zwar die Seed Wörter „Seiten“, „Buch“ und „Wörter“, jedoch erscheinen die anderen Begriffe zusammenhanglos, was zu einer geringen Interpretierbarkeit des Themas führt. Betrachtet man die wahrscheinlichsten Wörter der Themen „ankommen“ und „Verein“ ausgenommen der vorhandenen Seed Wörter, würde man davon ausgehen, dass diese eher Themen über Computerspiele beziehungsweise Elektronikartikel und über Familie darstellen und somit nicht die gewünschten Themen beschreiben. Die auftretenden Seed Wörter fallen hingegen aus der Reihe. Bezüglich des Themas „Geld“ ist auffallend, dass es sich bei den zehn Wörtern mit der höchsten Wahrscheinlichkeit ausschließlich um Seed Wörter handelt.

Tabelle 4.4: Zehn Wörter mit der höchsten Wahrscheinlichkeit in ausgewählten „Seeded Topics“ des Modells „S“, der Seeded LDA. Die Seed Wörter des entsprechenden „Seeded Topics“ wurden fett geschrieben und Seed Wörter für ein anderes „Seeded Topic“ mit einem Stern markiert.

ankommen	**Anwalt**	**Buch**	**Geld**	**Verein**
spielen	**prüfen**	Canan	**Euro**	**privat**
ankommen	Stream	**Seiten**	**Geld**	**Screenshot**
PC	Twitch	Wohnung	**€**	Canan
Game	**anzeigen**	Villa	**schicken**	**Club**
Handy	**Anwalt**	Problem	**senden**	**vereinen**
zocken	Twitter	**Buch**	**Kohlen**	Mutter
Rechner	Discord	Nikah	**PayPal**	Eltern
Steam	deutsch	Flug	**erhalten**	**Steffi**
sitzen	Marvin	**Wörter**	**überweisen**	Frau
BHF	lesen	beten	**verdienen**	Feek

Für diese Ergebnisse bestehen mehrere Erklärungsansätze. Zum einen stehen diese im Einklang mit bisherigen Erkenntnissen von Jagarlamudi u. a. [27], Kumar u. a. [356] und Andrzejewski und Zhu [357] bezüglich des Einsatzes von einer asymmetrischen a-priori-Verteilung für die Integration von Vorwissen in die Themenmodellierung, zu denen ebenfalls die Seeded LDA zählt. So wurde von Jagarlamudi u. a. [27] und Andrzejewski und Zhu [357] auf das Problem hingewiesen, dass diese Algorithmen dazu tendieren, Themen zu generieren, die nahezu nur Seed Wörter enthalten. Von Kumar u. a. [356] wurde zudem betont, dass durch die asymmetrische a-priori-Verteilung zwar das Vorwissen des Nutzers einen hohen Einfluss auf die Themen nehmen kann, jedoch diese inkohärent sein können. Zum anderen ist ein möglicher Grund für die hohe Anzahl an „Seed Wörtern“ unter den zehn wahrscheinlichsten Begriffen, insbesondere bei dem Thema „Geld“, in der Wahl des Wertes für den Hyperparameter μ zu sehen, der auf einen nach Watanabe und Baturo [15] hohen Wert von $\mu = 0.05$ gesetzt wurde. Daher weisen Seed Wörter a-priori eine hohe Wahrscheinlichkeit in dem Thema auf.

Auffallend ist zudem, dass es sich bei neun der zehn wahrscheinlichsten Wörter in dem „Seeded Topic“ „Geld“, mit Ausnahme von „PayPal“, um diejenigen Wörter der Seed Wortmenge handelt, die die höchste Frequenz im Datensatz aufweisen. Ebenfalls waren die Seed Wörter, die in den anderen „Seeded Topics“ unter den zehn wahrscheinlichsten Begriffen des Themas erschienen wie beispielsweise „prüfen“ in dem „Seeded Topic“ „Anwalt“ die häufigsten Wörter der entsprechenden Seed Wortmenge. Dies kann darauf zurückgeführt werden, dass, wie in Abschnitt 3.3.3.2 beschrieben wurde, die Stärke des Einflusses der Seed Wörter auf das entsprechende „Seeded Topic“ neben dem Parameter μ ebenfalls durch die Frequenz des Wortes im Datensatz bestimmt wird. Da hier $\mu = 0.05$ gesetzt wurde, ergab sich beispielsweise für „privat“' als Seed Wort für das Thema „Verein“, das insgesamt 74-mal in dem Nachrichten enthalten war, eine zusätzliche Pseudofrequenz von $\omega = 0.0021$, während die zusätzliche Pseudofrequenz von selteneren, meist spezielleren und relevanteren Begriffen deutlich geringer war. Beispielsweise betrug sie für die Seed Wörter „Vorständen“ und „Rechtspersönlichkeit“ des Themas „Verein“, die nur einmal in dem Datensatz auftraten, nur $\omega = 2.8732e - 05$.

Hinsichtlich des „Seeded Topics“ „Buch“ ist ferner erkennbar, dass Wörter wie „Villa“ und „Flug“ über eine hohe Wahrscheinlichkeit verfügen. Eine mögliche Ursache ist darin zu sehen, dass durch die Lemmatisierung sowohl das Wort „Buch“ als auch das Verb „buchen“ auf das Lemma „buchen“ abgebildet wurden und somit im Rahmen der Themenmodellierung als dasselbe Wort aufgefasst wurden. Es ist an dieser Stelle anzumerken, dass der Einsatz der Lemmatisierung für die Themenmodellierung umstritten ist [187, 318, 358–360]. Einerseits wurde von Churchill und Singh [187] betont, dass die Lemmatisierung dem Problem der hohen Sparsität entgegenwirkt und ferner wurde von Lau u. a. [318] und Martin und Johnson [358] hervorgehoben, dass diese zu kohärenteren und qualitativ hochwertigeren Themen führen kann, weshalb sie ebenfalls in dieser Arbiet angewendet wurde. Andererseits wurde von Laureate u. a. [359] und Brookes und McEnery [360] entgegnet, dass diese zwar darauf abzielt, Wörter mit derselben Bedeutung zusammenzufassen, jedoch in manchen Fällen auch Wörter beziehungsweise verschiedene morphologische Formen mit verschiedenen Bedeutungen fälschlicherweise als ein Wort angesehen werden und somit die Interpretierbarkeit der Themen erschwert wird. Allerdings bedarf es weiteren Untersuchungen, um festzustellen, ob tatsächlich die mangelhafte Interpretierbarkeit des Themas „Buch“ ausschließlich auf die Lemmatisierung zurückzuführen sind.

4.2.3 Auswirkungen der CluWords auf die Ergebnisse

Im folgenden Abschnitt wird auf die Ergebnisse eingegangen, die durch das keyATM und die Seeded LDA erreicht wurden, wenn die Dokumente durch die in Abschnitt 3.4 beschriebene CluWords Dokumentenrepräsentation angereichert wurden. In den Ergebnistabellen werden jeweils Wörter, die zu dem CluWord des Topic Labels des jeweiligen „Seeded Topic“ gehörten, durch ein „[C]“ markiert und Wörter, die in einem CluWord des Topic Labels eines anderen „Seeded Topics“ vorkamen, mit einem „[C*]“ kenntlich gemacht.

4.2.3.1 CluWords basierend auf Word Embedding Ähnlichkeit

In Tabelle 4.5 sind die zehn wahrscheinlichsten Begriffe der untersuchten „Seeded Topics“ des Modells „KaCF-8“ dargestellt, das keyATM als Algorithmus verwendet und CluWords basierend auf fastText Word Embeddings einbezieht. Wie in der Tabelle zu sehen ist, befinden sich bei den „Seeded Topics“ „ankommen“ und „Geld“ unter den zehn wahrscheinlichsten Begriffen ausschließlich die Wörter des CluWords des entsprechenden Topic Labels. Die Wörter, die eine hohe Wahrscheinlichkeit in dem Thema „ankommen“ aufweisen, wie „bestellen“ und „pünktlich“, stehen in Verbindung mit der Ankunft eines Pakets. Die wahrscheinlichsten Wörter des Themas „Geld“ würde man intuitiv mit dem entsprechenden Thema assoziieren. Jedoch fehlen bei diesem speziellere Begriffe wie beispielsweise Eigennamen oder Wörter, die mit konkreten Spendenaktionen in Verbindung stehen. Diese Ergebnisse deuten somit auf das in Abschnitt 2.4 beschriebene Problem bei der Integration von externen Word Embeddings in die Themenmodellierung hin, wonach hierdurch nicht aufgezeigt wird, wie ein Thema in dem konkreten Datensatz besprochen wurde, sondern wie es allgemein aufgefasst wird.

Unter den Wörtern mit höchster Wahrscheinlichkeit in den Themen „Buch“ und „Verein“ befindet sich hingegen höchstens ein Begriff des CluWords des entsprechenden Topic Labels. Abgesehen von dem Begriff „lesen“, der in dem CluWord von „Buch“ vorhanden war, beziehen sich die Top Wörter des Themas „Buch“ eher auf das Thema „Musik“. Dies lässt vermuten, dass das eigentlich gewünschte Thema nicht gefunden werden konnte und stattdessen von einem dominanteren Thema verdrängt wurde. Hingegen wurden die in Tabelle 4.5 aufgeführten Wörter zum Thema „Verein“ bei der Betrachtung ihres Kontextes im Datensatz als für Charity Stream Events des Vereins SWH identifiziert.

Eine mögliche Ursache dafür, dass nur bestimmte Themen von den CluWords dominiert wurden und in anderen Themen keine oder wenige Wörter des CluWords eine hohe Wahrscheinlichkeit erhielten, kann in der Dokumentenfrequenz der Topic Labels gesehen werden. Wie in Abschnitt 3.4 erklärt wurde, wurden nur die Topic Label jeder Seed Wortmenge durch ihr CluWord ersetzt. Somit wurden ausschließlich Dokumente um semantisch ähnliche Wörter angereichert, in denen ein Topic Label auftrat. Während Wörter wie „Buch“ nur in elf und „Verein“ beziehungsweise sein Lemma „vereinen“ in 14 Konversationsdokumenten auftraten, war der Begriff „Schwester“ in 90 Dokumenten vorhanden und der Begriff „Geld“ in 229 Konversationsdokumenten. Dementsprechend wurden die semantisch ähnlichen Begriffe der CluWords zu den Topic Labels „Geld“ und „Schwester“ in deutlich mehr Konversationsdokumente eingefügt, was nahelegt, dass diese einen größeren Einfluss auf die Themenmodellierung hatten als die Wörter der CluWord für andere Topic Labels.

Zudem stehen diese Ergebnisse im Einklang mit den Erkenntnissen aus der Durchführung von keyATM ohne Einbeziehung von CluWords, dass bestimmte gewünschte Themen eher hervorgebracht werden konnten als andere (siehe Abschnitt 4.2.1). Wie in Tabelle 4.2 gezeigt wurde, erwies es sich ebenfalls ohne Verwendung von CluWords als problematisch, die Themen „Buch“ und „Verein“ zu finden. Darüber hinaus ist auffallend, dass sich unter den Top Wörtern des Themas „Anwalt“ Wörter wie „Blaulicht“, „Notruf“ und „flüchten“ befanden, die zwar nicht zu dem CluWord des Topic Labels „Anwalt“ zählten, sich jedoch in dem CluWord

des Topic Labels eines anderen Themas, genauer gesagt des Themas „Polizei“, befanden. Die Wörter „Staatsanwaltschaft“ und „Justiz“ waren sowohl in dem CluWord von „Anwalt“ als auch von „Polizei“ vorhanden. Eine mögliche Ursache ist darin zu sehen, dass eine Überschneidung beziehungsweise Korrelation zwischen den Themen „Anwalt“ und „Polizei“ besteht. Um diese Hypothese zu überprüfen, sind weitere Analysen erforderlich. Hierbei sollte untersucht werden, wie hoch der Anteil der Dokumente ist, die in beiden Themen eine hohe Wahrscheinlichkeit aufweisen. Zudem sollten die Ergebnisse mit den Resultaten von Algorithmen wie das Community Topic Model (CTM) [314] und STM [11] verglichen werden, die speziell zur Erfassung von Korrelationen zwischen Themen entwickelt wurden. Gegen diese Annahme spricht jedoch die Tatsache, dass in dem Thema „Polizei“ die CluWords für das Topic Label „Anwalt“ keine hohe Wahrscheinlichkeit aufweisen, sondern vor allem allgemeine und wenig aussagekräftige Wörter wie „Handy“, „Mail“ und „melden“ in diesem vorkommen. Ein alternative mögliche Erklärung liegt in der Überschneidung zwischen den CluWords von „Anwalt“ und „Polizei“, da diese beide die Begriffe „Staatsanwaltschaft“ und „Justiz“ beinhalteten. Hierdurch könnte es sich bei dem „Seeded Topic“ „Anwalt“ um ein sogenanntes „Chained Topic“ handeln. Hierunter wird nach Mimno u. a. [19] verstanden, dass die wahrscheinlichsten Wörter eines Themas zwei verschiedene Konzepte beschreiben, die durch ein Wort miteinander verbunden werden. In diesem Fall könnte beispielsweise das Wort „Justiz“ eine Verbindung zwischen dem Thema „Anwalt“ und den Wörtern, die eigentlich charakteristisch für das Thema „Polizei“ sind, herstellen.

Tabelle 4.5: Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „KaCF-8“ (keyATM unter Einbeziehung von CluWord basierend auf Word Embedding Ähnlichkeit). Die Seed Wörter des entsprechenden „Seeded Topics“ wurden fett hervorgehoben und die Seed Wörter für ein anderes „Seeded Topic“ mit einem Stern markiert. Zudem wurden Wörter, die in dem CluWord des Topic Labels des jeweiligen Themas enthalten waren, mit einem „[C]“ und Wörter, die in einem CluWord des Topic Labels eines anderen Themas vorkamen, mit einem „[C*]“ gekennzeichnet.

ankommen	**Anwalt**	**Buch**	**Geld**	**Verein**
bestellen [C]	Staatsanwaltschaft [C]	Song	**Euro [C]**	Stream
Paket [C]	Justiz [C]	Lied	kaufen [C]	€*
abholen [C]	Blaulicht [C*]	lesen [C]	**überweisen [C]**	Twitch
warten [C]	Notruf [C*]	Musik	kosten [C]	Event
erreichen [C]	flüchten [C*]	Playlist	teuer [C]	Twitter
pünktlich [C]	Polizisten [C*]	Mukke	bezahlen [C]	Team
verschicken [C]	Feuerwehr [C*]	Filmen	zahlen [C]	Fuba
versenden [C]	Behörde [C*]	Ohrwurm	ausgeben [C]	Gaming
auspacken [C]	Beamte [C*]	gönnen	Konto [C]	Marvin
zurückschicken [C]	Behörden [C*]	Spotify	spenden [C]	Streamer

Die in Tabelle 4.5 aufgeführten Themen hatten gemeinsam, dass sich unter den zehn wahrscheinlichsten Wörtern kein beziehungsweise nur wenige Seed Wörter befanden. Hingegen wurden die meisten der in Tabelle 4.6 dargestellten „Seeded Topics“ des Modells „SCF-8“ von den Seed Wörtern dominiert. Dieses Modell unterscheidet sich von dem zuvor beschriebenen Modell „KaCF-8“ nur dadurch, dass statt des Algorithmus keyATM die Seeded LDA verwendet

wurde. Bei den Seed Wörtern mit einer hohen Wahrscheinlichkeit in den Themen „Buch“ und „Geld“ handelt es sich überwiegend um Wörter, die ebenfalls in dem CluWord des entsprechenden Topic Labels vorhanden waren. Die wahrscheinlichsten Seed Wörter des Themas „Anwalt“ umfassten sowohl Begriffe des CluWords von „Anwalt“ als auch des CluWords von „Polizei“, nämlich „ermitteln“, und des CluWords von „Geld“, nämlich „versteuern“. Zudem verfügten vor allem die häufigsten Seed Wörter der entsprechenden Seed Wortmenge wie beispielsweise „prüfen“ und „anzeigen“ im Thema „Anwalt“ und „privat“ und „Screenshot“ des Themas „Verein“ über eine hohe Wahrscheinlichkeit. Die Ergebnisse deuten erneut darauf hin, dass die CluWords teilweise einen starken Einfluss auf die Themen haben, sprechen aber auch für die in Abschnitt 4.2.2 aufgestellte Annahme, dass bei der Seeded LDA vor allem hochfrequente Seed Wörter die Themen beherrschen.

Tabelle 4.6: Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „„SCF-8“ (Seeded LDA unter Einbeziehung von CluWord basierend auf Word Embedding Ähnlichkeit). Die Seed Wörter des entsprechenden „Seeded Topics“ wurden fett geschrieben. Wörter, die zu dem CluWord des Topic Labels des entsprechenden Themas gehörten, wurden mit einem „[C]“ und Wörter, die in einem CluWord des Topic Labels eines anderen Themas vorhanden waren, mit einem „[C*]“ kenntlich gemacht.

ankommen	**Anwalt**	**Buch**	**Geld**	**Verein**
ankommen [C]	**versteuern [C*]**	**Seiten**	**Euro [C]**	**privat**
Canan	**prüfen**	Bild	**überweisen [C]**	**Screenshot**
Wohnung	Energy	**Duschenbuch [C]**	**verdienen [C]**	**Club**
Villa	**ermitteln [C*]**	**Kapitel [C]**	**money [C]**	Stream
Auto	**Klage [C]**	lesen [C]	**Vermögen [C]**	**Steffi**
warten [C]	**anzeigen**	**Leseratte [C]**	**Gelder [C]**	eintragen
Flug	**Anwalt [C]**	Star	**Kleingeld [C]**	Twitch
bestellen [C]	hungern	**Buch [C]**	**Geldern [C]**	**vereinen [C]**
abholen [C]	schmecken	**Wörter**	**Geldbündel [C]**	Twitter
erreichen [C]	Penny	zeigen	**Taschengeld [C]**	Event

Darüber hinaus tritt insbesondere beim Thema „Anwalt“ das bereits in Abschnitt 4.2.2 beschriebene Problem auf, dass die auftretenden Seed Wörter nicht zu den anderen Wörtern mit hoher Wahrscheinlichkeit in diesem Thema passen. Dieses Thema kann als „Chained Topic“ [19] bezeichnet werden. In diesem besitzen zum einen Wörter eine hohe Wahrscheinlichkeit, die tatsächlich für das Thema relevant sind, wie „versteuern“, „prüfen“, „Klage“ und „Anzeigen“, die entweder Wörter des CluWords oder Seed Wörter des Themas sind. Zum anderen verfügen ebenfalls irrelevante Wörter wie „hungern“, „schmecken“ und „Penny“ über eine hohe Wahrscheinlichkeit, die mit Themen wie „Lebensmittel“ oder „Nahrung“ verbunden werden. Wurden die 50 wahrscheinlichsten Begriffe betrachtet, zeigte sich, dass diese ebenfalls sowohl Wörter in Zusammenhang mit dem Thema „Anwalt“ als auch über Essen enthielten. Eine mögliche Ursache für dieses Ergebnis liegt in dem Problem der Polysemie, da das Wort „Gericht“ im CluWord von „Anwalt“ enthalten war. Dieses wurde somit in je-

des Konversationsdokument eingefügt, in dem das Wort „Anwalt“ enthalten war, trat jedoch weiterhin beispielsweise auch in Konversationsdokumenten über afrikanische Gerichte oder Fertiggerichte auf.

Für das Thema „Verein“ wiesen zwar ebenfalls vor allem Seed Wörter eine hohe Wahrscheinlichkeit auf, die anderen in Tabelle 4.6 aufgeführten Begriffe können jedoch intuitiv ebenfalls mit dem Thema in Verbindung gebracht werden. Zum Beispiel könnten Wörter wie „Stream“, „Twitch“ und „Twitter“ Social Media Kampagnen eines Vereins beschreiben. Dass die Wörter in diesem Thema eher als relevant empfunden wurden, stimmte mit den Erkenntnissen durch das Modell „KaCF-8“ überein, bei dem ebenfalls das Thema „Verein“ vor allem bedeutungsvolle Begriffe enthielt.

Bei den Themen „ankommen“ und „Geld“, die eine hohe Anzahl an CluWords enthielten, konnte erneut wie bei dem Modell „KaCF-8“ festgestellt werden, dass diese keine spezifischen Informationen über den Fall oder den Datensatz vermitteln. Stattdessen treten in dem Thema „ankommen“ Wörter mit einer hohen Wahrscheinlichkeit auf, die in Verbindung mit der Ankunft an einem Urlaubsziel sowie mit der Ankunft eines Pakets stehen. Hinsichtlich des Themas „Geld“ fehlten erneut speziellere, unerwartete Begriffe.

4.2.3.2 CluWords basierend auf paradigmatischen Relationen

Wenn die CluWords auf der Grundlage paradigmatischer Relationen anstelle von Word Embeddings erstellt wurden, konnte erneut festgestellt werden, dass insbesondere bei dem Algorithmus keyATM einige Themen durch die CluWords des jeweiligen Topic Labels dominiert werden und kaum Seed Wörter in den Themen vorkommen. Die wahrscheinlichsten Wörter der ausgewählten „Seeded Topics“ des Modells „KaCP-8“ werden in Tabelle 4.7 aufgeführt. Wie dieser entnommen werden kann, wurden bis auf das Thema „ankommen“ dieselben Themen wie bei dem Modell „KCF-8“ (keyATM mit CluWords basierend auf Word Embeddings) von den Begriffen der CluWords dominiert. Konkret wiesen vor allem in den Themen „Anwalt“ und „Geld“ die Begriffe der CluWords eine hohe Wahrscheinlichkeit auf.

Zu diesen zählten Wörter, die tatsächlich zu dem Thema passen könnten wie „belangen“ und „verpflichten“ in dem Thema „Anwalt“ und „Justizkasse“ und „Einzugsermächtigung“ in dem Thema „Geld“. Diese sind spezifischer und aussagekräftiger als Wörter wie „abholen“ in dem Thema „ankommen“ sowie „kaufen“ und „zahlen“ in dem Thema „Geld“ des Modells „KaCF“ (siehe Tabelle 4.5), welches Word Embeddings zur Bildung der CluWords verwendete. Jedoch führten die CluWords auch irrelevante Wörter in die Themen ein. So gehörten beispielsweise zu dem CluWord von „Anwalt“ Wörter wie „viiiiiiiiel“ und zu dem CluWord von „Geld“ die Begriffe „boahhh“ und „Hagenbeckstierpark“. Der wesentliche Unterschied zu den auf Word Embeddings basierenden CluWords besteht darin, dass der Ansatz basierend auf paradigmatischen Relationen die semantische Ähnlichkeit von Wörtern aus dem Trainingsdatensatz der Themenmodelle ableitet. Dementsprechend ist anzunehmen, dass die mangelhafte sprachliche Qualität der umgangssprachlichen Kurznachrichten und ihre hohe Anzahl an irrelevanten Noise Words auch einen negativen Einfluss auf die paradigmatischen Relationen und somit auf die Ergebnisse der Themenmodellierung hat.

