Digitale, automatisierte Analyse von Literaturdatenbanken in Public Health und Pflegewissenschaft – Quantitative Textanalyse großer Ergebnismengen mittels Topic Modeling

Eine Darstellung am Beispiel neuer Technologien in der Pflege

Dominik Domhoff
Universität Bremen
Bremen, Deutschland
ddomhoff@uni-bremen.de

Tobias Krick
Universität Bremen
Bremen, Deutschland
tkrick@uni-bremen.de

Kai Huter
Universität Bremen
Bremen, Deutschland
huter@uni-bremen.de

Abdallah El Ali
Centrum Wiskunde & Informatica
Amsterdam, Niederlande
abdallah.el.ali@cwi.nl

Tim Claudius Stratmann
Universität Oldenburg
Oldenburg, Deutschland
tim.claudius.stratmann@uol.de

Karin Wolf-Ostermann
Universität Bremen
Bremen, Deutschland
wolf-ostermann@uni-bremen.de

Heinz Rothgang
Universität Bremen
Bremen, Deutschland
rothgang@uni-bremen.de

Abstract—Systematische Literaturrecherchen stellen in den Gesundheitswissenschaften ein wichtiges Instrument zur Beurteilung des Forschungsstandes dar. Durch den stark zunehmenden Bestand an verfügbarer Literatur sind bei Wahrung anerkannter wissenschaftlicher Standards hohe zeitliche und personelle Ressourcen zur Durchführung dieser Literaturanalysen notwendig. In dieser Arbeit wird am Beispiel einer Recherche zum Thema innovativer Technologien in der Pflege gezeigt, welche Möglichkeiten die Methode des Topic Modeling zum automatisierten Clustern und inhaltlichen Erschließen wissenschaftlicher Literatur bietet. Es konnten 109.598 englischsprachige Titel und Abstracts aus sieben Datenbanken eingeschlossen und in 35 thematische Cluster eingeordnet werden. Als prominente Technologiebereiche in der Pflege ließen sich die elektronische Dokumentation, Monitoring & Sensorik, Sturzerkennung, Diabetes-Therapie, Wundversorgung und Robotik identifizieren. Es zeigte sich, dass Topic Modeling unter verringertem Ressourceneinsatz die Strukturierung wissenschaftlicher Literatur unterstützen kann. Für den Einsatz im Rahmen hochqualitativer systematischer Reviews ist eine Weiterentwicklung der Methodik und die Durchführung von Validierungsstudien notwendig.

Keywords—Review, Literaturdatenbanken, innovative Technologien, Pflege, Topic Modeling, Latent Dirichlet Allocation

I. EINLEITUNG

Systematische Literaturrecherchen sind in den Gesundheitswissenschaften unverzichtbar, um einen validen Überblick über eine bestehende Studienlage zu gewinnen. Sie dienen als Ausgangspunkt für weitergehende Forschung, sind entscheidend für die Umsetzung von Interventionen und Therapien und beeinflussen politische Entscheidungen. Durch einen rasant wachsenden Bestand an wissenschaftlicher Literatur können vorgefundene Ergebnismengen jedoch die für die Durchführung einer hochwertigen systematischen Übersichtsarbeit verfügbaren personellen und zeitlichen Ressourcen übersteigen. Dieses Problem betrifft auch andere Arten von Übersichtsarbeiten, welche ein Vier-Augen-Prinzip im Prozess

der Sichtung zur Wahrung der wissenschaftlichen Qualität erfordern.

II. Hintergrund und Zielsetzung

Im folgenden Beitrag wird dargestellt, wie die Methode des Topic Modeling eingesetzt werden kann, um automatisiert große Mengen wissenschaftlicher Publikationen in Cluster einzuteilen und auszuwerten. Dies erfolgt am Beispiel einer systematischen Literaturrecherche mit der inhaltlichen Fragestellung, welche innovativen Technologien und Lösungen im Bereich der Pflege in der wissenschaftlichen Literatur beschrieben werden. Aufgrund der großen Menge der einzubeziehenden Literatur mit mehr als 100.000 Treffern erweist sich hier die etablierte Vorgehensweise eines händischen Screenings nicht mehr als realisierbar, so dass die Zweckmäßigkeit einer technisch unterstützen Auswertung durch Topic Modeling ausgelotet werden soll.

III. Methoden

Im Folgenden werden zunächst die Such- und Recherchestrategie der zugrundeliegenden Literaturrecherche und notwendige Schritte zur Aufbereitung der Daten beschrieben. Anschließend wird die Analyse im Rahmen des Topic Modeling dargestellt.

Suchstrategie

Basierend auf der Forschungsfrage wurden Ein- und Ausschlusskriterien für die Literaturanalyse definiert. Eingeschlossen wurden demnach empirische Arbeiten in englischer Sprache. Ausgeschlossen wurden Technologien, die im Kontext des zugrundeliegenden Forschungsprojektes nicht von Interesse waren. Dies waren insbesondere biologische, chemische und chirurgische Technologien. Es erfolgte keine zeitliche Einschränkung bezüglich des Publikationsdatums.

Es wurde im *Block Building Approach* eine Suchstrategie entworfen, welche die Bereiche *Pflege* und *Technik* umfasste. Für beide Themengebiete wurden einschlägige Begriffe und Synonyme identifiziert. Für die ausgeschlossenen Gebiete wurden Begriffe mit größtmöglicher Spezifität gewählt. Die so erzeugte Suchanfrage lautete:

(care OR caring OR nursing)
AND
(technol* OR robot* OR intelligent OR smart OR assistive OR "decision support system" OR "ambient assisted living" OR sensor OR wearable OR "virtual reality" OR "mixed reality" OR tagging OR tracking OR "remote health monitoring" OR "fall detection" OR "human computer interaction" OR "human machine interaction" OR gerontotechnology OR gerontechnology OR "head mounted display" OR exoskeleton OR "augmented reality" OR "biomedical monitoring")
NOT
(biotech* OR nano* OR chromosome OR "sequencing technologies" OR polymerase OR RNA OR DNA OR enzym* OR "shampoo surfactant technologies" OR transplantation OR "radiotherapy technology" OR electrochemical OR molecul* OR "biological engineering" OR particle* OR chemical* OR prosthesis OR tomogra* OR agricultur* OR biomolecul* OR reproduct* OR lab-on-fiber OR cytolo* OR "tissue Engineering" OR genet* OR genom* OR pharmacogenomic OR gene-therapy OR proton OR neutron OR sequencing OR in-vitro)

Es wurden die Datenbanken Medline, Scopus, CINAHL, Cochrane Library, ACM Digital Library, IEEE Xplore und The Collection of Computer Science Bibliographies durchsucht. Der Suchterm wurde auf die Syntax der jeweiligen Suchmaschine angepasst. Die Suche erfolgte jeweils in Titel und Abstract. Insgesamt wurden 150.261 Treffer aus den Datenbanken exportiert.

Datenaufbereitung

Alle Ergebnisse wurden zunächst in Endnote importiert. Die Suchergebnisse wurden um Duplikate (n=25.595) bereinigt und Artikel ohne Abstract (n=7.503) wurden ausgeschlossen. Alle weiteren Schritte wurden mit der Programmiersprache Python in der Version 3.6.4 durchgeführt. Da auch nicht-englischsprachige Artikel in den Suchergebnisse vorhanden waren, wurde mittels Googles *language-detection* [1] die Sprache im Abstract ermittelt und nicht-englischsprachige Artikel (n=8.017) ausgeschlossen. Somit gingen 109.598 englischsprachige Artikel in die Auswertung ein.

Die automatisierte Analyse der Artikel erfolgte auf Basis von Titel und Abstract. Diese wurden zunächst verkettet und Stoppworte, die keinen Informationsgewinn aufweisen, wurden an Hand der Liste aus dem Programmppaket *nltk* [2], welches Werkzeuge für die maschinelle Verarbeitung natürlicher Sprache liefert, entfernt. Folgend mussten die Dokumente in ihre einzelnen Worte zerlegt werden (*bag-of-words*). Um in den Ergebnissen jedoch auch feststehende Begriffe identifizieren zu können, die aus mehr als einem Wort bestehen, wurden jene zwei- und dreifachen Wortkombinationen, die im gesamten Datensatz mehr als 100 mal vorkamen, mit der Phrasen-Erkennung des Paketes *gensim* [3], welches die Funktionen für das Topic Modeling liefert, jeweils zu zusammenhängenden Wortkombinationen transformiert. Abschließend wurden alle Worte mit dem *WordNet Lemmatizer* aus dem *nltk*-Paket [2] auf ihren Wordstamm reduziert. Hierdurch wird beispielsweise der Plural eines Wortes in den Singular überführt oder die Beugung eines Wortes entfernt, sodass Worte mit gleicher Bedeutung eine einheitliche Repräsentation aufweisen.