Beim Vergleich der Themenwörter der Modelle „KaCF“ in Tabelle 4.5 und „KaCP“ in Tabelle 4.7 fällt auf, dass bei dem Thema „ankommen“ im Modell „KaCF“ ausschließlich Begriffe des entsprechenden CluWords unter den zehn wahrscheinlichsten Wörtern vorhanden waren. Im Gegensatz dazu trat bei dem Modell „KaCP“ kein einziges CluWord unter den Wörtern mit der höchsten Wahrscheinlichkeit in dem Thema „ankommen“ auf. Hingegen würde man die wahrscheinlichsten Begriffe eher mit einem Thema über Gaming oder Veranstaltungen assoziieren. Möglicherweise könnte eine Ursache für den unterschiedlichen Einfluss der CluWords auf dieses Thema, je nach verwendeter Methode zur Bildung der CluWords, darin liegen, dass das mit fastText ermittelte CluWord über mehr Begriffe zum Topic Label „ankommen“ verfügte als das auf paradigmatischen Relationen basierende CluWord. Allerdings sind die Unterschiede gering. So umfasste das auf fastText basierende CluWord 19 Wörter, während das mit paradigmatischen Relationen gebildete CluWord elf Wörter umfasste. Eine alternative Erklärung ist, dass sich in dem fastText-basierten CluWord zum Topic Label „ankommen“ häufigere Wörter befanden, während das mit paradigmatischen Relationen erstellte CluWord vor allem seltene und ungewöhnliche Wörter wie „vermezler“ und „hättendann“ beinhaltete. Zum Beispiel wurde festgestellt, dass die Wörter des CluWords, gebildet mit paradigmatischen Relationen, eine maximale Dokumentenfrequenz von zwei aufwiesen. Im Gegensatz dazu traten Begriffe wie „warten“ und „bestellen“ des fastText-basierten CluWords in über 100 Konversationsdokumenten auf. Es liegt die Vermutung nahe, dass die auf paradigmatischen Relationen basierenden Wörter des CluWords trotz ihrer künstlichen Hinzufügung zu weiteren Dokumenten nicht präsent genug waren, um einen Einfluss auf die Themenmodellierung zu haben. Es bedarf jedoch weiterer Untersuchungen, um festzustellen, was ausschlaggebend dafür ist, dass die CluWords eine hohe Wahrscheinlichkeit in den Themen aufweisen.

Die wahrscheinlichsten Wörter der „Seeded Topics“ „Buch“ und „Verein“ erschienen überwiegend als irrelevant für diese Themen. Das Thema „Buch“ enthielt hauptsächlich Wörter, die thematisch zu Lebensmitteln passen, während das Thema „Verein“ von allgemeinen Begriffen dominiert wurde, zwischen denen kein Zusammenhang erkennbar war. In Bezug auf das Thema „Verein“ fällt besonders auf, dass es durch das Modell “KaT-8“ (siehe Tabelle 4.3) identifiziert werden konnte. Die Modelle „Ka-8“ (siehe Tabelle 4.2) und „KaCP-8“ (siehe Tabelle 4.6) konnten dieses Thema hingegen nicht erkennen. Ein möglicher Verbesserungsansatz wäre die Verwendung eines Termgewichtungsschemas, ähnlich wie beim Modell „kaT“. Ein konkreter Ansatz wäre die Integration der von Viegas u. a. [16] speziell für CluWords vorgeschlagenen Gewichtung, die ursprünglich für den Algorithmus NMF entwickelt wurde, in das Collapsed Gibbs Sampling der halbüberwachten LDA.

Tabelle 4.7: Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „KaCP-8“ (keyATM unter Berücksichtigung von CluWord basierend auf paradigmatischen Relationen). Die Seed Wörter des entsprechenden „Seeded Topics“ wurden fett hervorgehoben und die Seed Wörter für ein anderes „Seeded Topic“ mit einem Stern markiert. Das „[C]“ kennzeichnet Wörter, die in dem CluWord des Topic Labels des entsprechenden Themas enthalten waren, die in einem CluWord des Topic Labels eines anderen Themas auftraten.

ankommen	**Anwalt**	**Buch**	**Geld**	**Verein**
Fuba	Twitch	energy	**Euro** [C]	Kopf
Hamburg	lediglich [C]	schmecken	**Geld [C]**	Jahr
Event	Typ	kaufen	Justizkasse [C]	leben
Termin	belangen [C]	Penny	Rückfuhr [C]	merken
Oli	verpflichten [C]	lecker	tatsächloch [C]	nerven
Gaming	anlama [C]	hungern	boahhh [C]	Menschen
Gameventi-on	Vereinsgründer* [C]	kochen	Abschlag [C]	kennen
melden	Partner	wassern	asare [C]	erzählen
Festival	Internetangele-genheiten [C]	Brot	Einzugsermächtigung [C]	anfangen
Gespräch	viiiiiiiiel [C]	reisen	Hagenbeckstierpark [C]	stellen

Die Wörter mit der höchsten Wahrscheinlichkeit der „Seeded Topics“ des Modells „SCP-8“ können Tabelle 4.8 entnommen werden. Analog zu den Ergebnissen des Modells „SCF-8“ (Seeded LDA mit CluWords basierend auf fastText-Embeddings) befinden sich unter den wahrscheinlichsten Wörtern jedes dargestellten Themas sowohl mehrere Seed Wörter als auch Begriffe des CluWords des entsprechenden Topic Labels. Das Problem, dass druch die auf paradigmatischen Relationen basierenden CluWords Rauschen beziehungsweise irrelevante Wörter in die Themen eingebracht werden, ist bei der Seeded LDA geringer als bei den zuvor dargestellten Ergebnissen des Modells „KaCP-8“ (siehe Tabelle 4.7). Dies ist jedoch vor allem darauf zurückzuführen, dass die meisten Wörter, die in den CluWords in den Themen vorkommen, auch Seed-Wörter sind. Diese sind zwar relevant, liefern aber keine neuen Details zum gewünschten Thema.

Tabelle 4.8: Zehn wahrscheinlichste Wörter in ausgewählten „Seeded Topics“ des Modells „SCP-8“ (Seeded LDA unter Einbeziehung von CluWord basierend auf paradigmatischen Relationen). Bei den fett geschriebenen Wörtern handelt es sich um die Seed Wörter des entsprechenden „Seeded Topics“. Mit dem „[C]“ wurden Wörter markiert, die in dem CluWord des Topic Labels des entsprechenden Themas enthalten waren und mit dem „[C*]“ Wörter, die in einem CluWord des Topic Labels eines anderen Themas enthalten waren.

ankommen	**Anwalt**	**Buch**	**Geld**	**Verein**
Song	Woche	Bild	**Euro [C]**	**privat**
Briefes [C]	**prüfen**	**Seiten**	**Geld [C]**	**Screenshot**
ankommen [C]	melden	Video	**€**	**Club**
Bahnen	Hamburg	Typ	**schicken**	**Steffi**
fahren	**anzeigen**	**Wörter**	**senden**	**rechtsfähig [C]**
Lied	Termin	**Buch [C]**	**kohlen**	**Vereinsregister [C]**
Bus	**dokumnetie-ren [C]**	zeigen	**überweisen**	**Vereinsgründer [C*]**
Minuten	**Fachgebiet [C]**	Seite	**erhalten**	Stream
warten [C]	**gestriked [C]**	Insta	Justizkasse [C]	**Rechtspersönlichkeit [C]**
zuhause	**Anwalt [C]**	antworten	Rückfuhr [C]	eingetragen [C]

Bei Wörtern, die weder Seed Words noch Begriffe des CluWords sind und dennoch unter den zehn wahrscheinlichsten Begriffen vorkommen, hängt es vom jeweiligen Thema ab, ob diese dazu passen. Zum Beispiel beschreiben die meisten Wörter im Seeded Topic „ankommen“ tatsächlich eine Ankunft, beispielsweise eines Briefes oder eines Verkehrsmittels, mit Ausnahme der Wörter „Song“ und „Lied“. Allerdings tritt hier erneut das beispielsweise bereits in Abschnitt 4.2.1 beschriebene Problem auf, dass allgemeine Seed-Wörter wie „ankommen“ und „HBF“ zu Themen führen, bei denen es schwierig ist, die Relevanz des Themas für den Fall klar festzustellen. Wörter wie „Termin“ und „melden“ könnten im Zusammenhang mit dem Thema „Anwalt“ stehen. Gerade beim Thema „Buch“ zeigte sich jedoch erneut das Problem, dass die Verwendung von Wörtern wie „zeigen“ und „antworten“ die Interpretation erschwert.

4.2.4 Teilzusammenfassung

Zusammenfassend wurden im Rahmen der qualitativen Evaluierung die folgenden Feststellungen getroffen:

1. In Bezug auf den halbüberwachten Algorithmus keyATM haben die ausgewählten Seed-Wörter einen wesentlichen Einfluss auf die resultierenden „Seeded Topics“. Die Ergebnisse deuten darauf hin, dass, wie von Eshima u. a. [14] betont wurde, möglichst häufige und gleichzeitig für die Themen diskriminierende Seed Wörter gewählt werden sollten.
2. Durch die Einbeziehung des Termgewichts in keyATM konnten in einigen Fällen Themen gelernt werden, bei denen spezifische Begriffe unter den am häufigsten vorkommenden Wörtern erschienen.

3. Die Seeded LDA führte vor allem bei hohen Werten für den Hyperparameter μ zu Themen, bei denen die wahrscheinlichsten Wörter durch „Seed Wörter“ dominiert werden. Jedoch wurden diese intuitiv teilweise nicht mit den anderen Wörtern assoziiert.
4. Durch die CluWord-Repräsentation basierend auf fastText-Embeddings wurden Themen generiert, die vor allem die allgemeine Bedeutung des Themas präsentierten.
5. Die Einführung von CluWords basierend auf fastText-Embeddings führte zu Themen, bei denen sich unter den wahrscheinlichsten Wörtern ebenfalls speziellere Begriffe befanden, die man mit dem entsprechenden „Seeded Topic“ verbinden würde. Jedoch wurden hierdurch gerade bei keyATM ebenfalls umgangssprachliche, offensichtlich irrelevante Begriffe in die Themen eingeführt.
6. Die Anzahl der Begriffe der CluWords, die sich unter den wahrscheinlichsten Wörtern der Themen befanden, variierte stark je nach betrachtetem Thema. Mögliche Ursachen sind in der Dokumentenfrequenz der Topic Labels, der Anzahl der Wörter je CluWord und der Häufigkeit der Begriffe in den CluWords zu sehen. Allerdings sind weitere Untersuchungen erforderlich, um dies zu klären.

4.3 Quantitative, automatische Evaluierung

Im Folgenden werden die Ergebnisse präsentiert, die nach der quantitativen Evaluierung basierend auf der semantischen Kohärenz erreicht wurden. Tabelle 4.9 zeigt die mittlere semantische Kohärenz der einzelnen Themenmodelle, wobei für die Ansätze des Seed-Guided Topic Modellings sowohl der Mittelwert über die Kohärenzwerte über alle Themen des Modells, einschließlich der „Unseeded Topics“ als auch nur über die „Seeded Topics“ ermittelt wurde.

Wie in der Tabelle zu sehen ist, erzielten die Ansätze des Seed-Guided Topic Modelling, die zusätzlich die CluWords der Topic Labels einbeziehen, die höchsten mittleren Kohärenzwerte für die meisten Werte für die Anzahl der „Unseeded Topics“. Dabei erreichte das Modell „KCP-9“ die höchste mittlere semantische Kohärenz. Dieses verwendete als halbüberwachten Algorithmus zur Themenmodellierung keyATM und wurde auf Pseudodokumenten trainiert, in die CluWords basierend auf paradigmatischen Relationen eingefügt wurden. Die zweithöchste und dritthöchste semantische Kohärenz wurde durch die Modelle „KCP-7“ und anschließend „KCP-8“ erreicht. Diese unterschieden sich gegenüber dem Modell „KCP-9“ lediglich in der Anzahl der „Unseeded Topics“.

Die Baseline wurde für die meisten untersuchten Werte der „Unseeded Topics“ durch die Modelle „KCP“, „KCF“ und „SCP“ übertroffen. Diese haben gemeinsam, dass sie auf Pseudodokumenten trainiert wurden, in denen die Topic Labels durch ihr CluWord ersetzt wurden. Durch das Modell „KCF-7“ wurde zudem die vierthöchste mittlere semantische Kohärenz erzielt.

Das einzige halbüberwachte Modell, das ebenfalls CluWords integrierte, jedoch für keine Anzahl an „Unseeded Topics“ die Baseline übertreffen konnte, war „SCF“, für das die Seeded LDA als halbüberwachter Algorithmus gewählt wurde und die Konversationsdokumente mit CluWords basierend auf fastText-Word Embeddings angereichert wurden. Dennoch wies dieses Modell eine höhere mittlere Kohärenz als die halbüberwachten Modelle „K“, „S“ und „KT“

auf. Somit deuten die Ergebnisse der automatischen, quantitativen Evaluierung darauf hin, dass die Integration von CluWords in die halbüberwachte Themenmodellierung basierend auf keyATM und Seeded LDA zu kohärenteren Themen führt. Auffällig ist zudem, dass kein halbüberwachtes Modell ohne Einbeziehung von CluWords, d.h. die Modelle „K“, „S“ oder „KT“, die Baseline übertreffen konnte. Ob mit dem Modell „K“ (keyATM mit gewöhnlichem Collapsed Gibbs Sampling) oder „S“ (Seeded LDA) eine höhere mittlere Kohärenz erzielt wurde, hing von der Anzahl an „Unseeded Topics“ ab. Bei sieben und acht „Unseeded Topics“ erreichte keyATM eine höhere durchschnittliche semantische Kohärenz als Seeded LDA, während diese keyATM bei neun „Unseeded Topics“ geringfügig übertraf.

Die geringste mittlere Kohärenz wurde durch das Modell „KT-8“ erzielt, bei dem in keyATM ein Termgewichtungsschema integriert wurde. Ebenfalls über eine geringe mittlere semantische Kohärenz verfügte das Modell „S-7“, die unter der mittleren semantischen Kohärenz der Modelle „KT-7“ und „KT-9“ lag.

Darüber hinaus konnte festgestellt werden, dass es vom Modell abhing, bei welcher Anzahl an „Unseeded Topics“ dieses die höchste mittlere Kohärenz aufwies. Beispielsweise war bezüglich des Modells „KCF“ die mittlere Kohärenz bei sieben „Unseeded Topics“, bei dem Modell „K“ bei acht „Unseeded Topics“ und bei den Modellen „KCP“ und „S“ bei neun „Unseeded Topics“ am höchsten. Hierbei war weder eine Tendenz eines Algorithmus oder einer Vorgehensweise bezüglich der Bildung der CluWords für eine bestimmte Themenanzahl erkennbar.

Tabelle 4.9: Mittlere semantische Kohärenz aller Themen sowie der „Seeded Topics“ der untersuchten Modelle. Der linke Teil der Tabelle zeigt die mittlere semantische Kohärenz aller Themen der halbüberwachten Modelle sowie des Baseline-Systems, während der rechte Teil die mittlere Kohärenz der „Seeded Topics“ der halbüberwachten Themenmodelle enthält. „# UT“ steht für die Anzahl an „Unseeded Topics“ und „# GT“ für die Gesamtanzahl an Themen, mit denen der Algorithmus trainiert wurde. Ein höherer durchschnittlicher Kohärenzwert zeigt eine höhere Themenqualität an. Die höchsten erzielten Werte sowie das entsprechende Themenmodell, mit dem diese erzielt wurden, wurden in fett hervorgehoben.

Modell	**∅ Kohärenz**			**∅ Kohärenz „Seeded Topics“**		
# UT/ GT	7/ 15	8/ 16	9/ 17	7/ 15	8/ 16	9/ 17
LDA	-189.1183	-190.9525	-191.7222	-	-	-
K	-196.3179	-195.6066	-199.9291	-200.2574	-200.6043	-204.7911
KT	-198.8774	-205.3617	-200.3462	-194.859	-210.1596	-192.7125
KCF	-182.1798	-185.6687	-193.3893	-182.0187	-188.8275	-196.0022
KCP	-178.2448	-180.9875	**-172.4382**	-159.7928	-166.0204	**-151.2488**
S	-205.1136	-202.0467	-199.0984	-209.0337	-204.6664	-201.5256
SCF	-194.3595	-192.5342	-192.8242	-188.8223	-189.4315	-189.6191
SCP	-182.834	-190.0287	-191.1971	-176.6456	-187.0686	-179.9386

Die durchschnittliche semantische Kohärenz der „Seeded Topics“ ist in dem rechten Abschnitt von Tabelle 4.9 aufgeführt. Aus dieser geht hervor, dass das Modell „KCP-9“ ebenfalls die höchste mittlere Kohärenz über die „Seeded Topics“ aufweist, gefolgt von „KCP-7“ und „KCP-8“. Im Unterschied zur durchschnittlichen Kohärenz über alle Themen zeigte das Modell

„SCP-7“ anstelle des Modells „KCF-7“ die viertstärkste durchschnittliche Kohärenz über die „Seeded Topics“ auf. Zudem war die mittlere Kohärenz der „Seeded Topics“ der „SCP“ Modelle (für sieben, acht und neun „Unseeded Topics“) höher als bei allen anderen untersuchten halbüberwachten Modellen mit Ausnahme von „KCP“. Somit war die automatisch bestimmte mittlere semantische Kohärenz der „Seeded Topics“ der Modelle, die CluWords basierend auf paradigmatischen Relationen einbezogen, höher als die der Modelle, die CluWords basierend auf fastText-Embeddings bildeten. Unabhängig von der konkreten Methode zur Bildung der CluWords waren Modelle, die die CluWords enthielten, wie bereits bei der Kohärenzberechnung über alle Themen hinweg den anderen halbüberwachten Modellen überlegen. Die am wenigsten kohärenten „Seeded Topics“ wies „KT-8“ auf, das auch bereits die geringste mittlere Kohärenz aller Themen, einschließlich „Unseeded Topics“ lieferte.

Vergleicht man die mittlere Kohärenz über alle Themen und ausschließlich über „Seeded Topics“, fällt auf, dass bei den beiden halbüberwachten Modellen „K“ und „S“ ohne Einbeziehung von CluWords oder zusätzlicher Termgewichtung die Gesamtkohärenz höher war als die Kohärenz über „Seeded Topics“. Hingegen waren die Themen der halbüberwachten Modelle mit CluWords mit Ausnahme von „KCF“ besser, wenn nur die „Seeded Topics“ einbezogen wurden.

4.4 Evaluierung durch die Nutzerstudie

Der folgende Abschnitt beschreibt und diskutiert die Ergebnisse des in Abschnitt 2.11 erläuterten Word Intrusion Tests. Die Mean Model Precision, die als Maß für die Kohärenz der „Seeded Topics“ eines Modells dient, ist in Tabelle 4.10 dargestellt. In dieser wurde ebenfalls die ermittelte Mean Model Precision für das Baseline-System, die unüberwachte LDA, angegeben. Hierbei ist zu beachten, dass im Unterschied zu den halbüberwachten Themenmodellen für die LDA alle Themen (d.h. 15, 16 beziehungsweise 17 Themen) von den Annotatoren bewertet wurden. Zusätzlich wurde in Tabelle 4.10 gekennzeichnet, ob nach dem in Abschnitt 3.5.3 beschriebenem Signifikanztest die Ergebnisse der Annotatoren besser waren, als wenn sie den Eindringling zufällig erraten hätten.

Wie aus Tabelle 4.10 hervorgeht, wurde die höchste Mean Model Precision mit einem Wert von 0.9167 durch die Modelle „KCF-8“ und „SCF-8“ erzielt, die beide CluWords basierend auf fastText-Embeddings in die halbüberwachte Themenmodellierung integrierten und acht „Unseeded Topics“ umfassten. Das nächstbeste Modell war „KCF-9“. Die Modelle mit der höchsten Mean Model Precision hatten somit gemeinsam, dass sie auf Pseudodokumenten trainiert wurden, in die CluWords eingefügt wurden, die auf der Ähnlichkeit von fastText-Embeddings beruhten. Die meisten der Modelle „KCF“ und „SCF“, mit Ausnahme von „KCF-7“ und „SCF-7“, übertrafen zudem die Baseline. Dabei ist zu beachten, dass die Word Embeddings auf einem externen Datensatz von Tweets zu verschiedenen Themen trainiert wurden. Die hohen Ergebnisse dieser Modelle deuten somit daraufhin, dass die Integration von Wissen über die semantische Ähnlichkeit von Wörtern im allgemeinen Sprachgebrauch zu Themen führt, die für die menschlichen Annotatoren als interpretierbarer und kohärenter empfunden wurden.

Tabelle 4.10: Mean Model Precision der verschiedenen untersuchten Themenmodelle, gemessen mit dem Word Intrusion Test. Bezüglich der halbüberwachten Themenmodelle wurden nur die „Seeded Topics“ durch die Annotatoren bewertet, während bei der LDA alle Themen evaluiert wurden. „# UT“ steht für die Anzahl an „Unseeded Topics“ und „# GT“ für die Gesamtanzahl an Themen. Höhere Werte bedeuten eine höhere Kohärenz der „Seeded Topics“ des entsprechenden Modells. Die Werte, bei denen die Nullhypothese, dass die Annotatoren den Eindringling zufällig erraten haben, bei einem Signifikanzniveau von $\alpha = 0.05$ verworfen werden konnte, sind mit einem hochgestellten Sternchen markiert. Die besten Resultate sowie die Themenmodelle, die diese hervorbrachten, wurden in fett hervorgehoben.

Modell	**Mean Model Precision**		
# UT/ GT	7/ 15	8/ 16	9/ 17
K	0.7083*	0.625*	0.4583*
KT	0.4583*	0.7083*	0.4166*
KCF	0.6667*	**0.9167***	0.8333*
KCP	0.5833*	0.5833*	0.5000*
S	0.5833*	0.5000*	0.4583*
SCF	0.625*	**0.9167***	0.7083*
SCP	0.4583*	0.5833*	0.2916
LDA	0.7556*	0.5*	0.6078*

Darüber hinaus war auffallend, dass die Mean Model Precision der Modelle „S“ und „SCP“ für alle Werte bezüglich der Anzahl an „Unseeded Topics“ unter den Ergebnissen der Modelle „K“ und „KCP“ lag. Hinsichtlich der Modelle „S“, der Seeded LDA ohne Einbeziehung von CluWords wurde zudem festgestellt, dass diese der einzige Ansatz beziehungsweise Algorithmus war, der für keine Anzahl an „Unseeded Topics“ die Baseline übertreffen konnte. Eine mögliche Ursache ist in dem von Kumar u. a. [356] hervorgehobenen und in Abschnitt 4.2.2 beschriebenen Problem zu sehen, dass Themen von Algorithmen wie der Seeded LDA, die eine asymmetrische a-priori-Verteilung einführen, stark von den Seed Wörtern beeinflusst werden, jedoch teilweise eine geringe Kohärenz aufweisen. In Abschnitt 4.2.2 wurde bereits anhand einiger Beispiele gezeigt, dass die meisten Themen der Modelle „S“ sowie „SCP“ tatsächlich von Seed Wörtern dominiert wurden, die jedoch teilweise intuitiv nicht in Zusammenhang mit den anderen wahrscheinlichsten Begriffen des entsprechenden Themas gebracht werden konnten. Die Tatsache, dass es den Annotatoren schwerer fiel, den Eindringling der Themen der Seeded LDA zu identifizieren als bei anderen Themen deutet ebenfalls daraufhin, dass die Seeded LDA zu inkohärenten Themen führen kann.

Wurden die Ergebnisse des Signifikanztests betrachtet, konnte festgestellt werden, dass mit Ausnahme des Modells „SCP-9“ die Ergebnisse aller Modelle besser waren als eine zufällige Schätzung. Dieses wies die geringste Mean Model Precision von allen untersuchten Modellen mit einem Wert von 0.2916 auf. Das geringe Ergebnis dieses Modells war insofern überraschend, da in Abschnitt 4.3 festgestellt wurde, dass die „Seeded Topics“ von „SCP-9“ im Durchschnitt eine höhere Kohärenz als die anderen untersuchten Themenmodelle mit Ausnahme von „KCP“ aufwiesen. In Abschnitt 4.5 wird genauer auf die Unterschiede zwischen den Resultaten der automatischen Evaluierung und der Nutzerstudie eingegangen.