Topic Modeling

Zur Modellierung der dem Dokumentenbestand zu Grunde liegenden Themengebiete nutzten wir die Latent Dirichlet Allocation (LDA) [4] mit dem Programmpaket *gensim* [3]. Ausgangspunkt von LDA ist, dass Dokumenten latent Anteile verschiedener Themengebiete (*Topics*) zugrunde liegen, welche selbst durch bestimmte Wörter charakterisiert sind. Aus häufig zusammen in einem Dokument vorkommenden Begriffen werden dann thematische Cluster erzeugt. LDA erfordert dabei vorab die manuelle Festlegung der Anzahl von Themen im zu analysierenden Dokumentenkorpus. Da diese Information – insbesondere in Anbetracht der Forschungsfrage – nicht zur Verfügung stand, wurden mit *tmtoolkit [5]* Modelle für unterschiedliche Anzahlen von 2 bis 100 Themengebiete erzeugt und an Hand von vier verschiedenen, in der Literatur beschrieben Modellgütekriterien bewertet. Dies waren *Topic Coherence*-Maße nach Mimno et al. [6] und Röder et al. [7], die Methode nach Cao et al. [8], sowie die Loglikelihood der

Modelle. Da sich hierbei keine eindeutigen – zum Teil sogar gegenläufige – Hinweise auf eine statistisch korrekte Anzahl von Themengebieten, ergaben, wurde das Modell durch Bewertung und Abstimmung im Forschungsteam ausgewählt. Als Kriterium wurde dabei eine größtmögliche inhaltliche Konsistenz und Trennschärfe der Topics zugrunde gelegt. Es ergab sich ein einstimmiger Konsens für ein Modell mit 35 Topics.

Die LDA-Modellierung erfolgte mit *gensim* [3] in der Multicore-Version als online Training über den gesamten Korpus von 109.598 Dokumente bei 50 Durchgängen zur Erzielung von Konvergenz. Die Visualisierung zur Beurteilung der Modelle und zur Darstellung der relevantesten Begriffe für die identifizierten Themengebiete erfolgte mit *pyLDAvis* [9]. Dieses ermöglich die grafische Darstellung der Topics mit multidimensionaler Skalierung und gibt die relevanten Wörter je Topic aus. Hierbei wurde ein λ von 0,2 gewählt, um insbesondere die für das jeweilige Thema relevanten Wörter und Wortkombinationen zu erhalten. Auf diese Weise können Begriffe erhalten werden, die seltener im gesamten Textkorpus vorkommen, jedoch besonders häufig im jeweiligen Topic vorhanden sind [9]. Jedem Themengebiet wurde händisch auf Basis der 30 relevantesten Begriffe ein Titel zugewiesen, welcher im Forschungsteam konsentiert wurde.

Für alle in die Modellierung eingegangenen Dokumente wurde schließlich der Anteil des Dokumentes an jedem der 35 Themengebiete berechnet. Da jedes Dokument in LDA als eine Mischung aus verschiedenen Themen betrachtet wird, besitzt jedes Dokument Anteile aus verschiedenen Topics, welche sich für jedes Dokument auf 1 aufsummieren. Als primäres Thema wurde jenes definiert, welches den höchsten Anteil am Dokument aufwies.

Für einige der Themengebiete zeigte sich anhand der wichtigsten Begriffe eine hohe Relevanz für die Fragestellung, die identifizierten relevanten Begriffe ermöglichten jedoch keine Aussage zu den eingesetzten Technologien. Dies betrifft insbesondere das Cluster „Senioren, Pflegeheime, Demenz“. Für dieses Cluster wurde eine Subanalyse durchgeführt. Hierfür wurden alle Dokumente mit einem Anteil von mehr als 20 % am entsprechenden Topic als Teilmenge gemäß der zuvor dargestellten Methodik erneut untersucht.

IV. ERGEBNISSE

Die Titel der 35 Themengebiete mit der Häufigkeit, wie oft diese als primäres Topic identifiziert wurden und der Anzahl von Dokumenten mit einem Anteil von mehr als 20 % für dieses Thema sind in Tabelle I dargestellt. In der Gesamtschau zeigen sich stark unterschiedliche Themengebiete, von denen viele als nicht relevant für die Forschungsfrage erachtet werden können, wie etwa die Topics zu klinischer Medizin (#34), medizinischer Bildgebung (#2) oder Softwarearchitektur (#4). Zudem bewegen sich die Themen auf sehr unterschiedlichen Abstraktionsebenen. So fokussieren einige der gebildeten Topics auf konkrete Krankheitsbilder wie Diabetes (#19) oder Krebserkrankungen (#23), während andere Dokumente eher auf Grund des Studientyps zu Clustern (#3, #29, #35) zusammengeführt wurden.

TABELLE I. IDENTIFIZIERTE TOPICS

Topic	Inhalt	primäres Topic *Anteil*	primäres Topic *Anzahl*	Anzahl P > 20 %[a]
#12	Rechtliche und organisatorische Rahmenbedingungen	12,0 %	13.116	19.702
#34	klinische Medizin, Leitlinien	7,1 %	7.744	13.713
#4	Softwarearchitektur	6,7 %	7.293	12.017
#20	Statistische Methoden & Algorithmen	4,5 %	4.882	7.904
#32	Funk-/Elektrotechnik	4,4 %	4.794	7.150
#17	Elektronische Dokumentation/Patientenakten	4,3 %	4.728	8.193
#35	(qualitative) (Pflege-)Forschung	4,3 %	4.702	8.449
#1	Monitoring, Sensoren, Wearables	4,3 %	4.682	7.608
#33	Senioren, Pflegeheime, Demenz	4,3 %	4.665	7.383
#15	Telemedizin: technische Bezüge, mobile Kommunikationstechnik, Big Data	4,2 %	4.592	8.263
#14	Pflegebildung	3,7 %	4.092	6.048
#3	quantitative Forschung, Vergleichsstudien	3,5 %	3.852	7.252
#30	Ethik	3,1 %	3.380	6.365
#21	Netzwerke, Sicherheit	3,1 %	3.376	5.138
#9	Chirurgie	2,6 %	2.845	4.188
#8	Interventionen, Risikofaktoren: Rauchen, Übergewicht, Therapieadhärenz	2,6 %	2.798	4.348
#28	Stationäre (Notfall-)Versorgung	2,5 %	2.772	4.860
#2	Medizin, Zahnmedizin, Bildgebung	2,3 %	2.543	4.912
#22	Blutwerte/-bestandteile/-analyse	2,3 %	2.489	3.976
#27	Reha-Technik, Sturz-/Gangerkennung	2,1 %	2.340	3.936
#29	Systematische Übersichtsarbeiten, HTA	2,1 %	2.303	3.772
#18	Marktwirtschaftliche Aspekte	1,9 %	2.129	3.800
#10	Ökonomische Faktoren	1,9 %	2.089	3.789
#7	Herz-Kreislauf, Dialyse, Trauma	1,6 %	1.723	3.392
#5	Internet, Social Networks	1,6 %	1.712	3.276
#25	Pädiatrie, Palliativversorgung	1,5 %	1.597	2.787
#24	Entwicklungsländer, übertragbare Krankheiten, HIV	1,4 %	1.565	2.794
#13	Telemedizin, psychische Erkrankungen, Versorgung	0,9 %	1.013	2.078
#31	Schwangerschaft, Geburt	0,9 %	965	1.772
#19	Diabetes inkl. BZ-Monitoring	0,8 %	922	1.571
#23	Krebserkrankungen/Onkologie	0,7 %	809	1.573
#16	Wundversorgung, Dekubitus, Hygiene	0,5 %	546	967
#26	Robotik	0,2 %	269	657
#6	Medikation	0,2 %	168	396
#11	Ernährung, Brasilien	0,1 %	103	163