Hinsichtlich der Themenanzahl zeigte sich analog zu der quantitativen Evaluierung, dass die Modelle unterschiedliche Anzahlen an „Unseeded Topics“ bevorzugten. Keines der Modelle erreichte jedoch im Unterschied zu den in Abschnitt 4.3 aufgeführten Resultaten seine maximale Mean Model Precision bei neun „Unseeded Topics“. Die halbüberwachten Themenmodelle „K“ und „S“, die keine CluWords oder Termgewichtungsschemata integrierten, wiesen die höchste Mean Model Precision bei sieben „Unseeded Topics“ auf. Bei den anderen halbüberwachten Modellen war hingegen jeweils ihre Mean Model Precision höher, wenn sie mit acht „Unseeded Topics“ trainiert wurden. Je nach verwendeten Ansatz hatte zudem die Anzahl der „Unseeded Topics“ eine starke Auswirkung auf die Resultate des Word Intrusion Tests. Dies wird beispielsweise bei Betrachtung der Mean Model Precision des Modells „KT“ deutlich, bei dem es sich um die Version des Algorithmus keyATM handelt, bei der Collapsed Gibbs Sampling um das Termgewichtungsschema basierend auf dem PMI erweitert wurde. Bei sieben und neun „Unseeded Topics“ führte dieses Modell zu den geringsten Werten für die Mean Model Precision, abgesehen von dem Modell „SCP“, während bei acht „Unseeded Topics“ das Modell höhere Resultate als die meisten anderen Themenmodelle ohne Einbeziehung von CluWords basierend auf fastText hervorbrachte. Die Tatsache, dass bereits geringe Änderungen bezüglich der Anzahl der „Unseeded Topics“ sich bei den Ergebnissen des Word Intrusion Tests bemerkbar machen, widerspricht der Annahme von Eshima u. a. [14], die betonten, dass keyATM und seine Varianten robust gegenüber der Anzahl der zusätzlichen Themen ist. Allerdings basierten die Experimente der Autoren auf Gesetzesvorschlagen, bei denen es sich um sprachlich korrekte Dokumente ausreichender Länge handelt. Darüber hinaus sind weitere Experimente mit einer größeren Bandbreite an zusätzlichen Themen für die verschiedenen halbüberwachten Themenmodelle notwendig, um eine Beurteilung über den Grad der Auswirkungen zu ermöglichen.

4.5 Vergleich zwischen quantitativer Evaluierung und Nutzerstudie

Schließlich soll darauf eingegangen werden, inwiefern die automatisch ermittelte semantische Kohärenz mit den Ergebnissen des Word Intrusion Tests übereinstimmt, der darauf abzielt, die menschlich wahrgenommene Themenkohärenz zu messen. Wie in Abschnitt 3.5.3 erläutert wurde, wurde die Korrelation sowohl auf Modellebene als auch auf der Ebene von einzelnen Themen berechnet.

Hinsichtlich der **Korrelation auf Modellebene** konnte festgestellt werden, dass die mittlere semantische Kohärenz der Themenmodelle und ihre Mean Model Precision, die mit dem Intrusion Test ermittelt wurde, nicht korrelieren. Der Kendall-Tau-Korrelationskoeffizient betrug -0.0797 mit einem p-Wert von 0.6252. Bei einem Signifikanzniveau von $\alpha = 0.05$ kann somit die Nullhypothese, dass keine Korrelation zwischen der automatisch ermittelten mittleren semantischen Kohärenz der Themenmodelle und ihrer Kohärenz nach der Nutzerstudie besteht, nicht abgelehnt werden.

Was die **Korrelation auf Themenebene** betrifft, so betrug der Korrelationskoeffizient nach Kendall-Tau zwischen der Kohärenz aller „Seeded Topics“ und dem Prozentsatz der Annotatoren, die das entsprechende „Seeded Topic“ identifizieren konnten, 0.0283. Der Signifikanztest

ergab einen p-Wert von 0,6263, was bei einem Signifikanzniveau von $\alpha = 0.05$ erneut darauf verweist, dass die Beziehung zwischen der automatisch ermittelten Kohärenz und den Ergebnissen der Nutzerstudie nicht statistisch signifikant ist.

Somit konnte mit dem quantitativen Evaluierungsmaß weder eine verlässliche Aussage darüber getroffen werden, welche Themenmodelle nach der menschlichen Bewertung besonders interpretierbar sind, noch wie kohärent die einzelnen Themen wahrgenommen werden. Eine mögliche Ursache ist darin zu sehen, dass, wie in Abschnitt 4.1 erläutert wurde, die semantische Kohärenz auf einer Adaption des PMI basiert [19]. Die grundlegende Idee besteht darin, die Kohärenz von Themen basierend auf dem gemeinsamen Vorkommen der wahrscheinlichsten Wörter in den Dokumenten des Trainingsdatensatzes zu bewerten. Somit tritt analog zu traditionellen Algorithmen der Themenmodellierung (siehe Abschnitt 2.4) das Problem auf, dass Datensätze von kurzen Texten nicht ausreichend Informationen über Wortkookkurrenzen auf Dokumentenebene zulassen. Aus diesem Grund rieten beispielsweise Quan u. a. [157] davon ab, die semantische Kohärenz als Evaluierungsmaß für Themenmodelle, die auf Datensätze kurzer Texte trainiert wurden, zu verwenden. Das beschriebene Probleme wurde in dieser Arbeit bereits dadurch adressiert, dass die semantische Kohärenz nicht basierend auf dem gemeinsamen Auftreten von Wörtern innerhalb der kurzen Texte, sondern innerhalb der Konversationsdokumente berechnet wurde. Jedoch handelt es sich auch bei diesen mit einer durchschnittlichen Länge von 13 Wörtern nach der Vorverarbeitung beispielsweise nach der Definition von Bicalho u. a. [172] und Xun u. a. [361] weiterhin um Kurztexte.

Um weitere mögliche Ursachen für die unterschiedliche Bewertung von Themen nach der automatischen Evaluierung und der Nutzerstudie zu erkennen, wurden Themen näher betrachtet, die von der automatischen Evaluierung fälschlicherweise als kohärent eingestuft wurden. Exemplarisch sind in Tabelle 4.11 vier Themen abgebildet, die eine hohe semantische Kohärenz aufwiesen, bei denen aber höchstens ein Annotator den Eindringling korrekt identifizieren konnte. Genauer gesagt befanden sich diese Themen unter den 20 Themen mit der höchsten semantischen Kohärenz aller 168 Themen aller Modelle. Diese wurden jeweils mit Themenmodellen extrahiert, die CluWords basierend auf paradigmatischen Relationen integrierten.

Wie aus Tabelle 4.11 ersichtlich ist, sind die fünf wahrscheinlichsten Begriffe des Thema „ankommen“ des Modells „KCP-7“ primär türkische Wörter. Die deutschen Übersetzungen dieser Begriffe lassen intuitiv keine Verbindung erkennen. Die anderen beiden Wörter „verantwortlichen“ und „ux“ als Abkürzung für „User Experience“ können intuitiv ebenfalls nicht mit den türkischen Wörtern verbunden werden, weshalb es nahe liegt, dass keiner der drei Annotatoren „Vater“ als Eindringling unter den sechs präsentierten Begriffen identifizieren konnte. Es ist zu beachten, dass für die Berechnung der semantischen Kohärenz im Gegensatz zum Word Intrusion Test die zehn wahrscheinlichsten Wörter einbezogen wurden. Unter diesen befanden sich weitere türkische Begriffe wie „eşini“ (dt. „Ehepartner“) und „soruyor“ (dt. „er fragt“). Eine mögliche Ursache dafür, dass dieses Thema von allen Themen die höchste semantische Kohärenz aufwies, ist darin, zu sehen, dass fremdsprachige Wörter häufig gemeinsam in einem Dokument beziehungsweise Gespräch vorkommen [139, 140]. Da die semantische Kohärenz das gemeinsame Auftreten von Wörtern in Konversationen beurteilt, ist es naheliegend, dass dieses Thema nach der automatischen Evaluierung betrachtet wurde.

Wie jedoch bereits in Abschnitt 2.3.2 beschrieben wurde, müssen Wörter einer Sprache nicht dasselbe Thema beschreiben, was dazu führen kann, dass dieses Thema für die Annotatoren schwer interpretierbar war.

Hinsichtlich des Themas „Anwalt“ des Modells „KCP-8“ fällt auf, dass die Wörter „belangen“ und „verpflichten“ intuitiv miteinander assoziiert werden können. Jedoch verfügen in diesem Thema ebenfalls die beiden sehr allgemeinen Begriffe „lediglich“ und „Typ“ über eine hohe Wahrscheinlichkeit. Das Wort „lediglich“ wurde von zwei Annotatoren und das Wort „Typ“ von einem Annotator fälschlicherweise als Eindringling anstatt „spenden“ ausgewählt, was zeigt, dass diese nicht zu den anderen drei Wörtern passen. Dass das Thema „Anwalt“ dennoch über eine hohe semantische Kohärenz verfügt, könnte auf das bereits in Abschnitt 3.3.2.1 beschriebene und von Roberts u. a. [11] und Nikolenko u. a. [12] hervorgehobene Problem zurückgeführt werden, dass die semantische Kohärenz Themen bevorzugt, in denen allgemeine, hochfrequente Begriffe mit einer hohen Wahrscheinlichkeit auftreten. Darüber hinaus konnte bei Betrachtung der zehn wahrscheinlichsten Begriffe des Themas festgestellt werden, dass unter diesen Wörter wie „Internetangelegenheiten“ und „Vereinsgründer“auftraten. Das Wort „Internetangelegenheiten“ könnte intuitiv mit dem Live-Streaming-Videoportal „Twitch“ in Zusammenhang gebracht werden. Den Begriff „Vereinsgründer“ könnte man ebenfalls mit dem Live-Streaming-Videoportal verbinden, da aus den Chatnachrichten hervorging, dass der Verein SWH Streaming-Events veranstaltete. Wie in Abschnitt 3.3.2.1 erläutert wurde, wird zur Bestimmung der semantischen Kohärenz zwischen je zwei der wahrscheinlichsten Begriffe eines Themas ein Bestätigungsmaß berechnet und anschließend werden die einzelnen Werte summiert [19]. Somit bewertet sie lediglich die Assoziation zwischen jeweils zwei wahrscheinlichen Wörtern und nicht, ob alle zehn Wörter mit der höchsten Wahrscheinlichkeit ein gemeinsames Thema beschreiben. Es muss jedoch an dieser Stelle darauf hingewiesen werden, dass es weiteren Untersuchungen bedarf, um festzustellen, ob ein stärkerer Zusammenhang zwischen den Ergebnissen des Word Intrusion Tests und der semantischen Kohärenz besteht, wenn für die semantische Kohärenz ebenfalls nur die fünf wahrscheinlichsten Wörter berücksichtigt werden.

Ein weitere mögliche Ursache für die Diskrepanz zwischen der semantischen Kohärenz und den Resultaten der Nutzerstudie wird an dem Thema „Schwester“ des Modells „KCP-9“ deutlich. Bei diesem handelt es sich bei den Wörtern „Bruder“, „Mutter“ und „Eltern“ jeweils um Bezeichnungen für Familienangehörige. Hingegen ist „Feek“ Bestandteil der türkischen Phrase „Barakallahu feek“, was auf Deutsch „Möge Allah dich segnen“ bedeutet. Dass dieser Begriff von einem der Annotatoren als Eindringling identifiziert wurde, deutet daraufhin, dass, wie von Lau u. a. [318] vermutet wurde, Annotatoren bei mehreren unpassenden Begriffen dazu neigen, ungewöhnliche Wörter als Eindringling einzustufen. Lau u. a. [318] betonen in diesem Zusammenhang vor allem auch, dass der Word Intrusion Test bereits ein Thema als wenig kohärent betrachtet, wenn ein einzelner Begriff nicht in Verbindung mit den anderen Begriffen steht, selbst wenn die wahrscheinlichsten Wörter des Themas ansonsten kohärent wären. Dementsprechend reagiert er empfindlich auf Ausreißer, die mit dem tatsächlichen Eindringling verwechselt werden, während die semantische Kohärenz die Werte der Bestätigungsmaße zwischen mehreren Wortpaaren zusammenfasst [19]. Dementsprechend kann angenommen werden, dass diese sich weniger empfindlich gegenüber einem einzelnen unpassenden Wort verhält, wenn die Paare aus anderen Wörtern stark miteinander assoziiert werden können. In Bezug auf den männlichen türkischen Vornamen „Canan“, der ebenfalls

von einem Annotator anstelle des Wortes „filmen“ als Eindringling ausgewählt wurde, lässt sich ohne weitere Kontextinformationen für die Annotatoren schwer erkennen, ob der Name zu den anderen Wörtern passt beziehungsweise ob es sich bei „Canan“ um ein Familienmitglied handelt. Dies deutet daraufhin, dass, wie auch von Hoyle u. a. [18] hervorgehoben wurde, nicht nur die automatische Evaluierung der Themenmodellierung mit Problemen verbunden ist, sondern ebenfalls Nutzerstudien wie der Word Intrusion Test Herausforderungen mit sich bringen können. Hierbei weisen Hoyle u. a. [18] daraufhin, dass sich für die Annotatoren die Identifizierung des Eindringlings als schwierig erweist, wenn sie mit den Bedeutungen der Wörter nicht vertraut sind. Diesem Problem wurde zwar bereits dadurch entgegengewirkt, dass den Annotatoren erlaubt wurde, die Wörter mit einer Suchmaschine nachzuschlagen und für ungewöhnliche Begriffe Erklärungen bereitgestellt wurden. Gerade bei forensischen Kommunikationsdaten ist es jedoch schwierig, die Bedeutung bestimmter Begriffe zu erkennen, wenn man nicht umfassende Kenntnisse über den Inhalt der Kurznachrichten oder den spezifischen Fall hat.

Tabelle 4.11: Beispiele für Themen mit einer hohen semantischen Kohärenz und geringen Ergebnissen beim Word Intrusion Test. Dargestellt sind die fünf wahrscheinlichsten Wörter der Themen. Ebenfalls ist das Topic Label des entsprechenden „Seeded Topics“, das Modell, mit dem das Thema extrahiert wurde, und der Eindringling, den die Annotatoren beim Word Intrusion Test identifizieren mussten, angegeben. Für die Themenmodelle wurden die in Tabelle 4.1 erläuterten Bezeichnungen verwendet. SK steht für die semantische Kohärenz des Themas und WI für das Ergebnis beim Word Intrusion Test, konkret für den Prozentanteil der Annotatoren, die den Eindringling des Themas identifizieren konnten. Neben dem Wert für die semantische Kohärenz ist in Klammern der Rang des Themas unter allen 168 Themen aller Modelle nach der semantischen Kohärenz angegeben.

Thema	Modell	Themenwörter	Eindringling	SK	WI
ankommen	KCP-7	1. anaktari (dt. Schlüssel) 2. verantwortlichen 3. ux (User Experience) 4. vermezler (dt. sie werden nicht) 5. aramış (dt. gerufen)	Vater	15.25 (Rang 1)	0
Anwalt	KCP-8	1. Twitch 2. lediglich 3. Typ 4. belangen 5. verpflichten	spenden	-129.78 (Rang 12)	0
Schwester	KCP-9	1. Canan 2. Bruder 3. Mutter 4. Eltern 5. Feek	filmen	-150.91 (Rang 20)	33,33 %

Darüber hinaus sind in Tabelle 4.12 beispielhaft Themen abgebildet, bei denen alle drei Annotatoren den Eindringling erfolgreich identifizieren konnten, die sich jedoch unter den 20 Themen mit der geringsten semantischen Kohärenz befanden. Bei den Beispielen handelt es sich um Themen von Modellen, die CluWords basierend auf fastText-Embeddings integrierten. Wie in Abschnitt 4.4 bereits erläutert wurde, wiesen die beiden Themenmodelle unter Einbeziehung von CluWords basierend auf Word Embeddings die höchste Mean Model Precision auf, was darauf hindeutet, dass diese für die menschlichen Annotatoren besonders gut interpretierbar waren.

Betrachtet man die fünf wahrscheinlichsten Wörter des Themas „ankommen“ des Modells „KCF-9“, stellt man fest, dass diese die Ankunft eines Pakets beschreiben, während das Wort „Blaulicht“ nicht zu diesen passt und dementsprechend von allen Annotatoren korrekt als Eindringling ausgewählt wurde. Die fünf Wörter mit der höchsten Wahrscheinlichkeit in dem Thema „Terror“ würde man intuitiv eher mit dem Thema „Freizeit“ assoziieren, da es sich um Begriffe über das Hören von Musik oder das Spielen von Computerspielen handelt. Erneut erscheint der Eindringling, das Wort „Notruf“ unpassend. Hinsichtlich des Themas „Verein“ fällt der Eindringling „Duschenbuch“ vor allem dadurch auf, dass es sich um ein ungewöhnliches Wort handelt, was ein möglicher Grund dafür sein könnte, dass alle Annotatoren ihn als Ausreißer betrachteten.

Die Tatsache, dass diese Themen eine geringe semantische Kohärenz aufwiesen, obwohl bei diesen im Word Intrusion Test gute Ergebnisse erzielt wurden, könnte darauf zurückgeführt werden, dass die semantische Kohärenz im Gegensatz zu dem Word Intrusion Test die zehn wahrscheinlichsten Wörter berücksichtigt. Jedoch handelte es sich bei den weiteren Wörtern unter den zehn Begriffen mit der höchsten Wahrscheinlichkeit in dem Thema weiterhin um Begriffe, die intuitiv mit der den wahrscheinlichsten fünf Wörtern zusammenpassen wie beispielsweise „versenden“ und „auspacken“ bei dem Thema „ankommen“ sowie „Video“ und „Zocken“ bei dem Thema „Terror“, das eher als Thema „Freizeit“ betrachtet werden konnte. Wahrscheinlicher ist es, dass die geringe semantische Kohärenz damit begründet werden kann, dass diese das gemeinsame Auftreten der wahrscheinlichsten Wörter eines Themas in den ursprünglichen Konversationsdokumenten beurteilt (siehe Abschnitt 3.3.2.1). Jedoch bestand das Ziel bei der Integration der CluWords basierend auf den fastText-Embeddings gerade darin, in die Dokumente weitere Wörter einzufügen, die normalerweise nicht in diesen auftreten. Somit beruhen die entstandenen Themen nicht nur auf Kookkurrenzen von Wörtern auf Dokumentenebene, wodurch anzunehmen ist, dass die wahrscheinlichsten Wörter seltener als bei anderen Sätzen gemeinsam in Konversationsdokumenten erscheinen, was somit eine geringere semantische Kohärenz bewirken könnte. Es bedarf allerdings weiterer Untersuchungen, um festzustellen, ob es sich hierbei lediglich um Einzelfälle handelt oder ob die semantische Kohärenz generell ungeeignet für Themen unter Integration von CluWords ist, die auf externer Ähnlichkeit basieren. Es stellt sich außerdem die Frage, inwieweit die CluWords basierend auf paradigmatischen Relationen ebenfalls von diesem Problem betroffen sind.

Tabelle 4.12: Beispiele für Themen mit guten Ergebnissen beim Word Intrusion Test und einer geringen semantischen Kohärenz. Aufgeführt sind die fünf wahrscheinlichsten Wörter dieser Themen. Bei diesen gelang es 100 % der Annotatoren den korrekten Eindringling zu identifizieren. Ebenfalls ist das Topic Label des entsprechenden „Seeded Topics“, das Modell, mit dem das Thema gelernt wurde, und der Eindringling, den die Annotatoren beim Word Intrusion Test bestimmen mussten, angegeben. Für die Themenmodelle wurden die in Tabelle 4.1 erläuterten Bezeichnungen verwendet. SK steht für die semantische Kohärenz des Themas, wobei ebenfalls in Klammern der Rang des Themas unter allen 168 Themen aller Modelle nach der semantischen Kohärenz angegeben wurde.

Thema	Modell	Themenwörter	Eindringling	SK
ankommen	KCF-9	1. bestellen 2. Paket 3. abholen 4. warten 5. erreichen	Blaulicht	-239.98 (Rang 164)
Terror	KCF-9	1. Song 2. Lied 3. spielen 4. Game 5. testen	Notruf	-236.17 (Rang 161)
Verein	SCF-7	1. **privat** 2. **Screenshot** 3. **Club** 4. **Steffi** 5. Fuba	Duschenbuch	-233.24 (Rang 156)

Werden die wahrscheinlichsten Wörter des Themas „Terror“ in Tabelle 4.12 betrachtet, ist zudem ein Problem des Word Intrusion Tests bei Ansätzen des Seed-Guided Topic Modelings erkennbar: Zwar konnte von allen Annotatoren der Eindringling korrekt identifiziert werden, jedoch beschreiben die Wörter nicht das gewünschte Thema und stehen nicht mit dem Topic Label „Terror“ in Verbindung. Dies liegt daran, dass der Word Intrusion Test nur darauf abzielt, zu beurteilen, wie kohärent die wahrscheinlichsten Wörter des Themas sind, jedoch keine Informationen über die Beziehung zu Seed Wörtern bei der halbüberwachten Themenmodellierung mit einbezieht [14, 315]. Ein Lösungsansatz würde darin bestehen, zusätzlich zu dem gewöhnlichen Word Intrusion Test die Random 4 Word Set Intrusion (R4WSI) heranzuziehen, der ursprünglich von Ying u. a. [116] vorgeschlagen wurde und von für Eshima u. a. [14] für den Vergleich von halbüberwachten Algorithmen der Themenmodellierung adaptiert wurde. Die grundlegende Idee besteht darin, aus vier Wortmengen eine Wortmenge als Eindringling auszuwählen, die am wenigsten zu dem präsentierten Topic Label des „Seeded Topics“ passt. Jedoch ist dieser Test nicht geeignet, um die halbüberwachten Themenmodelle mit einem unüberwachten Baseline-System wie der LDA zu vergleichen. Darüber hinaus wurden bisher

häufig angewendete Kohärenzmaße [z.B. 19, 150, 308] nicht um die Einbeziehung des Topic Labels erweitert, weshalb die R4WSI für den Vergleich mit der automatischen Evaluierung nicht geeignet ist.

5 Zusammenfassung

Die umfassende Menge an Kurznachrichten, die auf mobilen Endgeräten gespeichert sind und im Rahmen von strafrechtlichen Ermittlungen ausgewertet werden müssen, erweist sich als zunehmend herausfordernd. Eine kompakte Zusammenfassung der zahlreichen Nachrichten würde dem Ermittler helfen, sich einen Überblick zu verschaffen. Eine Möglichkeit, eine solche Zusammenfassung automatisiert zu erhalten, bietet die Themenmodellierung. Für die Auswertung von Kommunikationsdaten im forensischen Kontext wurde sie bisher jedoch kaum eingesetzt, was darauf zurückzuführen ist, dass forensische Kommunikationsdaten besondere Herausforderungen an die Themenmodellierung stellen.

Ein wesentliches Ziel dieser Arbeit bestand darin, die verschiedenen Herausforderungen zu identifizieren und Lösungsstrategien aufzuzeigen. Zu diesem Zweck wurde eine umfassende systematische Literaturrecherche durchgeführt. Konkret wurden folgende Herausforderungen für die Themenextraktion aus forensischen Kommunikationsdaten beleuchtet:

- die geringe Länge der forensischen Kurznachrichten
- ihre mangelhafte sprachliche Qualität
- die Tatsache, dass die Themen durch den Kontext der Nachrichten wie der Zeitpunkt, zu dem eine Nachricht verschickt wurde, beeinflusst werden
- die hohe Variabilität im Vokabular von forensischen Kommunikationsdaten
- das Interesse des Ermittlers an bestimmten, oftmals seltenen Themen, die durch traditionelle Algorithmen nicht zwangsläufig identifiziert werden können
- die Tatsache, dass gerade bei organisierter Kriminalität und Bandenkriminalität die fallrelevanten Informationen auf den Mobilfunkgeräten mehrerer verdächtiger Personen verteilt sein können, was zu segmentierter Information führt
- die Mehrsprachigkeit der Datensätze von forensischen Kurznachrichten, gerade bei grenzüberschreitender Kriminalität

Es lassen sich verschiedene Oberkategorien von Themenmodellen unterscheiden, welche mit diesen Herausforderungen unterschiedlich gut umgehen können. Beispielsweise sind Ansätze des Fuzzy Clusterings, graphenbasierte Ansätze zur Themenmodellierung sowie das Exemplar-based Topic Modelling besonders geeignet für kurze und verrauschte Texte. Jedoch können durch diese nicht die weiteren Herausforderungen, wie beispielsweise die Bedeutung des Kontextes oder die Erwartungshaltung des Ermittlers an die Themen adressiert werden. Für diese kommen wiederum Erweiterungen traditioneller algebraischer und probabilistischer Ansätze sowie auf Deep Learning basierende Algorithmen zur Themenmodellierung infrage. So wurden beispielsweise externe Wissensquellen in diese Algorithmen integriert, um einer hohen Vielfalt des Vokabulars von Datensätzen zu begegnen. Wörterbücher wurden in den Prozess der Themenmodellierung einbezogen, um die Schwierigkeit der Mehrsprachigkeit zu überwinden. Kein Ansatz beziehungsweise Algorithmus konnte jedoch alle Probleme in Bezug auf die Themenmodellierung in forensischen Kommunikationsdaten lösen.

Im praktischen Teil dieser Arbeit lag der Fokus auf der Herausforderung, dass der Ermittler oftmals besonders an bestimmten, fallrelevanten Themen interessiert ist, die er bereits in dem Datensatz vermutet und für die er Indizien finden möchte. Hierfür wurden zwei halbüberwachte Ansätze der Themenmodellierung, das keyATM und die Seeded LDA, auf realen Falldaten untersucht und mit der unüberwachten LDA verglichen. Diese benötigten als Vorwissen lediglich einige charakteristische Begriffe zu den gewünschten Themen. Darüber hinaus wurden Kombinationen der halbüberwachten Algorithmen mit dem CluWords-Ansatz vorgestellt, der die Eingabedokumente für die Themenmodellierung um Informationen über die semantische Wortähnlichkeit anreichert. Zur Bestimmung der semantischen Ähnlichkeit wurden Word Embeddings und paradigmatische Relationen miteinander verglichen.

Die Evaluierung der verschiedenen Ansätze erfolgte zunächst qualitativ, indem Beispiele für Themen der verschiedenen untersuchten Modelle präsentiert und interpretiert wurden. Um zudem objektiv bewerten zu können, welcher Ansatz am besten für die forensischen Kommunikationsdaten geeignet ist, wurden die Themen mithilfe eines quantitativen, automatischen Evaluierungsmaßes, der semantischen Kohärenz, bewertet. Ein Problem automatischer Evaluierungsmaße besteht jedoch darin, dass sie je nach verwendetem Datensatz nicht notwendigerweise die tatsächliche menschliche Interpretierbarkeit der Themen wiedergeben. Daher verfolgte diese Arbeit ebenfalls das Ziel, die automatisch ermittelte Themenqualität mit den Ergebnissen einer Nutzerstudie, dem sogenannten Word Intrusion Test, zu vergleichen und die Korrelation zwischen den Ergebnissen zu ermitteln.

Es konnte festgestellt werden, dass nach der automatischen quantitativen Evaluierung das Themenmodell basierend auf dem keyATM-Algorithmus unter Einbeziehung von CluWords basierend auf paradigmatischen Relationen als am besten bewertet werden kann. Dieses führte zur höchsten mittleren semantischen Kohärenz der Themen. Zudem wurde gezeigt, dass die beiden halbüberwachten Algorithmen, keyATM und die Seeded LDA, nicht die Baseline übertreffen konnten, wenn keine CluWords integriert wurden. Nach der Nutzerstudie wurden die besten Resultate durch die beiden halbüberwachten Algorithmen bei Einbeziehung von CluWords basierend auf Word Embeddings erzielt. Die CluWords, welche auf paradigmatischen Relationen basierten, führten hingegen vor allem bei der Seeded LDA je nach verwendeter Themenanzahl im Vergleich zu den anderen Themenmodellen teilweise zu deutlich schlechteren Ergebnissen.

In Bezug auf die Frage, ob die automatisch ermittelte Themenkohärenz die tatsächliche Qualität der Themen widerspiegelt, zeigte sich, dass nach dem Kendall-Rangkorrelationskoeffizienten keine Korrelation zwischen der automatisch ermittelten semantischen Kohärenz und den Ergebnissen der Nutzerstudie bestand. Dies konnte unter anderem darauf zurückgeführt werden, dass die automatisch ermittelte semantische Kohärenz durch die geringe Länge der Kurznachrichten beeinträchtigt wird.

6 Ausblick

Diese Arbeit hat bereits mehrere Ansätze untersucht, um eine der zentralen Herausforderungen bei der Themenmodellierung in forensischen Kommunikationsdaten, das Interesse des Ermittlers an bestimmten, oft seltenen Themen, zu adressieren. Dennoch besteht weiteres Verbesserungspotential.

Insbesondere bei der qualitativen Evaluierung konnte festgestellt werden, dass die Themen der bisherigen Ansätze teilweise von hochfrequenten, irrelevanten Begriffen dominiert wurden. Daher bietet sich ein Vergleich der bisherigen Ergebnisse mit dem halbüberwachten Themenmodell GTM [130] an, dass speziell die Herausforderung der zahlreichen irrelevanten Wörter in umgangssprachlichen Texten adressiert. Hierbei ist geplant, zu untersuchen, ob dieses Modell eine Verbesserung gegenüber der in dieser Arbeit bereits verwendeten Version von dem halbüberwachten Algorithmus keyATM hervorbringen kann, die ebenfalls darauf abzielt, die Dominanz irrelevanter Wörter durch die Integration eines Termgewichtungsschemas zu verhindern. Eine weitere denkbare Möglichkeit zur Bewältigung dieses Problems besteht in der Verwendung des Targeted Topic Modellings, das sich ebenfalls zur Identifizierung von Themen eignet, die den Erwartungen des Ermittlers entsprechen. Im Unterschied zu den hier untersuchten Algorithmen werden jedoch irrelevante Dokumente oder Wörter des Datensatzes für die Themenmodellierung vollständig ausgeschlossen.

Die Ergebnisse der Nutzerstudie zeigten, dass halbüberwachte Ansätze der Themenmodellierung durch die Integration der Ähnlichkeit nach Word Embeddings verbessert werden können. Wünschenswert wäre ein Vergleich dieser Ansätze mit halbüberwachten Themenmodellen, die ausschließlich auf der Word-Embedding-Ähnlichkeit beruhen. Es bedarf jedoch weiterer Untersuchungen, um festzustellen, ob mithilfe von Word Embeddings, die auf einem externen Datensatz trainiert wurden, tatsächlich Themen gelernt werden können, die von hoher Relevanz für einen spezifischen Fall sind.

Darüber hinaus wurde in dieser Arbeit der Datensatz auf monolinguale Nachrichten beschränkt, da bisherige halbüberwachte Algorithmen nicht für mehrsprachige Datensätze geeignet sind. Dementsprechend gilt es für die weitere Forschung, die halbüberwachten Themenmodelle mit den beschriebenen Strategien zum Umgang mit mehrsprachigen Datensätzen zu kombinieren.

Zudem wurde festgestellt, dass für die einzelnen untersuchten Modelle sowohl nach der quantitativen Evaluierung als auch nach der Nutzerstudie jeweils eine unterschiedliche Anzahl an zusätzlichen Themen als optimal betrachtet wurde. Bisher wurden passende Anzahlen an zusätzlichen Themen basierend auf einem einzelnen Algorithmus, der unüberwachten LDA, ermittelt und anschießend nur drei verschiedene Werte von Themenanzahlen für die halbüberwachten Themenmodelle untersucht. Dementsprechend empfiehlt es sich, für jedes einzelne halbüberwachte Themenmodell die Bestimmung der optimalen Anzahl von Themen separat zu wiederholen.

Schließlich zeigten erste Experimente, dass die Resultate der semantischen Kohärenz als häufig eingesetztes Evaluierungsmaß auf dem untersuchten Datensatz nicht mit den Ergebnissen der Nutzerstudie korreliert waren. Um abschließend bewerten zu können, inwieweit dieses Maß zur Evaluierung von Themenmodellen in der Domäne der forensischen Kommunikationsdaten geeignet ist, ist eine umfassendere Untersuchung mit verschiedenen forensischen Datensätzen und einer größeren Anzahl menschlicher Annotatoren erforderlich. Es sollte auch untersucht werden, inwieweit die Anzahl der wahrscheinlichsten Wörter der Themen, die in die automatische Evaluierung sowie in die Nutzerstudie einbezogen werden, eine Auswirkung auf die Korrelation zwischen den Ergebnissen hat. Zudem ist es notwendig, weitere automatische Evaluierungsmaße bei dem Vergleich mit der Nutzerstudie zu berücksichtigen, um zu ermitteln, ob es Alternativen gibt, die die tatsächliche Interpretierbarkeit der Themen besser widerspiegeln.

Literaturverzeichnis

[1] N. Paulsen und S. Klöß, *Corona-Jahr 2021: 300 Milliarden Kurznachrichten in Deutschland*, https://www.bitkom.org/Presse/Presseinformation/Corona-Jahr-2021-300-Milliarden-Kurznachrichten-in-Deutschland, Apr. 2021. (besucht am 20. 10. 2023).

[2] M. Spranger, J. Xi, L. Jaeckel, J. Felser und D. Labudde, „MoNA: A Forensic Analysis Platform for Mobile Communication“, *Künstliche Intelligenz*, Jg. 36, Nr. 2, S. 163–169, Mai 2022, ISSN: 1610-1987. DOI: 10.1007/s13218-022-00762-w.

[3] M. Spranger, F. Heinke, L. Appelt, M. Puder und D. Labudde, „MoNA: Automated Identification of Evidence in Forensic Short Messages“, *International Journal on Advances in Security*, Jg. 9, Nr. 1 & 2, S. 14–24, Aug. 2016, ISSN: 1942-2636.

[4] D. M. Blei, A. Y. Ng und M. I. Jordan, „Latent Dirichlet Allocation“, *The Journal of Machine Learning Research*, Jg. 3, S. 993–1022, Mai 2003, ISSN: 1532-4435.

[5] C. Zhai und S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining* (ACM Books 12), 1. Aufl. San Rafael, Kalifornien, USA: Morgan & Claypool, Juni 2016, ISBN: 978-1-970001-17-4.

[6] S. Wang, Z. Chen, G. Fei, B. Liu und S. Emery, „Targeted Topic Modeling for Focused Analysis“, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York City, New York, USA: Association for Computing Machinery, Aug. 2016, S. 1235–1244, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939743.

[7] V. Rakesh, W. Ding, A. Ahuja, N. Rao, Y. Sun und C. K. Reddy, „A Sparse Topic Model for Extracting Aspect-Specific Summaries from Online Reviews“, in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, Ser. Track: Web Search and Mining, Lyon, Frankreich: Association for Computing Machinery, Apr. 2018, S. 1573–1582, ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186069.

[8] T.-C. Nguyen, T.-N. Pham, H.-Q. Le, T.-T. Nguyen, H.-N. Bui und Q.-T. Ha, „A Targeted Topic Model based Multi-Label Deep Learning Classification Framework for Aspect-based Opinion Mining“, in *Proceedings of the 12th International Conference on Knowledge and Systems Engineering*, Can Tho, Vietnam: IEEE Xplore, Nov. 2020, S. 165–170, ISBN: 978-1-72814-510-5. DOI: 10.1109/KSE50997.2020.9287397.

[9] Y. Tsou, D.-N. Chen und C.-Y. Lai, „A Cross-lingual Patent Topics Model for Trend Analysis“, in *Proceedings of the International Computer Symposium*, Tainan, Taiwan: IEEE, Dez. 2020, S. 525–528, ISBN: 978-1-72819-255-0. DOI: 10.1109/ICS51289.2020.00108.

[10] A. Karami, A. Gangopadhyay, B. Zhou und H. Kharrazi, „FLATM: A Fuzzy Logic Approach Topic Model for Medical Documents“, in *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) Held Jointly with 2015 5th World Conference on Soft Computing*, Redmond, Washington, USA: IEEE, Aug. 2015, S. 1–6, ISBN: 978-1-4673-7248-0. DOI: 10.1109/NAFIPS-WConSC.2015.7284190.

[11] M. E. Roberts u. a., „Structural Topic Models for Open-Ended Survey Responses“, *American Journal of Political Science*, Jg. 58, Nr. 4, S. 1064–1082, März 2014, ISSN: 0092-5853. DOI: 10.1111/ajps.12103.

[12] S. I. Nikolenko, S. Koltcov und O. Koltsova, „Topic Modelling for Qualitative Studies“, *Journal of Information Science*, Jg. 43, Nr. 1, S. 88–102, Feb. 2017, ISSN: 0165-5515. DOI: 10.1177/0165551515617393.

[13] S.-H. Kim, N. Lee und P. E. King, „Dimensions of Religion and Spirituality: A Longitudinal Topic Modeling Approach“, *Journal for the Scientific Study of Religion*, Jg. 59, Nr. 1, S. 62–83, Jan. 2020. DOI: 10.1111/jssr.12639.

[14] S. Eshima, K. Imai und T. Sasaki, „Keyword-Assisted Topic Models“, *American Journal of Political Science*, Jg. 0, Nr. 0, S. 1–21, Apr. 2023, ISSN: 1540-5907. DOI: 10.1111/ajps.12779.

[15] K. Watanabe und A. Baturo, „Seeded Sequential LDA: A Semi-Supervised Algorithm for Topic-Specific Analysis of Sentences“, *Social Science Computer Review*, Jg. 0, Nr. 0, S. 1–25, Mai 2023. DOI: 10.1177/08944393231178605.

[16] F. Viegas u. a., „CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling“, in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, Ser. Session 12: Text Understanding, Melbourne, Australien: Association for Computing Machinery, Jan. 2019, S. 753–761, ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3291032.

[17] C. Doogan und W. Buntine, „Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures“, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Juni 2021, S. 3824–3848, ISBN: 978-1-954085-46-6. DOI: 10.18653/v1/2021.naacl-main.300.

[18] A. Hoyle, P. Goel, D. Peskov, A. Hian-Cheong, J. Boyd-Graber und P. Resnik, „Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence“, in *Proceedings of 35th Conference on Neural Information Processing Systems*, Bd. 34, Online: Curran Associates, Dez. 2021, S. 1–16.

[19] D. Mimno, H. Wallach, E. Talley, M. Leenders und A. McCallum, „Optimizing Semantic Coherence in Topic Models“, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Schottland, Großbritannien: Association for Computational Linguistics, Juli 2011, S. 262–272.

[20] S. Likhitha, B. S. Harish und H. M. K. Kumar, „A Detailed Survey on Topic Modeling for Document and Short Text Data“, *International Journal of Computer Applications*, Jg. 178, Nr. 39, S. 1–9, Aug. 2019. DOI: 10.5120/ijca2019919265.

[21] R. Alghamdi und K. Alfalqi, „A Survey of Topic Modeling in Text Mining“, *International Journal of Advanced Computer Science and Applications*, Jg. 6, Nr. 1, S. 147–153, 2015, ISSN: 2158107X. DOI: 10.14569/IJACSA.2015.060121.

[22] J. Qiang, Z. Qian, Y. Li, Y. Yuan und X. Wu, „Short Text Topic Modeling Techniques, Applications, and Performance: A Survey“, *IEEE Transactions on Knowledge and Data Engineering*, Jg. 34, Nr. 3, S. 1427–1445, März 2022, ISSN: 1558-2191. DOI: 10.1109/TKDE.2020.2992485.

[23] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki und H. M. Abdulwahab, „Short Text Topic Modelling Approaches in the Context of Big Data: Taxonomy, Survey, and Analysis“, *Artificial Intelligence Review*, Jg. 56, S. 5133–5260, Okt. 2022, ISSN: 1573-7462. DOI: 10.1007/s10462-022-10254-w.

[24] R. Churchill und L. Singh, „The Evolution of Topic Modeling“, *ACM Computing Surveys*, Jg. 54, Nr. 10, S. 1–35, Nov. 2022, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3507900.

[25] N. Döring, „Forschungsstand und theoretischer Hintergrund“, in *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, 6. Aufl., Berlin, Deutschland: Springer, 2022, S. 161–181, ISBN: 978-3-662-64762-2. DOI: 10.1007/978-3-662-64762-2_6.

[26] M. J. Bates, „Where Should the Person Stop and the Information Search Interface Start?“, *Information Processing & Management*, Jg. 26, Nr. 5, S. 575–591, Jan. 1990, ISSN: 03064573. DOI: 10.1016/0306-4573(90)90103-9.

[27] J. Jagarlamudi, H. Daumé III und R. Udupa, „Incorporating Lexical Priors into Topic Models“, in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, Frankreich: Association for Computational Linguistics, Apr. 2012, S. 204–213.

[28] D. Ramage, D. Hall, R. Nallapati und C. D. Manning, „Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora“, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapur, Singapur: Association for Computational Linguistics, Aug. 2009, S. 248–256.

[29] J. Tang, M. Zhang und Q. Mei, „One Theme in All Views: Modeling Consensus Topics in Multiple Contexts“, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA: Association for Computing Machinery, Aug. 2013, S. 5–13, ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487682.

[30] C. Wang, H. Zhang, B. Chen, D. Wang, Z. Wang und M. Zhou, „Deep Relational Topic Modeling via Graph Poisson Gamma Belief Network“, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Bd. 42, Vancouver, Kanada: Curran Associates, Dez. 2020, S. 488–500.

[31] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer und R. Harshman, „Indexing by Latent Semantic Analysis“, *Journal of the American Society for Information Science*, Jg. 41, Nr. 6, S. 391–407, Sep. 1990, ISSN: 0002-8231. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[32] P. Paatero, „Least Squares Formulation of Robust Non-Negative Factor Analysis“, *Chemometrics and Intelligent Laboratory Systems*, Jg. 37, Nr. 1, S. 23–35, Mai 1997, ISSN: 0169-7439. DOI: 10.1016/S0169-7439(96)00044-5.

[33] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat und A. Hassan Yousef, „Topic Modeling Algorithms and Applications: A Survey“, *Information Systems*, Jg. 112, Nr. C, S. 1–33, Dez. 2022. DOI: 10.1016/j.is.2022.102131.

[34] F. Shahnaz, M. W. Berry, V. Pauca und R. J. Plemmons, „Document Clustering Using Nonnegative Matrix Factorization“, *Information Processing & Management*, Jg. 42, Nr. 2, S. 373–386, März 2006, ISSN: 03064573. DOI: 10.1016/j.ipm.2004.11.005.

[35] X. Chen, Y. Qi, B. Bai, Q. Lin und J. G. Carbonell, „Sparse Latent Semantic Analysis“, in *Proceedings of the 2011 SIAM International Conference on Data Mining*, Phoenix, Arizona, USA: Society for Industrial and Applied Mathematics, Apr. 2011, S. 474–485, ISBN: 978-0-89871-992-5 978-1-61197-281-8. DOI: 10.1137/1.9781611972818.41.

[36] J. Choo, C. Lee, C. K. Reddy und H. Park, „UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization“, *IEEE Transactions on Visualization and Computer Graphics*, Jg. 19, Nr. 12, S. 1992–2001, Dez. 2013, ISSN: 1941-0506. DOI: 10.1109/TVCG.2013.212.

[37] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi und F. A. Ghanem, „Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling“, *IEEE Access*, Jg. 10, S. 105 328–105 351, Okt. 2022, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3211396.

[38] Y. Chen, H. Zhang, R. Liu, Z. Ye und J. Lin, „Experimental Explorations on Short Text Topic Mining Between Lda and Nmf Based Schemes“, *Knowledge-Based Systems*, Jg. 163, S. 1–13, Jan. 2019, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2018.08.011.

[39] R. Albalawi, T. H. Yeap und M. Benyoucef, „Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis“, *Frontiers in Artificial Intelligence*, Jg. 3, Nr. 42, S. 1–14, Juli 2020, ISSN: 2624-8212. DOI: 10.3389/frai.2020.00042.

[40] N. Harada, K. Yamashita, Y. Motomura und Y. Kano, „Applying Statistical Approach to Topic Analysis for more Comprehensive and Appropriate Modeling“, in *Proceedings of 6th Conference on Data Science and Machine Learning Applications*, Riad, Saudi-Arabien: IEEE, März 2020, S. 13–18, ISBN: 978-1-72812-746-0. DOI: 10.1109/CDMA47397.2020.00008.

[41] T. Hofmann, „Probabilistic Latent Semantic Indexing“, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ser. SIGIR '99, Berkeley, Kalifornien, USA: Association for Computing Machinery, Aug. 1999, S. 50–57, ISBN: 1-58113-096-1. DOI: 10.1145/312624.312649.

[42] Z. Abbasi, S. Latif, F. Shafait und R. Latif, „Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling“, *IEEE Access*, Jg. 9, S. 127 531–127 547, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3112620.

[43] M. Zhou, Y. Kong und J. Lin, „Financial Topic Modeling Based on the BERT-LDA Embedding“, in *Proceedings of the 20th International Conference on Industrial Informatics*, Perth, Australien: IEEE, Juli 2022, S. 495–500, ISBN: 978-1-72817-568-3. DOI: 10.1109/INDIN51773.2022.9976145.

[44] T. Shi, K. Kang, J. Choo und C. K. Reddy, „Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations“, in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, Lyon, Frankreich: ACM Press, Apr. 2018, S. 1105–1114, ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186009.

[45] L. Hong und B. D. Davison, „Empirical study of topic modeling in Twitter“, in *Proceedings of the First Workshop on Social Media Analytics*, Washington D.C., USA: Association for Computing Machinery, Juli 2010, S. 80–88, ISBN: 978-1-4503-0217-3. DOI: 10.1145/1964858.1964870.

[46] A. Srivastava und C. Sutton, „Autoencoding Variational Inference For Topic Models“, in *Proceedings of the 5th International Conference on Learning Representations*, Ser. Machine Learning and Data Mining, Toulon, Frankreich: ICLR Press, Apr. 2017, S. 1–12.

[47] T. L. Griffiths und M. Steyvers, „Finding Scientific Topics“, *Proceedings of the National Academy of Sciences*, Jg. 101, Nr. suppl_1, S. 5228–5235, Apr. 2004. DOI: 10.1073/pnas.0307752101.

[48] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du und W. Buntine, „Topic Modelling Meets Deep Neural Networks: A Survey“, in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Montreal, Kanada: AAAI Press, Aug. 2021, S. 4713–4720, ISBN: 978-0-9992411-9-6. DOI: 10.24963/ijcai.2021/638.

[49] Y. Miao, L. Yu und P. Blunsom, „Neural Variational Inference for Text Processing“, in *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan und K. Q. Weinberger, Hrsg., Ser. Proceedings of Machine Learning Research, Bd. 48, New York, New York, USA: Journal of Machine Learning Research, Juni 2016, S. 1727–1736.

[50] B. Zhu, Y. Cai und H. Ren, „Graph Neural Topic Model with Commonsense Knowledge“, *Information Processing & Management*, Jg. 60, Nr. 2, S. 1–13, März 2023, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2022.103215.

[51] R. Wang, D. Zhou und Y. He, „ATM: Adversarial-Neural Topic Model“, *Information Processing & Management*, Jg. 56, Nr. 6, S. 102 098, Nov. 2019, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2019.102098.

[52] S. Lauly, Y. Zheng, A. Allauzen und H. Larochelle, „Document Neural Autoregressive Distribution Estimation“, *The Journal of Machine Learning Research*, Jg. 18, Nr. 1, S. 4046–4069, Jan. 2017, ISSN: 1532-4435.

[53] Y. Ge und X. Hu, „Enhancing Graph Variational Autoencoder for Short Text Topic Modeling with Mutual Information Maximization“, in *Proceedings of the International Conference on Knowledge Graph*, Orlando, Florida, USA: IEEE, Nov. 2022, S. 64–70, ISBN: 978-1-66545-102-4. DOI: 10.1109/ICKG55886.2022.00016.

[54] I. Reyad, M. Rashad und M. Abdelfatah, „A Comparative Study of Topic Modeling Methods for Document Retrieval“, in *Proceedings of 32nd International Conference on Computer Theory and Applications*, Alexandria, Ägypten: IEEE, Dez. 2022, S. 74–79, ISBN: 979-8-3503-2019-0. DOI: 10.1109/ICCTA58027.2022.10206075.

[55] R. Ding, R. Nallapati und B. Xiang, „Coherence-Aware Neural Topic Modeling“, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brüssel, Belgien: Association for Computational Linguistics, Okt. 2018, S. 830–836, ISBN: 978-1-948087-84-1. DOI: 10.18653/v1/D18-1096.

[56] D. T. K. Geeganage, Y. Xu und Y. Li, „Semantic-based Topic Representation Using Frequent Semantic Patterns“, *Knowledge-Based Systems*, Jg. 216, S. 1–10, März 2021, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2021.106808.

[57] C. E. Moody, *Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec*, Mai 2016. DOI: 10.48550/arXiv.1605.02019. arXiv: 1605.02019 [cs].

[58] D. E. Rumelhart, G. E. Hintont und R. J. Williams, „Learning Representations by Back-Propagating Errors“, in *Neurocomputing: Foundations of Research*, Bd. 323, Cambridge, Massachusetts, USA: MIT Press, Jan. 1988, S. 696–699, ISBN: 0-262-01097-6.

[59] S. Lai, K. Liu, S. He und J. Zhao, „How to Generate a Good Word Embedding“, *IEEE Intelligent Systems*, Jg. 31, Nr. 6, S. 5–14, Mai 2016, ISSN: 1941-1294. DOI: 10.1109/MIS.2016.45.

[60] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun und Z. Ma, „Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings“, *ACM Transactions on Information Systems*, Jg. 36, Nr. 2, S. 1–30, Aug. 2017. DOI: 10.1145/3091108.

[61] Q. Le und T. Mikolov, „Distributed Representations of Sentences and Documents“, in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing und T. Jebara, Hrsg., Bd. 32, Peking, China: MIT Press, Juni 2014, S. 1188–1196.

[62] A. B. Dieng, F. J. R. Ruiz und D. M. Blei, „Topic Modeling in Embedding Spaces“, *Transactions of the Association for Computational Linguistics*, Jg. 8, S. 439–453, Juli 2020, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00325.

[63] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado und J. Dean, „Distributed Representations of Words and Phrases and their Compositionality“, in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Bd. 2, Lake Tahoe, Nevada, USA: Curran Associates, Dez. 2013, S. 3111–3119.

[64] M. Grootendorst, *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*, März 2022. arXiv: 2203.05794 [cs].

[65] D. Angelov, *Top2Vec: Distributed Representations of Topics*, Aug. 2020. DOI: 10.48550/arXiv.2008.09470. arXiv: 2008.09470 [cs, stat].

[66] J. Devlin, M.-W. Chang, K. Lee und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Juni 2019, S. 4171–4186. DOI: 10.18653/v1/N19-1423.

[67] L. McInnes, J. Healy und S. Astels, „Hdbscan: Hierarchical Density Based Clustering“, *The Journal of Open Source Software*, Jg. 2, Nr. 11, S. 205–206, März 2017, ISSN: 2475-9066. DOI: 10.21105/joss.00205.

[68] N. Gialitsis, G. Giannakopoulos und M. Athanasouli, „Evaluation of Distributed DNA Representations on the Classification of Conserved Non-Coding Elements“, in *Proceedings of the 11th Hellenic Conference on Artificial Intelligence*, Athen, Griechenland: Association for Computing Machinery, Sep. 2020, S. 41–47, ISBN: 978-1-4503-8878-8. DOI: 10.1145/3411408.3411463.

[69] R. Egger und J. Yu, „A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts“, *Frontiers in Sociology*, Jg. 7, S. 886 498, Mai 2022, ISSN: 2297-7775. DOI: 10.3389/fsoc.2022.886498.

[70] R. K. Srivastava, S. Sharma und P. P. Singh, „Exploring Latent Themes-Analysis of Various Topic Modelling Algorithms“, *International Journal of Advanced Research in Science, Communication and Technology*, Jg. 3, Nr. 5, S. 225–229, Juni 2023, ISSN: 2581-9429. DOI: 10.48175/IJARSCT-11635.

[71] H. F. de Arruda, L. d. F. Costa und D. R. Amancio, „Topic Segmentation Via Community Detection in Complex Networks“, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, Jg. 26, Nr. 6, S. 1–11, Juni 2016, ISSN: 1054-1500. DOI: 10.1063/1.4954215.

[72] A. Hamm und S. Odrowski, „Term-Community-Based Topic Detection with Variable Resolution“, *Information*, Jg. 12, Nr. 6, S. 221–252, Juni 2021, ISSN: 2078-2489. DOI: 10.3390/info12060221.

[73] R. Churchill und L. Singh, „Percolation-Based Topic Modeling for Tweets“, in *Proceedings of the 9th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining*, San Diego, USA: Association for Computing Machinery, Aug. 2020, S. 1–8.

[74] I. Ganguli, J. Sil und N. Sengupta, „Non-parametric Method of Topic Identification Using Granularity Concept and Graph-Based Modeling“, in *Proceedings of the 6th International Conference on Soft Computing & Machine Intelligence*, Johannesburg, Südafrika: IEEE, Nov. 2019, S. 78–82, ISBN: 978-1-72814-577-8. DOI: 10.1109/ISCMI47871.2019.9004380.

[75] W. Wang, H. Zhou, K. He und J. E. Hopcroft, „Learning Latent Topics from the Word Co-occurrence Network“, in *Proceedings of the 35th National Conference of Theoretical Computer Science*, D. Du, L. Li, E. Zhu und K. He, Hrsg., Ser. Communications in Computer and Information Science (CCIS), Bd. 768, Singapur, Singapur: Springer Nature, Okt. 2017, S. 18–30, ISBN: 978-981-10-6893-5. DOI: 10.1007/978-981-10-6893-5_2.

[76] D. Paranyushkin, „InfraNodus: Generating Insight Using Text Network Analysis“, in *Proceedings of the World Wide Web Conference*, New York City, New York, USA: Association for Computing Machinery, Mai 2019, S. 3584–3589, ISBN: 978-1-4503-6674-8. DOI: 10.1145/3308558.3314123.

[77] G. Zhou und G. Chen, „Hierarchical Latent Semantic Mapping for Automated Topic Generation“, in *Proceedings of the 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Shanghai, China: IEEE, Mai 2016, S. 57–63, ISBN: 978-1-5090-2239-7. DOI: 10.1109/SNPD.2016.7515878.

[78] S. Yang, Q. Sun, H. Zhou, Z. Gong, Y. Zhou und J. Huang, „A Topic Detection Method Based on KeyGraph and Community Partition“, in *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, Ser. ICCAI 2018, New York City, New York, USA: Association for Computing Machinery, März 2013, S. 30–34, ISBN: 978-1-4503-6419-5. DOI: 10.1145/3194452.3194474.

[79] G. Tolegen, A. Toleu, R. Mussabayev und A. Krassovitskiy, „A Clustering-based Approach for Topic Modeling via Word Network Analysis“, in *Proceedings of the 7th International Conference on Computer Science and Engineering*, Diyarbakir, Türkei: IEEE, Sep. 2022, S. 192–197, ISBN: 978-1-66547-010-0. DOI: 10.1109/UBMK55850.2022.9919530.

[80] S. Yang, G. Huang und B. Cai, „Discovering Topic Representative Terms for Short Text Clustering“, *IEEE Access*, Jg. 7, S. 92 037–92 047, Juli 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2927345.

[81] A. Lancichinetti, M. I. Sirer, J. X. Wang, D. Acuna, K. Körding und L. A. N. Amaral, „High-Reproducibility and High-Accuracy Method for Automated Topic Classification“, *Physical Review X*, Jg. 5, Nr. 1, S. 1–11, Jan. 2015, ISSN: 2160-3308. DOI: 10.1103/PhysRevX.5.011007.

[82] V. D. Blondel, J.-L. Guillaume, R. Lambiotte und E. Lefebvre, „Fast Unfolding of Communities in Large Networks“, *Journal of Statistical Mechanics: Theory and Experiment*, Jg. 08, Nr. 10, S. 1–13, Okt. 2008. DOI: 10.1088/1742-5468/2008/10/P10008.

[83] G. Amati u. a., „Topic Modeling by Community Detection Algorithms“, in *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, Ser. OASIS '21, New York City, New York, USA: Association for Computing Machinery, Okt. 2021, S. 15–20, ISBN: 978-1-4503-8632-6. DOI: 10.1145/3472720.3483622.

[84] V. Traag, L. Waltman und N. J. van Eck, „From Louvain to Leiden: Guaranteeing Well-Connected Communities“, *Scientific Reports*, Jg. 9, Nr. 1, S. 5233–5245, März 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z.

[85] A. Srivastava, A. J. Soto und E. Milios, „A Graph-Based Topic Extraction Method Enabling Simple Interactive Customization“, in *Proceedings of the ACM Symposium on Document Engineering*, Florenz, Italien: Association for Computing Machinery, Sep. 2013, S. 71–80, ISBN: 978-1-4503-1789-4. DOI: 10.1145/2494266.2494280.

[86] J. Rashid, S. M. A. Shah und A. Irtaza, „Fuzzy Topic Modeling Approach for Text Mining Over Short Text“, *Information Processing & Management*, Jg. 56, Nr. 6, S. 1–19, Nov. 2019, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2019.102060.

[87] J. C. Bezdek, R. Ehrlich und W. Full, „FCM: The Fuzzy C-Means Clustering Algorithm“, *Computers & Geosciences*, Jg. 10, Nr. 2, S. 191–203, Jan. 1984, ISSN: 0098-3004. DOI: 10.1016/0098-3004(84)90020-7.

[88] K. Pearson, „LIII. on Lines and Planes of Closest Fit to Systems of Points in Space“, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Jg. 2, Nr. 11, S. 559–572, Nov. 1901, ISSN: 1941-5982. DOI: 10.1080/14786440109462720.

[89] A. Elbagoury, R. Ibrahim, A. Farahat, M. Kamel und F. Karray, „Exemplar-Based Topic Detection in Twitter Streams“, in *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, Oxford, Großbritannien: AAAI Press, Aug. 2021, S. 610–613, ISBN: 978-1-57735-733-9. DOI: 10.1609/icwsm.v9i1.14651.

[90] R. J. Gallagher, K. Reing, D. Kale und G. Ver Steeg, „Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge“, *Transactions of the Association for Computational Linguistics*, Jg. 5, S. 529–542, Dez. 2017, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00078.

[91] G. Ver Steeg und A. Galstyan, „Discovering Structure in High-Dimensional Data Through Correlation Explanation“, in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Bd. 1, Montreal, Kanada: MIT Press, Dez. 2014, S. 577–585, ISBN: 978-3-030-63835-1.

[92] T. M. Cover und J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications). New York City, New York, USA: John Wiley & Sons, Inc., Juni 1991, ISBN: 978-0-471-06259-2.

[93] G. E. Noel und G. L. Peterson, „Applicability of Latent Dirichlet Allocation to multi-disk search“, *Digital Investigation*, Jg. 11, Nr. 1, S. 43–56, März 2014, ISSN: 17422876. DOI: 10.1016/j.diin.2014.02.001.

[94] A. de Waal, J. Venter und E. Barnard, „Applying Topic Modeling to Forensic Data“, in *Advances in Digital Forensics IV*, Ser. IFIP — The International Federation for Information Processing, 4. Aufl., Bd. 285, New York City, New York, USA: Springer Science+Business Media, Jan. 2008, S. 115–126, ISBN: 978-0-387-84926-3. DOI: 10.1007/978-0-387-84927-0_10.

[95] J. Li, W.-H. Chen, Q. Xu, N. Shah und T. Mackey, „Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology“, in *Proceedings of the Global Humanitarian Technology Conference*, Seattle, WA, USA: IEEE, Okt. 2019, S. 1–4, ISBN: 978-1-72811-780-5. DOI: 10.1109/GHTC46095.2019.9033129.

[96] L. Busso, M. Petyko, S. Atkins und T. Grant, „Operation Heron: Latent Topic Changes in an Abusive Letter Series“, *Corpora*, Jg. 17, Nr. 2, S. 225–258, Aug. 2022, ISSN: 1749-5032, 1755-1676. DOI: 10.3366/cor.2022.0255.

[97] X. Yan, J. Guo, Y. Lan und X. Cheng, „A Biterm Topic Model for Short Texts“, in *Proceedings of the 22nd International Conference on World Wide Web*, Rio de Janeiro, Brazil: Association for Computing Machinery, Mai 2013, S. 1445–1456, ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488514.

[98] M. Petyko, *TextCrimes*, http://fold.aston.ac.uk/handle/123456789/13, Mai 2021. (besucht am 26. 09. 2023).

[99] P. Joseph und P. Viswanathan, „SDOT: Secure Hash, Semantic Keyword Extraction, and Dynamic Operator Pattern-Based Three-Tier Forensic Classification Framework“, *IEEE Access*, Jg. 11, S. 3291–3306, Jan. 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3234434.

[100] J. S. Okolica, G. L. Peterson und R. F. Mills, „Using Author Topic to Detect Insider Threats from Email Traffic“, *Digital Investigation*, Jg. 4, Nr. 3, S. 158–164, Sep. 2007, ISSN: 1742-2876. DOI: 10.1016/j.diin.2007.10.002.

[101] M. Yang, F. Xu und K.-P. Chow, „Interest Profiling for Security Monitoring and Forensic Investigation“, in *Proceedings of the 21st Australasian Conference on Information Security and Privacy*, J. K. Liu und R. Steinfeld, Hrsg., Ser. Lecture Notes in Computer Science, Bd. 9723, Cham, Schweiz: Springer International Publishing, Juni 2016, S. 457–464, ISBN: 978-3-319-40367-0. DOI: 10.1007/978-3-319-40367-0_30.

[102] M. Rosen Zvi, T. Griffiths, M. Steyvers und P. Smyth, „The Author-Topic Model for Authors and Documents“, in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Ser. UAI '04, Arlington, Virginia, USA: AUAI Press, Juli 2004, S. 487–494, ISBN: 978-0-9749039-0-3.

[103] M. Bérubé, T.-U. Tang, F. Fortin, S. Ozalp, M. L. Williams und P. Burnap, „Social Media Forensics Applied to Assessment of Post–Critical Incident Social Reaction: The Case of the 2017 Manchester Arena Terrorist Attack“, *Forensic Science International*, Jg. 313, S. 1–37, Aug. 2020, ISSN: 0379-0738. DOI: 10.1016/j.forsciint.2020.110364.

[104] W. Rule, W. Duan, N. Prakash, N. Zhuang, R. C. Alvarado und D. E. Brown, „Social Pressure Analysis of Local Events Using Social Media Data“, in *Proceedings of the Conference on Systems and Information Engineering Design Symposium*, Charlottesville, Virginia, USA: IEEE, Apr. 2018, S. 277–281, ISBN: 978-1-5386-6343-1. DOI: 10.1109/SIEDS.2018.8374751.

[105] H. Suryotrisongko, H. Ginardi, H. T. Ciptaningtyas, S. Dehqan und Y. Musashi, „Topic Modeling for Cyber Threat Intelligence (CTI)“, in *Proceedings of the Seventh International Conference on Informatics and Computing*, Denpasar, Bali, Indonesien: IEEE, Dez. 2022, S. 1–7, ISBN: 979-8-3503-4572-8. DOI: 10.1109/ICIC56845.2022.10006988.

[106] I. Deliu, C. Leichter und K. Franke, „Collecting Cyber Threat Intelligence from Hacker Forums via a Two-Stage, Hybrid Process using Support Vector Machines and Latent Dirichlet Allocation“, in *Proceedings of the International Conference on Big Data*, Seattle, Washington, USA: IEEE, Dez. 2018, S. 5008–5013, ISBN: 978-1-5386-5035-6. DOI: 10.1109/BigData.2018.8622469.

[107] S. Samtani, K. Chinn, C. Larson und H. Chen, „AZSecure Hacker Assets Portal: Cyber Threat Intelligence and Malware Analysis“, in *Proceedings of the IEEE Conference on Intelligence and Security Informatics*, Tucson, Arizona, USA: IEEE, Sep. 2016, S. 19–24, ISBN: 978-1-5090-3865-7. DOI: 10.1109/ISI.2016.7745437.

[108] K. Porter, „Analyzing the DarkNetMarkets Subreddit for Evolutions of Tools and Trends Using Lda Topic Modeling“, *Digital Investigation*, Jg. 26, S. 87–97, Juli 2018, ISSN: 1742-2876. DOI: 10.1016/j.diin.2018.04.023.

[109] Z. Fang u. a., „Exploring Key Hackers and Cybersecurity Threats in Chinese Hacker Communities“, in *Proceedings of the IEEE Conference on Intelligence and Security Informatics*, Tucson, Arizona, USA: IEEE, Sep. 2016, S. 13–18, ISBN: 978-1-5090-3865-7. DOI: 10.1109/ISI.2016.7745436.

[110] T. Vahedi, B. Ampel, S. Samtani und H. Chen, „Identifying and Categorizing Malicious Content on Paste Sites: A Neural Topic Modeling Approach“, in *Proceedings of the International Conference on Intelligence and Security Informatics*, San Antonio, Texas, USA: IEEE, Nov. 2021, S. 1–6, ISBN: 978-1-66543-838-4. DOI: 10.1109/ISI53945.2021.9624765.

[111] D. Kuang, J. Brantingham und A. L. Bertozzi, „Crime Topic Modeling“, *Crime Science*, Jg. 6, Nr. 12, S. 1–20, Dez. 2017, ISSN: 2193-7680. DOI: 10.1186/s40163-017-0074-0.

[112] R. Pandey und G. O. Mohler, „Evaluation of Crime Topic Models: Topic Coherence vs Spatial Crime Concentration“, in *Proceedings of the International Conference on Intelligence and Security Informatics*, Miami, Florida, USA: IEEE, Nov. 2018, S. 76–78, ISBN: 978-1-5386-7848-0. DOI: 10.1109/ISI.2018.8587384.

[113] X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Nguyen, S. Horiguchi und Q.-T. Ha, „A Hidden Topic-Based Framework toward Building Applications with Short Web Documents“, *IEEE Transactions on Knowledge and Data Engineering*, Jg. 23, Nr. 7, S. 961–976, Juli 2011, ISSN: 1558-2191. DOI: 10.1109/TKDE.2010.27.

[114] D. Y. Sylfania, F. P. Juniawan, L. Laurentinus und H. A. Pradana, „SMS Security Improvement using RSA in Complaints Application on Regional Head Election's Fraud“, *Jurnal Teknologi dan Sistem Komputer*, Jg. 7, Nr. 3, S. 116–120, Juli 2019, ISSN: 2338-0403. DOI: 10.14710/jtsiskom.7.3.2019.116-120.

[115] K. Boczek und L. Koppers, „What's New about Whatsapp for News? A Mixed-Method Study on News Outlets' Strategies for Using WhatsApp“, *Digital Journalism*, Jg. 8, Nr. 1, S. 126–144, Jan. 2020, ISSN: 2167-0811, 2167-082X. DOI: 10.1080/21670811.2019.1692685.

[116] L. Ying, J. M. Montgomery und B. M. Stewart, „Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures“, *Political Analysis*, Jg. 30, Nr. 4, S. 570–589, Okt. 2022, ISSN: 1047-1987. DOI: 10.1017/pan.2021.33.

[117] I. Vayansky und S. A. P. Kumar, „A Review of Topic Modeling Methods“, *Information Systems*, Jg. 94, S. 101 582, Dez. 2020, ISSN: 0306-4379. DOI: 10.1016/j.is.2020.101582.

[118] J. Tang, Z. Meng, X. Nguyen, Q. Mei und M. Zhang, „Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis“, in *Proceedings of the 31st International Conference on Machine Learning*, Ser. ICML'14, Bd. 32, Peking, China: Journal of Machine Learning Research, Juni 2014, S. 190–198.

[119] Y. Zuo u. a., „Topic Modeling of Short Texts: A Pseudo-Document View“, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, Kalifornien, USA: Association for Computing Machinery, Aug. 2016, S. 2105–2114, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939880.

[120] W. Liu, Y. Huang, Y. Guo, Y. Wang, B. Fang und Q. Liao, „Topic Modeling for Short Texts Via Dual View Collaborate optimization“, in *Proceedings of the 7th IEEE International Conference on Data Science in Cyberspace*, Guilin, China: IEEE, Juli 2022, S. 160–166, ISBN: 978-1-66547-480-1. DOI: 10.1109/DSC55868.2022.00028.

[121] Y. Xu, Y. Li und D. T. K. Geeganage, „Investigation of the Quality of Topic Models for Noisy Data Sources“, in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Santiago, Chile: IEEE Computer Society, Dez. 2018, S. 488–493, ISBN: 978-1-5386-7325-6. DOI: 10.1109/WI.2018.00-48.

[122] J. A. Al-Ani und M. Fasli, „Probabilistic Relational Supervised Topic Modelling using Word Embeddings“, in *Proceedings of IEEE International Conference on Big Data*, Seattle, Washington, USA: IEEE, Dez. 2018, S. 2035–2043, ISBN: 978-1-5386-5035-6. DOI: 10.1109/BigData.2018.8622326.

[123] X. Li, Y. Wang, A. Zhang, C. Li, J. Chi und J. Ouyang, „Filtering Out the Noise in Short Text Topic Modeling“, *Information Sciences*, Jg. 456, S. 83–96, Aug. 2018, ISSN: 0020-0255. DOI: 10.1016/j.ins.2018.04.071.

[124] R. Churchill und L. Singh, „Topic-Noise Models: Modeling Topic and Noise Distributions in Social Media Post Collections“, in *Proceedings of IEEE International Conference on Data Mining*, Auckland, Neuseeland: IEEE, Dez. 2021, S. 71–80, ISBN: 978-1-66542-398-4. DOI: 10.1109/ICDM51629.2021.00017.

[125] A. Jayaweera, Y. Senanayake und P. S. Haddela, „Dynamic Stopword Removal for Sinhala Language“, in *Proceedings of National Information Technology Conference*, Colombo, Sri Lanka: IEEE, Okt. 2019, S. 81–86, ISBN: 978-1-72815-569-2. DOI: 10.1109/NITC48475.2019.9114476.

[126] N. Akhtar, M. M. Sufyan Beg und H. Javed, „Topic Modelling with Fuzzy Document Representation“, in *Proceedings of the Third International Conference on Advances in Computing and Data Sciences (ICACDS)*, M. Singh, P. Gupta, V. Tyagi, J. Flusser, T. Ören und R. Kashyap, Hrsg., Ser. Communications in Computer and Information Science (CCIS), Bd. 1046, Singapur, Singapur: Springer Nature, Juli 2019, S. 577–587, ISBN: 9789811399428. DOI: 10.1007/978-981-13-9942-8_54.

[127] X. Li, A. Zhang, C. Li, J. Ouyang und Y. Cai, „Exploring Coherent Topics by Topic Modeling with Term Weighting“, *Information Processing & Management*, Jg. 54, Nr. 6, S. 1345–1358, Nov. 2018, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2018.05.009.

[128] Q. Mei und C. Zhai, „A Mixture Model for Contextual Text Mining“, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, Pennsylvania, USA: Association for Computing Machinery, Aug. 2006, S. 649–655, ISBN: 978-1-59593-339-3. DOI: 10.1145/1150402.1150482.

[129] H. Walder, T. Hansjakob, T. E. Grundlach und P. Straub, *Kriminalistisches Denken* (Grundlagen der Kriminalistik), 11. Aufl. Heidelberg: C.F. Müller GmbH, 2020, ISBN: 978-3-7832-0043-0.

[130] R. Churchill, L. Singh, R. Ryan und P. Davis-Kean, „A Guided Topic-Noise Model for Short Texts“, in *Proceedings of the ACM Web Conference 2022*, Lyon, Frankreich: Association for Computing Machinery, Apr. 2022, S. 2870–2878, ISBN: 978-1-4503-9096-5. DOI: 10.1145/3485447.3512007.

[131] Y. Zuo, J. Zhao und K. Xu, „Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts“, *Knowledge and Information Systems*, Jg. 48, Nr. 2, S. 379–398, Sep. 2015, ISSN: 0219-3116. DOI: 10.1007/s10115-015-0882-z.

[132] J. Xi, M. Spranger und D. Labudde, „A Concept for a Comprehensive Understanding of Communication in Mobile Forensics“, in *Proceedings of the Tenth International Conference on Data Analytics*, Barcelona, Spanien: IARIA Press, Okt. 2021, S. 74–76, ISBN: 978-1-61208-891-4.

[133] M. Spranger, E. Zuchantke und D. Labudde, „Semantic Tools for Forensics: Towards Finding Evidence in Short Messages“, in *Proceedings of the Fourth International Conference on Advances in Information Mining and Management (IMMM)*, Paris, Frankreich: IARIA Press, 2014, S. 1–4, ISBN: 978-1-61208-364-3.

[134] M. Spranger und D. Labudde, „Semantic Tools for Forensics: Approaches in Forensic Text Analysis“, in *Proceedings of the Third International Conference on Advances in Information Mining and Management (IMMM)*, Lissabon, Portugal: IARIA Press, Nov. 2013, S. 97–100, ISBN: 978-1-61208-311-7. DOI: 10.13140/RG.2.1.2342.7685.

[135] L. Hilte, W. Daelemans und R. Vandekerckhove, „Lexical Patterns in Adolescents' Online Writing: The Impact of Age, Gender, and Education“, *Written Communication*, Jg. 37, Nr. 3, S. 365–400, Juli 2020, ISSN: 0741-0883. DOI: 10.1177/0741088320917921.

[136] S. T. Gries, „Polysemy“, in *Cognitive Linguistics - Key Topics*, E. Dąbrowska und D. Divjak, Hrsg., Berlin, Deutschland: De Gruyter, Juli 2019, S. 23–43, ISBN: 978-3-11-062643-8. DOI: 10.1515/9783110626438-002.

[137] M. Bernt und A. Schloenhardt, „Übereinkommen der Vereinten Nationen gegen die grenzüberschreitende organisierte Kriminalität“, in *Vertragliche Umsetzungskontrolle im Transnationalen Strafrecht*, M. Bernt und A. Schloenhardt, Hrsg., Berlin, Heidelberg: Springer, Aug. 2021, S. 101–140, ISBN: 978-3-662-63277-2. DOI: 10.1007/978-3-662-63277-2_5.

[138] J. Boyd-Graber und D. M. Blei, „Multilingual Topic Models for Unaligned Text“, in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, Kanada: AUAI Press, Juni 2009, S. 75–82, ISBN: 978-0-9749039-5-8.

[139] D. Zhang, Q. Mei und C. Zhai, „Cross-Lingual Latent Topic Extraction“, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Schweden: Association for Computational Linguistics, Juli 2010, S. 1128–1137, ISBN: 978-1-932432-66-4.

[140] J. Jagarlamudi und H. Daumé, „Extracting Multilingual Topics from Unaligned Comparable Corpora“, in *Proceedings of the 32nd European Conference on IR Research (ECIR)*, D. Hutchison u. a., Hrsg., Ser. Lecture Notes in Computer Science, Bd. 5993, Milton Keynes, Großbritannien: Springer Berlin Heidelberg, März 2010, S. 444–456, ISBN: 978-3-642-12274-3 978-3-642-12275-0. DOI: 10.1007/978-3-642-12275-0_39.

[141] D. Q. Nguyen, R. Billingsley, L. Du und M. Johnson, „Improving Topic Models with Latent Feature Word Representations“, *Transactions of the Association for Computational Linguistics*, Jg. 3, S. 299–313, Juni 2015, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00140.

[142] M. Jiang, R. Liu und F. Wang, „Word Network Topic Model Based on Word2Vector“, in *Proceedings of the Fourth International Conference on Big Data Computing Service and Applications*, Bamberg, Deutschland: IEEE, März 2018, S. 241–247, ISBN: 978-1-5386-5119-3. DOI: 10.1109/BigDataService.2018.00043.

[143] Y. Zuo, C. Li, H. Lin und J. Wu, „Topic Modeling of Short Texts: A Pseudo-Document View With Word Embedding Enhancement“, *IEEE Transactions on Knowledge and Data Engineering*, Jg. 35, Nr. 1, S. 972–985, Jan. 2023, ISSN: 1558-2191. DOI: 10.1109/TKDE.2021.3073195.

[144] K. Nigam, A. K. McCallum, S. Thrun und T. Mitchell, „Text Classification from Labeled and Unlabeled Documents using EM“, *Machine Learning*, Jg. 39, Nr. 2–3, S. 103–134, Mai 2000, ISSN: 0885-6125. DOI: 10.1023/A:1007692713085.

[145] W. X. Zhao u. a., „Comparing Twitter and Traditional Media Using Topic Models“, in *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*, Ser. Lecture Notes in Computer Science, Bd. 6611, Dublin, Irland: Springer Science+Business Media, Apr. 2011, S. 338–349, ISBN: 978-3-642-20160-8. DOI: 10.1007/978-3-642-20161-5.

[146] J. Yin und J. Wang, „A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering“, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Ser. KDD '14, New York City, New York, USA: Association for Computing Machinery, Aug. 2014, S. 233–242, ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623715.

[147] K. Sasaki, T. Yoshikawa und T. Furuhashi, „Online Topic Model for Twitter Considering Dynamics of User Interests and Topic Trends“, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Katar: Association for Computational Linguistics, Okt. 2014, S. 1977–1985. DOI: 10.3115/v1/D14-1212.

[148] T. Lin, W. Tian, Q. Mei und H. Cheng, „The Dual-Sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text“, in *Proceedings of the 23rd International Conference on World Wide Web*, Seoul, Südkorea: Association for Computing Machinery, Apr. 2014, S. 539–550, ISBN: 978-1-4503-2744-2. DOI: 10.1145/2566486.2567980.

[149] X. Li, Y. Wang, J. Ouyang und M. Wang, „Topic Extraction from Extremely Short Texts with Variational Manifold Regularization“, *Machine Language*, Jg. 110, Nr. 5, S. 1029–1066, Mai 2021, ISSN: 0885-6125. DOI: 10.1007/s10994-021-05962-3.

[150] D. Newman, J. H. Lau, K. Grieser und T. Baldwin, „Automatic Evaluation of Topic Coherence“, in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, Los Angeles, Kalifornien, USA: Association for Computational Linguistics, Juni 2010, S. 100–108.

[151] O. Jin, N. N. Liu, K. Zhao, Y. Yu und Q. Yang, „Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering“, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, Glasgow, Schottland, Großbritannien: Association for Computing Machinery, Okt. 2011, S. 775–784, ISBN: 978-1-4503-0717-8. DOI: 10.1145/2063576.2063689.

[152] P. Gupta, Y. Chaudhary und H. Schütze, „Multi-source Neural Topic Modeling in Multi-view Embedding Spaces“, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Juni 2021, S. 4205–4217, ISBN: 978-1-71383-013-9. DOI: 10.18653/v1/2021.naacl-main.332.

[153] R. Mehrotra, S. Sanner, W. Buntine und L. Xie, „Improving LDA Topic Models for Microblogs Via Tweet Pooling and Automatic Labeling“, in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Irland: Association for Computing Machinery, Juli 2013, S. 889–892, ISBN: 978-1-4503-2034-4. DOI: 10.1145/2484028.2484166.

[154] J. Weng, E.-P. Lim, J. Jiang und Q. He, „TwitterRank: Finding Topic-Sensitive Influential Twitterers“, in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, New York, New York, USA: Association for Computing Machinery, Feb. 2010, S. 261–270, ISBN: 978-1-60558-889-6. DOI: 10.1145/1718487.1718520.

[155] J. Qiang, P. Chen, T. Wang und X. Wu, „Topic Modeling over Short Texts by Incorporating Word Embeddings“, in *Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, J. Kim, K. Shim, L. Cao, J.-G. Lee, X. Lin und Y.-S. Moon, Hrsg., Ser. Lecture Notes in Computer Science, Cham, Schweiz: Springer International Publishing, Apr. 2017, S. 363–374, ISBN: 978-3-319-57529-2. DOI: 10.1007/978-3-319-57529-2_29.

[156] F. Yi, B. Jiang und J. Wu, „Topic Modeling for Short Texts via Word Embedding and Document Correlation“, *IEEE Access*, Jg. 8, S. 30 692–30 705, Feb. 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2973207.

[157] X. Quan, C. Kit, Y. Ge und S. J. Pan, „Short and Sparse Text Topic Modeling Via Self-Aggregation“, in *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*, Buenos Aires, Argentinien: AAAI Press, Juli 2015, S. 2270–2276, ISBN: 978-1-57735-738-4.

[158] W. Liang, R. Feng, X. Liu, Y. Li und X. Zhang, „GLTM: A Global and Local Word Embedding-Based Topic Model for Short Texts“, *IEEE Access*, Jg. 6, S. 43 612–43 621, Aug. 2018, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2863260.

[159] X. Yan, J. Guo, S. Liu, X. Cheng und Y. Wang, „Learning Topics in Short Texts by Non-negative Matrix Factorization on Term Correlation Matrix“, in *Proceedings of the SIAM International Conference on Data Mining (SDM)*, Austin, Texas, USA: Society for Industrial and Applied Mathematics, Mai 2013, S. 749–757, ISBN: 978-1-61197-262-7 978-1-61197-283-2. DOI: 10.1137/1.9781611972832.83.

[160] Y. Xia, N. Tang, A. Hussain und E. Cambria, „Discriminative Bi-Term Topic Model for Headline-Based Social News Clustering“, in *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, Hollywood, Florida, USA: Association for the Advancement of Artificial Intelligence (AAAI), Mai 2015, S. 311–316, ISBN: 978-1-57735-730-8.

[161] X. Yan, J. Guo, Y. Lan, J. Xu und X. Cheng, „A Probabilistic Model for Bursty Topic Discovery in Microblogs“, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas, USA: AAAI Press, Jan. 2015, S. 353–359, ISBN: 978-0-262-51129-2.

[162] H.-Y. Lu, G.-J. Ge, Y. Li, C.-J. Wang und J.-Y. Xie, „Exploiting Global Semantic Similarity Biterms for Short-Text Topic Discovery“, in *Proceedings of the 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, Volos, Griechenland: IEEE Computer Society, Nov. 2018, S. 975–982, ISBN: 978-1-5386-7450-5. DOI: 10.1109/ICTAI.2018.00151.

[163] X. Li, A. Zhang, C. Li, L. Guo, W. Wang und J. Ouyang, „Relational Biterm Topic Model: Short-Text Topic Modeling using Word Embeddings“, *The Computer Journal*, Jg. 62, Nr. 3, S. 359–372, März 2019, ISSN: 0010-4620, 1460-2067. DOI: 10.1093/comjnl/bxy037.

[164] F. Wang, R. Liu, Y. Zuo, H. Zhang, H. Zhang und J. Wu, „Robust Word-Network Topic Model for Short Texts“, in *Proceedings of the 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, Kalifornien, USA: IEEE Computer Society, Nov. 2016, S. 852–856, ISBN: 978-1-5090-4460-3. DOI: 10.1109/ICTAI.2016.0132.

[165] W. Chen, J. Wang, Y. Zhang, H. Yan und X. Li, „User Based Aggregation for Biterm Topic Model“, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, Bd. 2, Peking, China: Association for Computational Linguistics, Juli 2015, S. 489–494, ISBN: 978-1-941643-73-0. DOI: 10.3115/v1/P15-2080.

[166] X. Wu und C. Li, „Short Text Topic Modeling with Flexible Word Patterns“, in *Proceedings of the International Joint Conference on Neural Networks*, Budapest, Ungarn: IEEE, Juli 2019, S. 1–7, ISBN: 978-1-72811-985-4. DOI: 10.1109/IJCNN.2019.8852366.

[167] L. Li, Y. Sun und C. Wang, „Semantic Augmented Topic Model over Short Text“, in *Proceedings of the 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, Nanjing, China: IEEE, Nov. 2018, S. 652–656, ISBN: 978-1-5386-6005-8. DOI: 10.1109/CCIS.2018.8691313.

[168] L. Zhen, S. Yabin und Y. Ning, „A Short Text Topic Model Based on Semantics and Word Expansion“, in *Proceedings of the 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI)*, Peking, China: IEEE, Mai 2022, S. 60–64, ISBN: 978-1-66549-663-6. DOI: 10.1109/CCAI55564.2022.9807822.

[169] L. Jiang, H. Lu, M. Xu und C. Wang, „Biterm Pseudo Document Topic Model for Short Text“, in *Proceedings of the 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, Kalifornien, USA: IEEE Computer Society, Nov. 2016, S. 865–872, ISBN: 978-1-5090-4459-7. DOI: 10.1109/ICTAI.2016.0134.

[170] C. Li, H. Wang, Z. Zhang, A. Sun und Z. Ma, „Topic Modeling for Short Texts with Auxiliary Word Embeddings“, in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italien: Association for Computing Machinery, Juli 2016, S. 165–174, ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2911499.

[171] F. Viegas, W. Cunha, C. Gomes, A. Pereira, L. Rocha und M. Goncalves, „CluHTM - Semantic Hierarchical Topic Modeling based on CluWords“, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Juli 2020, S. 8138–8150, ISBN: 978-1-952148-25-5. DOI: 10.18653/v1/2020.acl-main.724.

[172] P. Bicalho, M. Pita, G. Pedrosa, A. Lacerda und G. L. Pappa, „A General Framework to Expand Short Text for Topic Modeling“, *Information Sciences*, Jg. 393, S. 66–81, Juli 2017, ISSN: 0020-0255. DOI: 10.1016/j.ins.2017.02.007.

[173] R. Murakami und B. Chakraborty, „Neural Topic Models for Short Text Using Pretrained Word Embeddings and Its Application To Real Data“, in *Proceedings of the 4th International Conference on Knowledge Innovation and Invention (ICKII)*, Taichung, Taiwan: IEEE, Juli 2021, S. 146–150, ISBN: 978-1-66542-307-6. DOI: 10.1109/ICKII51822.2021.9574752.

[174] R. Zhao und K. Mao, „Fuzzy Bag-of-Words Model for Document Representation“, *IEEE Transactions on Fuzzy Systems*, Jg. 26, Nr. 2, S. 794–804, Apr. 2018, ISSN: 1941-0034. DOI: 10.1109/TFUZZ.2017.2690222.

[175] S. Limwattana und S. Promon, „Topic Modeling Enhancement using Word Embeddings“, in *Proceedings of the 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Lampang, Thailand: IEEE, Juni 2021, S. 1–5, ISBN: 978-1-66543-831-5. DOI: 10.1109/JCSSE53117.2021.9493816.

[176] H. M. Mahmoud, *Pólya Urn Models* (Texts in Statistical Science). District of Columbia, USA: CRC Press, Juni 2008, ISBN: 978-1-4200-5984-7.

[177] F. Zhang, W. Gao, Y. Fang und B. Zhang, „Enhancing Short Text Topic Modeling with FastText Embeddings“, in *Proceedings of the International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Fuzhou, China: IEEE, Juni 2020, S. 255–259, ISBN: 978-1-72816-499-1. DOI: 10.1109/ICBAIE49996.2020.00060.

[178] Y. Liu, A. Tang, Z. Sun, W. Tang, F. Cai und C. Wang, „An Integrated Retrieval Framework for Similar Questions: Word-Semantic Embedded Label Clustering – LDA with Question Life Cycle“, *Information Sciences*, Jg. 537, S. 227–245, Okt. 2020, ISSN: 0020-0255. DOI: 10.1016/j.ins.2020.05.014.

[179] P. Bojanowski, E. Grave, A. Joulin und T. Mikolov, „Enriching Word Vectors with Subword Information“, *Transactions of the Association for Computational Linguistics*, Jg. 5, S. 135–146, Dez. 2017, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00051.

[180] Y. Qiu, H. Li, S. Li, Y. Jiang, R. Hu und L. Yang, „Revisiting Correlations between Intrinsic and Extrinsic Evaluations of Word Embeddings“, in *Proceedings of the China National Conference on Chinese Computational Linguistics*, M. Sun, T. Liu, X. Wang, Z. Liu und Y. Liu, Hrsg., Ser. Lecture Notes in Computer Science, Cham, Schweiz: Springer International Publishing, Okt. 2018, S. 209–221, ISBN: 978-3-030-01716-3. DOI: 10.1007/978-3-030-01716-3_18.

[181] T. Mikolov, A. Deoras, D. Povey, L. Burget und J. Černocký, „Strategies for Training Large Scale Neural Network Language Models“, in *Proceeding of the Workshop on Automatic Speech Recognition & Understanding*, Waikoloa, Hawaii, USA: IEEE, Dez. 2011, S. 196–201, ISBN: 978-1-4673-0366-8. DOI: 10.1109/ASRU.2011.6163930.

[182] J. Wang, L. Chen, L. Qin und X. Wu, „ASTM: An Attentional Segmentation Based Topic Model for Short Texts“, in *Proceedings of the International Conference on Data Mining (ICDM)*, Singapur, Singapur: IEEE Computer Society, Nov. 2018, S. 577–586, ISBN: 978-1-5386-9159-5. DOI: 10.1109/ICDM.2018.00073.

[183] Y. Zhao und G. Karypis, „Criterion Functions for Document Clustering: Experiments and Analysis“, University of Minnesota Digital Conservancy, Minneapolis, Minnesota, USA, Report, Nov. 2001, S. 1–30.

[184] R. Huang, G. Yu, Z. Wang, J. Zhang und L. Shi, „Dirichlet Process Mixture Model for Document Clustering with Feature Partition“, *IEEE Transactions on Knowledge and Data Engineering*, Jg. 25, Nr. 8, S. 1748–1759, Aug. 2013, ISSN: 1558-2191. DOI: 10.1109/TKDE.2012.27.

[185] C. Rashtchian, P. Young, M. Hodosh und J. Hockenmaier, „Collecting Image Annotations Using Amazon's Mechanical Turk“, in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, Kalifornien, USA: Association for Computational Linguistics, Juni 2010, S. 139–147.

[186] K. Gorro u. a., „Qualitative Data Analysis of Disaster Risk Reduction Suggestions Assisted by Topic Modeling and Word2vec“, in *Proceeding of International Conference on Asian Language Processing (IALP)*, Singapur, Singapur: IEEE, Dez. 2017, S. 293–297, ISBN: 978-1-5386-1981-0. DOI: 10.1109/IALP.2017.8300601.

[187] R. Churchill und L. Singh, „textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data“, in *Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA)*, Online: Science and Technology Publications (SCITEPRESS), Juli 2021, S. 60–70, ISBN: 978-989-758-521-0. DOI: 10.5220/0010559000600070.

[188] E. Ukkonen, „Algorithms for Approximate String Matching“, *Information and Control*, Jg. 64, Nr. 1-3, S. 100–118, Jan. 1985, ISSN: 00199958. DOI: 10.1016/S0019-9958(85)80046-2.

[189] A. T. Wilson und P. A. Chew, „Term weighting schemes for Latent Dirichlet Allocation“, in *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, Kalifornien, USA: Association for Computational Linguistics, Juni 2010, S. 465–473, ISBN: 978-1-932432-65-7.

[190] K. Yang, Y. Cai, Z. Chen, H.-f. Leung und R. Lau, „Exploring Topic Discriminating Power of Words in Latent Dirichlet Allocation“, in *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, Osaka, Japan: The COLING 2016 Organizing Committee, Dez. 2016, S. 2238–2247.

[191] K. W. Church und P. Hanks, „Word Association Norms, Mutual Information, and Lexicography“, in *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, Bd. 16, Vancouver, Kanada: Association for Computational Linguistics, Juni 1989, S. 76–83. DOI: 10.3115/981623.981633.

[192] C. E. Shannon, „A Mathematical Theory of Communication“, *The Bell System Technical Journal*, Jg. 27, Nr. 3, S. 379–423, Juli 1948, ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[193] S. Geman und D. Geman, „Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images“, *Transactions on Pattern Analysis and Machine Intelligence*, Jg. PAMI-6, Nr. 6, S. 721–741, Nov. 1984, ISSN: 1939-3539. DOI: 10.1109/TPAMI.1984.4767596.

[194] T. Wang, Y. Cai, H.-f. Leung, Z. Cai und H. Min, „Entropy-Based Term Weighting Schemes for Text Categorization in VSM“, in *Proceedings of 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, Vietri sul Mare, Italien: IEEE Computer Society, Nov. 2015, S. 325–332, ISBN: 978-1-5090-0163-7. DOI: 10.1109/ICTAI.2015.57.

[195] Q. Mei und C. Zhai, „A Note on EM Algorithm for Probabilistic Latent Semantic Analysis“, in *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, Atlanta, Georgia, USA: Association for Computing Machinery, Okt. 2001, S. 1–4, ISBN: 978-1-58113-436-0.

[196] C. Chemudugunta, P. Smyth und M. Steyvers, „Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model“, in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, Bd. 19, Vancouver, Kanada: MIT Press, Dez. 2006, S. 241–248, ISBN: 978-0-262-25691-9.

[197] G. Bouma, „Normalized (Pointwise) Mutual Information in Collocation Extraction“, in *Proceedings of the Biennal German Vorfeld for Local Coherence (GSCL)*, Potsdam, Deutschland: Gunter Narr Verlag, Jan. 2009, S. 31–40, ISBN: 978-3-8233-6511-2.

[198] R. Mihalcea und P. Tarau, „TextRank: Bringing Order into Text“, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spanien: Association for Computational Linguistics, Juli 2004, S. 404–411.

[199] I. Derényi, G. Palla und T. Vicsek, „Clique Percolation in Random Networks“, *Physical Review Letters*, Jg. 94, Nr. 16, S. 160 202, Apr. 2005, ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.94.160202.

[200] R. Churchill, „Modernizing Topic Models: Accounting for Noise, Time, and Domain Knowledge“, Diss., Georgetown University, Washington, D.C., USA, Dez. 2021, ISBN: 9798780612063.

[201] E. M. Williams, D. Levin und I. McCulloh, „Improving LDA Topic Modeling with Gamma and Simmelian Filtration“, in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Den Haag, Niederlande: IEEE, Dez. 2020, S. 692–696, ISBN: 978-1-72811-056-1. DOI: 10.1109/ASONAM49781.2020.9381330.

[202] Y. He, C. Wang und C. Jiang, „Modeling Document Networks with Tree-Averaged Copula Regularization“, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, Cambridge, Großbritannien: Association for Computing Machinery, Feb. 2017, S. 691–699, ISBN: 978-1-4503-4675-7. DOI: 10.1145/3018661.3018666.

[203] D. M. Blei und J. D. Lafferty, „Dynamic topic models“, in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, Juni 2006, S. 113–120, ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143859.

[204] A. McCallum, A. Corrada-Emmanuel und X. Wang, „Topic and Role Discovery in Social Networks“, in *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Juli 2005, S. 786–791.

[205] J. Risch und R. Krestel, „My Approach = Your Apparatus?“, in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18)*, Fort Worth, Texas, USA: Association for Computing Machinery, Mai 2018, S. 283–292, ISBN: 978-1-4503-5178-2. DOI: 10.1145/3197026.3197038.

[206] X. Wang und A. McCallum, „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*, Philadelphia, Pennsylvania, USA: Association for Computing Machinery, Aug. 2006, S. 424–433, ISBN: 978-1-59593-339-3. DOI: 10.1145/1150402.1150450.

[207] H. S. Banu und S. Chitrakala, „Trending Topic Analysis Using Novel Sub Topic Detection Model“, in *Proceedings of the 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, Indien: IEEE, Feb. 2016, S. 157–161, ISBN: 978-1-4673-9745-2. DOI: 10.1109/AEEICB.2016.7538263.

[208] M. Yang, Q. Qu, X. Chen, W. Tu, Y. Shen und J. Zhu, „Discovering Author Interest Evolution in Order-Sensitive and Semantic-Aware Topic Modeling“, *Information Sciences*, Jg. 486, S. 271–286, Juni 2019, ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.02.040.

[209] D. C. Zhang und H. W. Lauw, „Variational Graph Author Topic Modeling“, in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, Washington DC, USA: Association for Computing Machinery, Aug. 2022, S. 2429–2438, ISBN: 978-1-4503-9385-0. DOI: 10.1145/3534678.3539310.

[210] D. Li u. a., „Adding Community and Dynamic to Topic Models“, *Journal of Informetrics*, Jg. 6, Nr. 2, S. 237–253, Apr. 2012, ISSN: 1751-1577. DOI: 10.1016/j.joi.2011.11.004.

[211] M. G. Lozano, J. Schreiber und J. Brynielsson, „Tracking Geographical Locations Using a Geo-Aware Topic Model for Analyzing Social Media Data“, *Decision Support Systems*, Jg. 99, S. 18–29, Juli 2017, ISSN: 01679236. DOI: 10.1016/j.dss.2017.05.006.

[212] M. Paul und R. Girju, „Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models“, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapur, Singapur: Association for Computational Linguistics, Aug. 2009, S. 1408–1417.

[213] J. Chang und D. Blei, „Relational Topic Models for Document Networks“, in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, Bd. 5, Clearwater Beach, Florida, USA: MIT Press, Apr. 2009, S. 81–88.

[214] S. Terragni, E. Fersini und E. Messina, „Constrained Relational Topic Models“, *Information Sciences*, Jg. 512, S. 581–594, Feb. 2020, ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.09.039.

[215] Q. Mei, D. Cai, D. Zhang und C. Zhai, „Topic Modeling with Network Regularization“, in *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, Peking, China: Association for Computing Machinery, Apr. 2008, S. 101–110, ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367512.

[216] D. Duan, Y. Li, R. Li, R. Zhang, X. Gu und K. Wen, „LIMTopic: A Framework of Incorporating Link Based Importance into Topic Modeling“, *IEEE Transactions on Knowledge and Data Engineering*, Jg. 26, Nr. 10, S. 2493–2506, Okt. 2014, ISSN: 1558-2191. DOI: 10.1109/TKDE.2013.2297912.

[217] Y. Liu und S. Xu, „A Local Context-Aware LDA Model for Topic Modeling in a Document Network“, *Journal of the Association for Information Science and Technology*, Jg. 68, Nr. 6, S. 1429–1448, Apr. 2017, ISSN: 2330-1643. DOI: 10.1002/asi.23822.

[218] B. Hu, Z. Song und M. Ester, „User Features and Social Networks for Topic Modeling in Online Social Media“, in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, Türkei: IEEE Computer Society, Aug. 2012, S. 202–209, ISBN: 978-1-4673-2497-7. DOI: 10.1109/ASONAM.2012.43.

[219] D. M. Blei und J. D. McAuliffe, „Supervised Topic Models“, in *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*, Vancouver, Kanada: Curran Associates, Dez. 2007, S. 121–128, ISBN: 978-1-60560-352-0.

[220] Y. Zhang und W. Wei, „A Jointly Distributed Semi-Supervised Topic Model“, *Neurocomputing*, Special Issue on the 2011 Sino-Foreign-Interchange Workshop on Intelligence Science and Intelligent Data Engineering (IScIDE 2011), Jg. 134, S. 38–45, Juni 2014, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2012.12.077.

[221] H. Kim, D. Choi, B. Drake, A. Endert und H. Park, „TopicSifter: Interactive Search Space Reduction through Targeted Topic Modeling“, in *Proceedings of the Conference on Visual Analytics Science and Technology*, Vancouver, Kanada: IEEE, Okt. 2019, S. 35–45, ISBN: 978-1-72812-284-7. DOI: 10.1109/VAST47406.2019.8986922.

[222] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim und C. Collins, „Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework“, *IEEE Transactions on Visualization and Computer Graphics*, Jg. 24, Nr. 1, S. 382–391, Jan. 2018, ISSN: 1941-0506. DOI: 10.1109/TVCG.2017.2745080.

[223] D. Andrzejewski, X. Zhu und M. Craven, „Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors“, in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, Bd. 382, Montreal, Quebec, Canada: Association for Computing Machinery, Juni 2009, S. 25–32, ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553378.

[224] H. Kobayashi, H. Wakaki, T. Yamasaki und M. Suzuki, „Topic Models with Logical Constraints on Words“, in *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*, Chissarja, Bulgarien: Association for Computational Linguistics, Sep. 2011, S. 33–40, ISBN: 978-954-452-017-5.

[225] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos und R. Ghosh, „Exploiting Domain Knowledge in Aspect Extraction“, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Karlsruhe, Deutschland: Association for Computational Linguistics, Mai 2013, S. 1655–1667, ISBN: 978-1-937284-97-8. DOI: 10.1109/ICRA.2013.6630861.

[226] J. He, L. Li, Y. Wang und X. Wu, „Targeted Aspects Oriented Topic Modeling for Short Texts“, *Applied Intelligence*, Jg. 50, Nr. 8, S. 2384–2399, März 2020, ISSN: 1573-7497. DOI: 10.1007/s10489-020-01672-w.

[227] J. Chen, Z. Gong, W. Wang, W. Liu, M. Yang und C. Wang, „TAM: Targeted Analysis Model With Reinforcement Learning on Short Texts“, *IEEE Transactions on Neural Networks and Learning Systems*, Jg. 32, Nr. 6, S. 2772–2781, Juni 2021, ISSN: 2162-2388. DOI: 10.1109/TNNLS.2020.3009247.

[228] J. Wang, L. Chen, L. Li und X. Wu, „BiTTM: A Core Biterms-Based Topic Model for Targeted Analysis“, *Applied Sciences*, Jg. 11, Nr. 21, S. 10 162–10 184, Okt. 2021, ISSN: 2076-3417. DOI: 10.3390/app112110162.

[229] J. He, L. Li, Y. Wang und X. Wu, „Hierarchical Features-Based Targeted Aspect Extraction from Online Reviews“, *Intelligent Data Analysis*, Jg. 25, Nr. 1, S. 205–223, Jan. 2021. DOI: 10.3233/IDA-194952.

[230] D. Zha und C. Li, „Multi-Label Dataless Text Classification with Topic Modeling“, *Knowledge and Information Systems*, Jg. 61, Nr. 1, S. 137–160, Dez. 2018, ISSN: 0219-1377. DOI: 10.1007/s10115-018-1280-0.

[231] J. Li, Y. Qin und R. Huang, „A User-oriented Semi-supervised Probabilistic Topic Model“, in *Proceedings of the 2nd International Conference on Computer and Communications (ICCC)*, Chengdu, China: IEEE, Okt. 2016, S. 262–268, ISBN: 978-1-4673-9026-2. DOI: 10.1109/CompComm.2016.7924706.

[232] B. Lu, M. Ott, C. Cardie und B. K. Tsou, „Multi-aspect Sentiment Analysis with Topic Models“, in *Proceedings of the 11th International Conference on Data Mining Workshops*, Vancouver, Kanada: IEEE Computer Society, Dez. 2011, S. 81–88, ISBN: 978-1-4673-0005-6. DOI: 10.1109/ICDMW.2011.125.

[233] B. Harandizadeh, H. Priniski und F. Morstatter, „Keyword Assisted Embedded Topic Model“, in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, Arizona, USA: Association for Computing Machinery, Feb. 2022, S. 372–380, ISBN: 978-1-4503-9132-0. DOI: 10.1145/3488560.3498518.

[234] Y. Feng, J. Feng und Y. Rao, „Reward-Modulated Adversarial Topic Modeling“, in *Proceedings of the 25th International Conference on Database Systems for Advanced Applications*, Ser. Lecture Notes in Computer Science, Bd. 12112, Jeju, Südkorea: Springer Nature Switzerland AG, Sep. 2020, S. 689–697, ISBN: 978-3-030-59409-1. DOI: 10.1007/978-3-030-59410-7_47.

[235] Y. Meng u. a., „Discriminative Topic Mining via Category-Name Guided Text Embedding“, in *Proceedings of The Web Conference (WWW '20)*, Taipei, Taiwan: Association for Computing Machinery, Apr. 2020, S. 2121–2132, ISBN: 978-1-4503-7023-3. DOI: 10.1145/3366423.3380278.

[236] Y. Zhang, Y. Meng, X. Wang, S. Wang und J. Han, „Seed-Guided Topic Discovery with Out-of-Vocabulary Seeds“, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, USA: Association for Computational Linguistics, Juli 2022, S. 279–290, ISBN: 978-1-955917-71-1. DOI: 10.18653/v1/2022.naacl-main.21.

[237] Y. Zhang, Y. Zhang, M. Michalski, Y. Jiang, Y. Meng und J. Han, „Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts“, in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, Singapur, Singapur: Association for Computing Machinery, Feb. 2023, S. 429–437, ISBN: 978-1-4503-9407-9. DOI: 10.1145/3539597.3570475.

[238] D. Andrzejewski, X. Zhu, M. Craven und B. Recht, „A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic“, in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11)*, Barcelona, Spanien: AAAI Press, Juli 2011, S. 1171–1177, ISBN: 978-1-57735-514-4.

[239] Z. Chen und B. Liu, „Mining Topics in Documents: Standing on the Shoulders of Big Data“, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, New York City, New York, USA: Association for Computing Machinery, Aug. 2014, S. 1116–1125, ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623622.

[240] Z. Zhai, B. Liu, H. Xu und P. Jia, „Constrained LDA for Grouping Product Features in Opinion Mining“, in *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, J. Z. Huang, L. Cao und J. Srivastava, Hrsg., Ser. Lecture Notes in Artificial Intelligence, Bd. 6634, Shenzhen, China: Springer Science+Business Media, Mai 2011, S. 448–459, ISBN: 978-3-642-20840-9 978-3-642-20841-6. DOI: 10.1007/978-3-642-20841-6.

[241] M. Xu, R. Yang, S. Harenberg und N. F. Samatova, „A Lifelong Learning Topic Model Structured Using Latent Embeddings“, in *Proceedings of the 11th International Conference on Semantic Computing*, San Diego, Kalifornien, USA: IEEE Computer Society, Jan. 2017, S. 260–261, ISBN: 978-1-5090-4284-5. DOI: 10.1109/ICSC.2017.15.

[242] M. T. Khan, N. Azam, S. Khalid und J. Yao, „A Three-Way Approach for Learning Rules in Automatic Knowledge-Based Topic Models“, *International Journal of Approximate Reasoning*, Jg. 82, S. 210–226, März 2017, ISSN: 0888-613X. DOI: 10.1016/j.ijar.2016.12.011.

[243] Y. Chen, J. Wu, J. Lin, R. Liu, H. Zhang und Z. Ye, „Affinity Regularized Non-Negative Matrix Factorization for Lifelong Topic Modeling“, *Transactions on Knowledge and Data Engineering*, Jg. 32, Nr. 7, S. 1249–1262, Juli 2020, ISSN: 1558-2191. DOI: 10.1109/TKDE.2019.2904687.

[244] Z. Chen, A. Mukherjee und B. Liu, „Aspect Extraction with Automated Prior Knowledge Learning“, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Bd. 1, Baltimore, Maryland, USA: Association for Computational Linguistics, Juni 2014, S. 347–358, ISBN: 978-1-937284-72-5. DOI: 10.3115/v1/P14-1033.

[245] M. T. Khan, S. Yar, S. Khalid und F. Aziz, „Evolving Long-Term Dependency Rules in Lifelong Learning Models“, in *Proceedings of the International Conference on Knowledge Engineering and Applications*, Singapur, Singapur: IEEE, Sep. 2016, S. 93–97, ISBN: 978-1-5090-3471-0. DOI: 10.1109/ICKEA.2016.7802999.

[246] M. T. Khan, S. Khalid und F. Aziz, „Graph Clustering Based Size Varying Rules for Lifelong Topic Modeling“, in *Proceedings of the 5th International Conference on Bioinformatics Research and Applications (ICBRA '18)*, New York, NY, USA: Association for Computing Machinery, Dez. 2018, S. 73–77, ISBN: 978-1-4503-6611-3. DOI: 10.1145/3309129.3309146.

[247] M. T. Khan, S. Yar und S. Khalid, „Histogram Based Rule Verification in Lifelong Learning Models“, in *Proceedings of the 19th International Multi-Topic Conference (INMIC)*, Islamabad, Pakistan: IEEE, Dez. 2016, S. 1–5, ISBN: 978-1-5090-4300-2. DOI: 10.1109/INMIC.2016.7840096.

[248] X. Qin, Y. Lu, Y. Chen und Y. Rao, „Lifelong Learning of Topics and Domain-Specific Word Embeddings“, in *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, Online: Association for Computational Linguistics, Aug. 2021, S. 2294–2309, ISBN: 978-1-954085-54-1. DOI: 10.18653/v1/2021.findings-acl.202.

[249] S. Wang, Z. Chen und B. Liu, „Mining Aspect-Specific Opinion using a Holistic Lifelong Topic Model“, in *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*, Montréal, Québec, Kanada: International World Wide Web Conferences Steering Committee, Apr. 2016, S. 167–176, ISBN: 978-1-4503-4143-1. DOI: 10.1145/2872427.2883086.

[250] M. T. Khan und S. Khalid, „Multimodal Rule Transfer into Automatic Knowledge Based Topic Models“, in *Proceedings of the 19th International Multi-Topic Conference (INMIC)*, Islamabad, Pakistan: IEEE, Dez. 2016, S. 1–6, ISBN: 978-1-5090-4300-2. DOI: 10.1109/INMIC.2016.7840095.

[251] P. Gupta, Y. Chaudhary, T. Runkler und H. Schütze, „Neural Topic Modeling with Continual Lifelong Learning“, in *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, Bd. 119, Online: Journal of Machine Learning Research, Juli 2020, S. 3907–3917.

[252] Z. Lei, H. Liu, J. Yan, Y. Rao und Q. Li, „NMTF-LTM: Towards an Alignment of Semantics for Lifelong Topic Modeling“, *Transactions on Knowledge and Data Engineering*, S. 1–16, Apr. 2023, ISSN: 1558-2191. DOI: 10.1109/TKDE.2023.3267496.

[253] M. T. Khan, M. Durrani, S. Khalid und F. Aziz, „Online Knowledge-Based Model for Big Data Topic Extraction“, *Computational Intelligence and Neuroscience*, Jg. 2016, S. 1–11, Apr. 2016, ISSN: 1687-5265. DOI: 10.1155/2016/6081804.

[254] Z. Chen und B. Liu, „Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data“, in *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML'14)*, Bd. 32, Peking, China: Journal of Machine Learning Research, Juni 2014, S. 703–711.

[255] B. Liu, W. Hsu und Y. Ma, „Mining Association Rules with Multiple Minimum Supports“, in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '99)*, San Diego, Kalifornien, USA: Association for Computing Machinery, Aug. 1999, S. 337–341, ISBN: 978-1-58113-143-7. DOI: 10.1145/312129.312274.

[256] R. Agrawal und R. Srikant, „Fast Algorithms for Mining Association Rules in Large Databases“, in *Proceedings of the 20th International Conference On Very Large Data Base (VLDB '94)*, Santiago, Chile: Morgan Kaufmann Publishers Inc., Sep. 1994, S. 487–499, ISBN: 978-1-55860-153-6.

[257] H. Larochelle und S. Lauly, „A Neural Autoregressive Topic Model“, in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12)*, Bd. 25, Lake Tahoe, Kalifornien, USA: Curran Associates, Dez. 2012, S. 2708–2716.

[258] J. Felser, J. Xi, C. Demus, D. Labudde und M. Spranger, „Recommendation of Query Terms for Colloquial Texts in Forensic Text Analysis“, in *Proceedings of the International Workshop On Digital Forensics (IWDF)*, Hamburg, Deutschland: Gesellschaft für Informatik (GI), Sep. 2022, S. 35–47, ISBN: 978-3-88579-720-3. DOI: 10.18420/inf2022_02.

[259] Q. Wang, D. Song und X. Li, „Incorporating Entity Correlation Knowledge into Topic Modeling“, in *Proceedings of the International Conference on Big Knowledge (ICBK)*, Hefei, China: IEEE Computer Society, Aug. 2017, S. 254–258, ISBN: 978-1-5386-3120-1. DOI: 10.1109/ICBK.2017.33.

[260] M. Allahyari und K. Kochut, „Discovering Coherent Topics with Entity Topic Models“, in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Omaha, Nebraska, USA: IEEE Computer Society, Okt. 2016, S. 26–33, ISBN: 978-1-5090-4470-2. DOI: 10.1109/WI.2016.0015.

[261] D. Song, J. Gao, J. Pang, L. Liao und L. Qin, „Knowledge Base Enhanced Topic Modeling“, in *Proceedings of the International Conference on Knowledge Graph (ICKG)*, Nanjing, China: IEEE, Aug. 2020, S. 380–387, ISBN: 978-1-72818-156-1. DOI: 10.1109/ICBK50248.2020.00061.

[262] N. Van Linh, T. X. Bach und K. Than, „A Graph Convolutional Topic Model for Short and Noisy Text Streams“, *Neurocomputing*, Jg. 468, S. 345–359, Jan. 2022, ISSN: 09252312. DOI: 10.1016/j.neucom.2021.10.047.

[263] G. A. Miller, „WordNet: A Lexical Database for English“, *Communications of the ACM*, Jg. 38, Nr. 11, S. 39–41, Nov. 1995, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/219717.219748.

[264] T. N. Kipf und M. Welling, *Semi-Supervised Classification with Graph Convolutional Networks*, Toulon, Frankreich, Feb. 2017. DOI: 10.48550/arXiv.1609.02907. arXiv: 1609.02907 [cs, stat].

[265] C. Bizer u. a., „DBpedia - a Crystallization Point for the Web of Data“, *Journal of Web Semantics*, Jg. 7, Nr. 3, S. 154–165, Juli 2009, ISSN: 15708268. DOI: 10.1016/j.websem.2009.07.002.

[266] K. Bollacker, C. Evans, P. Paritosh, T. Sturge und J. Taylor, „Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge“, in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*, Vancouver, Kanada: Association for Computing Machinery, Juni 2008, S. 1247–1250, ISBN: 978-1-60558-102-6. DOI: 10.1145/1376616.1376746.

[267] W. Wu, H. Li, H. Wang und K. Q. Zhu, „Probase: A Probabilistic Taxonomy for Text Understanding“, in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD '12)*, Scottsdale, Arizona, USA: Association for Computing Machinery, Mai 2012, S. 481–492, ISBN: 978-1-4503-1247-9. DOI: 10.1145/2213836.2213891.

[268] D. Milne und I. H. Witten, „An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links“, in *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, Chicago, Illinois, USA: AAAI Press, Juli 2008, S. 25–30, ISBN: 978-1-57735-383-6.

[269] Y. Tian, D. Lo und J. Lawall, „Sewordsim: Software-Specific Word Similarity Database“, in *Proceedings of the 36th International Conference on Software Engineering*, Hyderabad, Indien: Association for Computing Machinery, Mai 2014, S. 568–571, ISBN: 978-1-4503-2768-8. DOI: 10.1145/2591062.2591071.

[270] D. Bollegala, Y. Matsuo und M. Ishizuka, „Measuring Semantic Similarity between Words Using Web Search Engines“, in *Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Kanada: Association for Computing Machinery, Mai 2007, S. 757–766, ISBN: 978-1-59593-654-7. DOI: 10.1145/1242572.1242675.

[271] N. Limsopatham und N. Collier, „Bidirectional LSTM for Named Entity Recognition in Twitter Messages“, in *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, Osaka, Japan: The COLING 2016 Organizing Committee, Dez. 2016, S. 145–152, ISBN: 978-4-87974-707-5. DOI: 10.17863/CAM.7201.

[272] L. Derczynski u. a., „Analysis of Named Entity Recognition and Linking for Tweets“, *Information Processing & Management*, Jg. 51, Nr. 2, S. 32–49, März 2015, ISSN: 03064573. DOI: 10.1016/j.ipm.2014.10.006.

[273] C.-H. Chang und S.-Y. Hwang, „A Word Embedding-Based Approach to Cross-Lingual Topic Modeling“, *Knowledge and Information Systems*, Jg. 63, Nr. 6, S. 1529–1555, Apr. 2021, ISSN: 0219-3116. DOI: 10.1007/s10115-021-01555-7.

[274] T. Zhang, K. Liu und J. Zhao, „Cross Lingual Entity Linking with Bilingual Topic Model“, in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*, Peking, China: AAAI Press, Aug. 2013, S. 2218–2224, ISBN: 978-1-57735-633-2.

[275] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith und A. McCallum, „Polylingual Topic Models“, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, Singapur, Singapur: Association for Computational Linguistics, Aug. 2009, S. 880–889, ISBN: 978-1-932432-62-6. DOI: 10.3115/1699571.1699627.

[276] X. Ni, J.-T. Sun, J. Hu und Z. Chen, „Mining Multilingual Topics from Wikipedia“, in *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, Madrid, Spanien: Association for Computing Machinery, Apr. 2009, S. 1155–1156, ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.1526904.

[277] S. Zoghbi, I. Vulić und M.-F. Moens, „Latent Dirichlet Allocation for Linking User-Generated Content and E-Commerce Data“, *Information Sciences*, Jg. 367, Nr. C, S. 573–599, Nov. 2016, ISSN: 0020-0255. DOI: 10.1016/j.ins.2016.05.047.

[278] G. Heyman, I. Vulić und M.-F. Moens, „C-Bilda Extracting Cross-Lingual Topics from Non-Parallel Texts by Distinguishing Shared from Unshared Content“, *Data Mining and Knowledge Discovery*, Jg. 30, Nr. 5, S. 1299–1323, Nov. 2015, ISSN: 1573-756X. DOI: 10.1007/s10618-015-0442-x.

[279] D. M. Blei und M. I. Jordan, „Modeling Annotated Data“, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*, Toronto, Kanada: Association for Computing Machinery, Juli 2003, S. 127–134, ISBN: 978-1-58113-646-3. DOI: 10.1145/860435.860460.

[280] R. Cai, M. Chen und H. Wang, „Nonparametric Symmetric Correspondence Topic Models for Multilingual Text Analysis“, in *Proceedings of the 4th Conference on Natural Language Processing and Chinese Computing (NLPCC)*, J. Li, H. Ji, D. Zhao und Y. Feng, Hrsg., Ser. Lecture Notes in Computer Science, Bd. 9362, Nanchang, China: Springer International Publishing, Okt. 2015, S. 270–281, ISBN: 978-3-319-25207-0. DOI: 10.1007/978-3-319-25207-0_23.

[281] Y. Sakata und K. Eguchi, „Relation Prediction in Multilingual Data Based on Multimodal Relational Topic Models“, *IEICE Transactions on Information and Systems*, Jg. E100.D, Nr. 4, S. 741–749, Apr. 2017, ISSN: 1745-1361. DOI: 10.1587/transinf.2016DAP0021.

[282] I. H. Musa, K. Xu und I. Zamit, „Multilingual Document Concept Topic Modeling“, in *Proceedings of the European Conference on Natural Language Processing and Information Retrieval (ECNLPIR)*, Hangzhou, China: IEEE Computer Society, Juli 2022, S. 84–91, ISBN: 978-1-66547-382-8. DOI: 10.1109/ECNLPIR57021.2022.00027.

[283] K. Asnani und J. D. Pawar, „Automatic Aspect Extraction using Lexical Semantic Knowledge in Code-Mixed Context“, *Procedia Computer Science*, Jg. 112, Nr. C, S. 693–702, Sep. 2017, ISSN: 1877-0509. DOI: 10.1016/j.procs.2017.08.146.

[284] S. Hao und M. J. Paul, „Learning Multilingual Topics from Incomparable Corpora“, in *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, S. 2595–2609, ISBN: 978-1-948087-50-6.

[285] T. Ma und T. Nasukawa, „Inverted Bilingual Topic Models for Lexicon Extraction from Non-parallel Data“, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'17)*, Melbourne, Australia: AAAI Press, Aug. 2017, S. 4075–4081, ISBN: 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/569.

[286] I. H. Musa, K. Xu, F. Liu, I. Zamit, W. A. Abro und G. Qi, „A Cross-Lingual Sentiment Topic Model Evolution Over Time“, *Intelligent Data Analysis*, Jg. 24, Nr. 2, S. 253–266, Jan. 2020, ISSN: 1088-467X. DOI: 10.3233/IDA-184449.

[287] W. Yang, J. Boyd-Graber und P. Resnik, „A Multilingual Topic Model for Learning Weighted Topic Links Across Corpora with Low Comparability“, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, S. 1243–1248, ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/D19-1120.

[288] M. Yuan, B. Van Durme und J. L. Ying, „Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages“, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*, Bd. 31, Montreal, Kanada: Curran Associates, Dez. 2018, S. 8667–8677.

[289] Y. Hu, K. Zhai, V. Eidelman und J. Boyd-Graber, „Polylingual Tree-Based Topic Models for Translation Domain Adaptation“, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA: Association for Computational Linguistics, Juni 2014, S. 1166–1176, ISBN: 978-1-937284-72-5. DOI: 10.3115/v1/P14-1110.

[290] T. Piccardi und R. West, „Crosslingual Topic Modeling with WikiPDA“, in *Proceedings of the Web Conference (WWW'21)*, Ljubljana, Slowenien: Association for Computing Machinery, Juni 2021, S. 3032–3041, ISBN: 978-1-4503-8312-7. DOI: 10.1145/3442381.3449805.

[291] C.-H. Chan u. a., „Reproducible Extraction of Cross-lingual Topics (rectr)“, *Communication Methods and Measures*, Jg. 14, Nr. 4, S. 285–305, Okt. 2020, ISSN: 1931-2458, 1931-2466. DOI: 10.1080/19312458.2020.1812555.

[292] C.-H. Chang, S.-Y. Hwang und T.-H. Xui, „Incorporating Word Embedding into Cross-Lingual Topic Modeling“, in *Proceedings of the International Congress on Big Data*, San Francisco, Kalifornien, USA: IEEE, Juli 2018, S. 17–24, ISBN: 978-1-5386-7232-7. DOI: 10.1109/BigDataCongress.2018.00010.

[293] X. Li, Z. Zeng, J. Zhang und S. Jiang, „Chinese-Thai Cross-Language Topic Extraction and Alignment“, in *Proceedings of the International Conference on Asian Language Processing (IALP)*, Singapur, Singapur: IEEE, Dez. 2017, S. 239–242, ISBN: 978-1-5386-1981-0. DOI: 10.1109/IALP.2017.8300588.

[294] Q. Xie, X. Zhang, Y. Ding und M. Song, „Monolingual and Multilingual Topic Analysis Using LDA and BERT Embeddings“, *Journal of Informetrics*, Jg. 14, Nr. 3, S. 1–16, Aug. 2020, ISSN: 17511577. DOI: 10.1016/j.joi.2020.101055.

[295] F. Bianchi, S. Terragni, D. Hovy, D. Nozza und E. Fersini, „Cross-lingual Contextualized Topic Models with Zero-shot Learning“, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Apr. 2021, S. 1676–1683, ISBN: 978-1-954085-02-2. DOI: 10.18653/v1/2021.eacl-main.143.

[296] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou und T. Mikolov, *FastText.Zip: Compressing Text Classification Models*, Dez. 2016. DOI: 10.48550/arXiv.1612.03651. arXiv: 1612.03651 [cs].

[297] X. Wang, M.-C. Chang, L. Wang und S. Lyu, „Efficient Algorithms for Graph Regularized PLSA for Probabilistic Topic Modeling“, *Pattern Recognition*, Jg. 86, S. 236–247, Feb. 2019, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2018.09.004.

[298] D. Jiang, Y. Tong und Y. Song, „Cross-Lingual Topic Discovery From Multilingual Search Engine Query Log“, *ACM Transactions on Information Systems*, Jg. 35, Nr. 2, S. 1–28, Apr. 2017, ISSN: 1046-8188, 1558-2868. DOI: 10.1145/2956235.

[299] S. Hao und M. J. Paul, „An Empirical Study on Crosslingual Transfer in Probabilistic Topic Models“, *Computational Linguistics*, Jg. 46, Nr. 1, S. 95–134, März 2020, ISSN: 0891-2017, 1530-9312. DOI: 10.1162/coli_a_00369.

[300] S. Hao, J. Boyd-Graber und M. J. Paul, *Lessons from the Bible on Modern Topics: Low-Resource Multilingual Topic Model Evaluation*, Apr. 2018. DOI: 10.48550/arXiv.1804.10184. arXiv: 1804.10184 [cs].

[301] W. De Smet, J. Tang und M.-F. Moens, „Knowledge Transfer across Multilingual Corpora via Latent Topics“, in *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'11)*, J. Z. Huang, L. Cao und J. Srivastava, Hrsg., Ser. Lecture Notes in Computer Science, Shenzhen, China: Springer Science+Business Media, Mai 2011, S. 549–560, ISBN: 978-3-642-20841-6. DOI: 10.1007/978-3-642-20841-6_45.

[302] D. Vilariño, D. Pinto, B. Beltrán, S. León, E. Castillo und M. Tovar, „A Machine-Translation Method for Normalization of SMS“, in *Proceedings of the 4th Mexican Conference on Pattern Recognition*, D. Hutchison u. a., Hrsg., Ser. Lecture Notes in Computer Science, Huatulco, Mexiko: Springer Science+Business Media, Juni 2012, S. 293–302, ISBN: 978-3-642-31148-2 978-3-642-31149-9. DOI: 10.1007/978-3-642-31149-9.

[303] S. Arora u. a., „A Practical Algorithm for Topic Modeling with Provable Guarantees“, in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta und D. McAllester, Hrsg., Ser. Proceedings of Machine Learning Research, Bd. 28, Atlanta, Georgia, USA: MIT Press, Juni 2013, S. 280–288.

[304] J. Lund, C. Cook, K. Seppi und J. Boyd-Graber, „Tandem Anchoring: A Multiword Anchor Approach for Interactive Topic Modeling“, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada: Association for Computational Linguistics, Juli 2017, S. 896–905, ISBN: 978-1-945626-75-3. DOI: 10.18653/v1/P17-1083.

[305] M. Aliramezani, E. Doostmohammadi, M. H. Bokaei und H. Sameti, „Persian Sentiment Analysis Without Training Data Using Cross-Lingual Word Embeddings“, in *Proceedings of the 10th International Symposium onTelecommunications (IST)*, Tehran, Iran: IEEE, Dez. 2020, S. 78–82, ISBN: 978-1-72818-012-0. DOI: 10.1109/IST50524.2020.9345882.

[306] P. H. Schönemann, „A Generalized Solution of the Orthogonal Procrustes Problem“, *Psychometrika*, Jg. 31, Nr. 1, S. 1–10, März 1966, ISSN: 1860-0980. DOI: 10.1007/BF02289451.

[307] G. Lample, A. Conneau, M. Ranzato, L. Denoyer und H. Jégou, „Word Translation Without Parallel Data“, in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Kanada: ICLR Press, Feb. 2018, S. 1–14.

[308] M. Röder, A. Both und A. Hinneburg, „Exploring the Space of Topic Coherence Measures“, in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, Shanghai, China: Association for Computing Machinery, Feb. 2015, S. 399–408, ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685324.

[309] N. Aletras und M. Stevenson, „Evaluating Topic Coherence Using Distributional Semantics“, in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, Potsdam, Deutschland: Association for Computational Linguistics, März 2013, S. 13–22.

[310] E. M. Airoldi und J. M. Bischof, „Improving and Evaluating Topic Models and Other Models of Text“, *Journal of the American Statistical Association*, Jg. 111, Nr. 516, S. 1381–1403, Okt. 2016, ISSN: 0162-1459. DOI: 10.1080/01621459.2015.1051182.

[311] J. M. Bischof und E. M. Airoldi, „Summarizing Topical Content with Word Frequency and Exclusivity“, in *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML'12)*, Edinburgh, Schottland, Großbritannien: Omnipress, Juni 2012, S. 9–16, ISBN: 978-1-4503-1285-1.

[312] D. Johnson und S. Sinanovic, „Symmetrizing the Kullback-Leibler Distance“, *IEEE Transactions on Information Theory*, S. 1–10, März 2001.

[313] M. Yuan, P. Lin, L. Rashidi und J. Zobel, „Assessment of the Quality of Topic Models for Information Retrieval Applications“, in *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23)*, Taipei, Taiwan: Association for Computing Machinery, Aug. 2023, S. 265–274, ISBN: 979-8-4007-0073-6. DOI: 10.1145/3578337.3605118.

[314] D. M. Blei und J. D. Lafferty, „A Correlated Topic Model of Science“, *The Annals of Applied Statistics*, Jg. 1, Nr. 1, S. 17–35, Juni 2007, ISSN: 1932-6157. DOI: 10.1214/07-AOAS114. JSTOR: 4537420.

[315] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang und D. Blei, „Reading Tea Leaves: How Humans Interpret Topic Models“, in *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09)*, Vancouver, Kanada: Curran Associates, Dez. 2009, S. 288–296, ISBN: 978-1-61567-911-9.

[316] T.-N. Doan und T.-A. Hoang, „Benchmarking Neural Topic Models: An Empirical Study“, in *Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021)*, Online: Association for Computational Linguistics, Aug. 2021, S. 4363–4368, ISBN: 978-1-954085-54-1. DOI: 10.18653/v1/2021.findings-acl.382.

[317] P. Tijare und J. Rani, „Exploring Popular Topic Models“, *Journal of Physics: Conference Series*, Jg. 1706, Nr. 1, S. 1–11, Aug. 2020, ISSN: 1742-6596. DOI: 10.1088/1742-6596/1706/1/012171.

[318] J. H. Lau, D. Newman und T. Baldwin, „Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality“, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Schweden: Association for Computational Linguistics, Apr. 2014, S. 530–539, ISBN: 978-1-937284-78-7. DOI: 10.3115/v1/E14-1056.

[319] J. Vosecky, D. Jiang, K. W.-T. Leung, K. Xing und W. Ng, „Integrating Social and Auxiliary Semantics for Multifaceted Topic Modeling in Twitter“, *ACM Transactions on Internet Technology*, Jg. 14, Nr. 4, S. 1–24, Dez. 2014, ISSN: 1533-5399, 1557-6051. DOI: 10.1145/2651403.

[320] G. H. Ball und D. J. Hall, „ISODATA, a Novel Method of Data Analysis and Pattern Classification“, Defense Technical Information Center, Stanford, Kalifornien, USA, Technical Report, Apr. 1965, S. 1–61.

[321] M. Ester, H.-P. Kriegel, J. Sander und X. Xu, „A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise“, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, Portland, Oregon, USA: AAAI Press, Aug. 1996, S. 226–231.

[322] J. Riesa, *Compact Language Detector v3 (Cld3)*, https://github.com/google/cld3, Aug. 2023. (besucht am 10. 06. 2023).

[323] J. Riesa und I. Giuliani, *Compact Language Detector 2*, https://github.com/CLD2Owners/cld2, Aug. 2023. (besucht am 10. 06. 2023).

[324] M. Lui und T. Baldwin, „Accurate Language Identification of Twitter Messages“, in *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, Gothenburg, Schweden: Association for Computational Linguistics, Apr. 2014, S. 17–25, ISBN: 978-1-937284-92-3. DOI: 10.3115/v1/W14-1303.

[325] G. Diaz, *Stopwords ISO*, https://github.com/stopwords-iso/stopwords-iso, Sep. 2020. (besucht am 10. 06. 2023).

[326] Lingolia, *Die 50 wichtigsten Verben im Deutschen*, https://deutsch.lingolia.com/de/50-verben-deutsch, Okt. 2023. (besucht am 28. 09. 2023).

[327] S. Nazim, *Redewendungen*, https://hingabe.at/redewendungen/, Okt. 2023. (besucht am 29. 09. 2023).

[328] C. Benjamine, *2023's most used internet abbreviations for tweeting and texting*, https://preply.com/en/blog/the-most-used-internet-abbreviations-for-texting-and-tweeting/, Mai 2020. (besucht am 29. 09. 2023).

[329] H. Schmid, „Probabilistic Part-of-Speech Tagging Using Decision Trees“, in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, Großbritannien: Routledge, 1994, S. 154–164, ISBN: 1-85728-711-8.

[330] H. Schmid, „Improvements in Part-of-Speech Tagging with an Application to German“, in *Natural Language Processing Using Very Large Corpora*, Ser. Text, Speech and Language Technology 11, N. Ide u. a., Hrsg., Dordrecht, Deutschland: Springer Science+Business Media, 1999, S. 13–25, ISBN: 978-90-481-5349-7 978-94-017-2390-9. DOI: 10.1007/978-94-017-2390-9_2.

[331] L. Hickman, S. Thapa, L. Tay, M. Cao und P. Srinivasan, „Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations“, *Organizational Research Methods*, Jg. 25, Nr. 1, S. 114–146, Jan. 2022, ISSN: 1094-4281. DOI: 10.1177/1094428120971683.

[332] S. J. Weston, I. Shryock, R. Light und P. A. Fisher, „Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial“, *Advances in Methods and Practices in Psychological Science*, Jg. 6, Nr. 2, S. 1–13, Apr. 2023, ISSN: 2515-2459, 2515-2467. DOI: 10.1177/25152459231160105.

[333] M. E. Roberts, B. M. Stewart und D. Tingley, „STM: An R Package for Structural Topic Models“, *Journal of Statistical Software*, Jg. 91, Nr. 2, S. 1–40, Okt. 2019, ISSN: 1548-7660. DOI: 10.18637/jss.v091.i02.

[334] D. Kokkinakis, R. M. Sánchez, S. Bruinsma und M.-M. Hammarlin, „Investigating the Effects of MWE Identification in Structural Topic Modelling“, in *Proceedings of the 19th Workshop on Multiword Expressions*, Dubrovnik, Kroatien: Association for Computational Linguistics, Mai 2023, S. 36–44, ISBN: 978-1-959429-59-3. DOI: 10.18653/v1/2023.mwe-1.7.

[335] J. D. Lee und K. Kolodge, „Exploring Trust in Self-Driving Vehicles Through Text Analysis“, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Jg. 62, Nr. 2, S. 260–277, Sep. 2019, ISSN: 0018-7208, 1547-8181. DOI: 10.1177/0018720819872672.

[336] R. A. Rutkowski, J. D. Lee, R. J. Coller und N. E. Werner, „How Can Text Mining Support Qualitative Data Analysis?“, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Bd. 66, Atlanta, Georgia, USA: Sage Publications, Sep. 2022, S. 2319–2323. DOI: 10.1177/1071181322661535.

[337] M. Endres, P. Roocks und W. Kießling, „Scalagon: An Efficient Skyline Algorithm for All Seasons“, in *Proceedings of the 20th International Conference on Database Systems for Advanced Applications*, M. Renz, C. Shahabi, X. Zhou und M. A. Cheema, Hrsg., Bd. 9050, Hanoi, Vietnam: Springer International Publishing, Apr. 2015, S. 292–308, ISBN: 978-3-319-18122-6 978-3-319-18123-3. DOI: 10.1007/978-3-319-18123-3_18.

[338] P. Roocks, „Computing Pareto Frontiers and Database Preferences with the rPref Package“, *The R Journal*, Jg. 8, Nr. 2, S. 393–404, Dez. 2016, ISSN: 2073-4859. DOI: 10.32614/RJ-2016-054.

[339] T. Preisinger und W. Kiessling, „The Hexagon Algorithm for Pareto Preference Queries“, in *Proceedings of 3rd Multidisciplinary Workshop on Advances in Preference Handling in conjunction with VLDB*, Wien, Oesterreich, 2007.

[340] N. Kratzke, *Monthly Samples of German Tweets*, https://zenodo.org/record/6624514, Juni 2022. (besucht am 23. 09. 2023).

[341] F. Poldi, *TWINT - Twitter Intelligence Tool*, https://github.com/twintproject/twint, Sep. 2023. (besucht am 23. 09. 2023).

[342] F. de Saussure, *Cours de linguistique generale. 1: Reprod. de l'ed. originale*. Wiesbaden: Harrassowitz, 1989, ISBN: 978-3-447-00798-6.

[343] C. Biemann, G. Heyer und U. Quasthoff, *Wissensrohstoff Text: Eine Einführung in das Text Mining*, 2. Aufl. Wiesbaden, Deutschland: Springer Naturre, 2022, ISBN: 978-3-658-35968-3 978-3-658-35969-0. DOI: 10.1007/978-3-658-35969-0.

[344] G. A. Miller und W. G. Charles, „Contextual Correlates of Semantic Similarity“, *Language and Cognitive Processes*, Jg. 6, Nr. 1, S. 1–28, Jan. 1991, ISSN: 0169-0965. DOI: 10.1080/01690969108406936.

[345] G. Heyer, U. Quasthoff und T. Wittig, *Text Mining: Wissensrohstoff Text - Konzepte, Algorithmen, Ergebnisse* (Informatik), 2. Nachdr. Herdecke: W3L-Verl, 2012, ISBN: 978-3-937137-30-8.

[346] C. Biemann und M. Riedl, „Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity“, *Journal of Language Modelling*, Jg. 1, Nr. 1, S. 55–95, Juli 2013, ISSN: 2299-8470. DOI: 10.15398/jlm.v1i1.60.

[347] S. Bordag, „A Comparison of Co-occurrence and Similarity Measures as Simulations of Context“, in *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'08)*, A. Gelbukh, Hrsg., Ser. Lecture Notes in Computer Science, Bd. 4919, Haifa, Israel: Springer Science+Business Media, Feb. 2008, S. 52–63, ISBN: 978-3-540-78134-9 978-3-540-78135-6. DOI: 10.1007/978-3-540-78135-6_5.

[348] T. Dunning, „Accurate Methods for the Statistics of Surprise and Coincidence“, *Computational Linguistics*, Jg. 19, Nr. 1, S. 61–74, 1993.

[349] P. G. Otero, „Comparing Different Properties Involved in Word Similarity Extraction“, in *Proceedings of the 14th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence (EPIA '09)*, L. S. Lopes, N. Lau, P. Mariano und L. M. Rocha, Hrsg., Ser. Lecture Notes in Artificial Intelligence, Aveiro, Portugal: Springer Berlin Heidelberg, Okt. 2009, S. 634–645, ISBN: 978-3-642-04685-8 978-3-642-04686-5. DOI: 10.1007/978-3-642-04686-5_52.

[350] G. Lapesa, S. Evert und S. S. Im Walde, „Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models“, in *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (SEM 2014)*, Dublin, Irland: Association for Computational Linguistics und Dublin City University, Aug. 2014, S. 160–170.

[351] P. Gamallo und S. Bordag, „Is Singular Value Decomposition Useful for Word Similarity Extraction?“, *Language Resources and Evaluation*, Jg. 45, Nr. 2, S. 95–119, Mai 2011, ISSN: 1574-0218. DOI: 10.1007/s10579-010-9129-5.

[352] C.-h. Chan und M. Sältzer, „Oolong: An R Package for Validating Automated Content Analysis Tools“, *Journal of Open Source Software*, Jg. 5, Nr. 55, S. 2461–2470, Nov. 2020, ISSN: 2475-9066. DOI: 10.21105/joss.02461.

[353] A. Edwards, „R.A. Fischer, Statistical Methods for Research Workers“, in *Landmark Writings in Western Mathematics 1640-1940*, 1. Aufl., Amsterdam, Niederlande: Elsevier, 2005, S. 856–870, ISBN: 978-0-444-50871-3. DOI: 10.1016/B978-044450871-3/50148-0.

[354] M. G. Kendall, „A New Measure of Rank Correlation“, *Biometrika*, Jg. 30, Nr. 1/2, S. 81–93, Juni 1938, ISSN: 00063444. DOI: 10.2307/2332226. JSTOR: 2332226.

[355] H. L. Costner, „Criteria for Measures of Association“, *American Sociological Review*, Jg. 30, Nr. 3, S. 341–353, Juni 1965, ISSN: 0003-1224. DOI: 10.2307/2090715. JSTOR: 2090715.

[356] V. Kumar, A. Smith-Renner, L. Findlater, K. Seppi und J. Boyd-Graber, „Why Didn't You Listen to Me? Comparing User Control of Human-in-the-Loop Topic Models“, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florenz, Italien: Association for Computational Linguistics, Juli 2019, S. 6323–6330, ISBN: 978-1-950737-48-2. DOI: 10.18653/v1/P19-1637. arXiv: 1905.09864 [cs].

[357] D. Andrzejewski und X. Zhu, „Latent Dirichlet Allocation with Topic-in-Set Knowledge“, in *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, Boulder, Colorado, USA: Association for Computational Linguistics, Juni 2009, S. 43–48, ISBN: 978-1-932432-38-1.

[358] F. Martin und M. Johnson, „More Efficient Topic Modelling Through a Noun Only Approach“, in *Proceedings of the Australasian Language Technology Association Workshop*, Bd. 13, Parramatta, Australien: ALTA, Dez. 2015, S. 111–115, ISBN: 1834-7037.

[359] C. D. P. Laureate, W. Buntine und H. Linger, „A Systematic Review of the Use of Topic Models for Short Text Social Media Analysis“, *Artificial Intelligence Review*, Jg. 56, Nr. 12, S. 14 223–14 255, Mai 2023, ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-023-10471-x.

[360] G. Brookes und T. McEnery, „The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation“, *Discourse Studies*, Jg. 21, Nr. 1, S. 3–21, Dez. 2018, ISSN: 1461-4456, 1461-7080. DOI: 10.1177/1461445618814032.

[361] G. Xun, V. Gopalakrishnan, F. Ma, Y. Li, J. Gao und A. Zhang, „Topic Discovery for Short Texts Using Word Embeddings“, in *Proceedings of The16th International Conference on Data Mining (ICDM)*, Barcelona, Spanien: IEEE Computer Society, Dez. 2016, S. 1299–1304, ISBN: 978-1-5090-5474-9. DOI: 10.1109/ICDM.2016.0176.

Eidesstattliche Erklärung

Hiermit versichere ich – Jenny Maria Felser – an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach Publikationen oder Vorträgen anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt oder anderweitig veröffentlicht.

Mittweida, 22. November 2023 [redacted]

Ort, Datum Jenny Maria Felser, B.Sc.