[a] Anzahl Dokumente mit einem Anteil von mehr als 20 % des jeweiligen Topics

Im Rahmen der durchgeführten Suchstrategie weist das Gebiet der rechtlichen und organisatorischen Rahmenbedingungen den größten Anteil der Dokumente auf. Dieses stellt für 12,0 % (n=13.116) der Dokumente das primäre Topic dar. Bei 19.702 Dokumenten beträgt der Anteil von rechtlichen und organisatorischen Rahmenbedingungen über 20 %. Diese höhere Zahl liegt darin begründet, dass Dokumente ein abweichendes primäres Thema aufweisen, jedoch auch einen bedeutsamen Anteil des jeweiligen Gebietes aufweisen.

Bezogen auf konkrete Technologien stellen sich vor allem die elektronische Dokumentation (#17), Monitoring & Sensorik (#1), Sturz-/Gangerkennung (#27) Diabetes mit Blutzucker-Monitoring (#19), Wundversorgung (#16) und Robotik (#26) als potenziell pflegerelevante Themengebiete dar. Auf jedes dieser Themen entfallen jeweils weniger als 5 % der gesamten Dokumentenmenge bei Betrachtung der primären Themen. Die zehn relevantesten Begriffe dieser Topics sind in Tabelle II dargestellt.

Daneben zeigen sich weitere Themengebiete, welche potenziell relevante Dokumente beinhalten können. Hierzu zählen die Cluster zu wissenschaftlichen Studien (#3, #29 & #35), allen voran jedoch das Themengebiet Senioren, Pflegeheime, Demenz (#33). Da entsprechend der Suchstrategie eine Nennung von Technologie- bzw. Technikgebieten zu erwarten ist, wurde für die 7.383 Dokumente, in denen dieses Thema einen Anteil von mehr als 20 % aufweist, eine Subanalyse durchgeführt, die hier bespielhaft beschrieben wird.

TABELLE II. RELEVANTESTE BEGRIFFE FÜR AUSGEWÄHLTE TOPICS

Topic		
#1	***#16***	***#17***
monitoring	wound	EHR (electronic health record)
alarm	skin	CDS (Clinical decision support)
wearable	burn	HIT
ECG	dressing	electronic health records
sensor	textile	electronic health record
vital sign	ulcer	electronic
monitor	fabric	EMR (electronic medical record)
heart rate	pressure ulcers	EHRs
remote monitoring	healing	clinical decision support
bluetooth	pressure ulcer	physician
#19	***#26***	***#27***
diabetes	robot	rehabilitation
glucose	robotics	motor
blood glucose	animal	waking
insulin	ECMO	fall
infusion	companion	motion
CGM	mobile robot	gait
type diabetes	rat	movement
glycemic control	paro	wheelchair
hypoglycemia	service robots	posture
diabetic	BI	muscle

Ausgabe von *pyLDAvis*[9], λ=0,2

Tabelle III zeigt die relevantesten Begriffe der Subanalyse des Topics #33 zu Senioren, Pflegeheime, Demenz mit fünf Subtopics. Hier lassen sich eindeutiger die zugrundeliegenden Technologien erkennen: Robotik (#33.1), Ambient Assisted Living (AAL), Telecare und Kommunikationstechnik (#33.2), Interventionen bei Demenz mit assistiver Technologie und Paro[1] (#33.3), Spielen und Mensch-Computer-Interaktion (#33.4) sowie Monitoring inklusive Sturzerkennung (#33.5). Tabelle IV zeigt die zugehörige Häufigkeitsverteilung mit dem Schwerpunkt auf AAL und Telecare mit einem Anteil von ca. 30 % der Dokumente und nahezu identischen Anteilen für die übrigen 4 Topics.

[1] ein therapeutischer Roboter in Gestalt einer Robbe

TABELLE III. SUBTOPICS VON TOPIC #33: RELEVANTE BEGRIFFE

Topic				
#33.1	***#33.2***	***#33.3***	***#33.4***	***#33.5***
robot	service	intervention	game	sensor
robotic	telecare	dementia	design	monitoring
interaction	care	people dementia	student	detect
wheelchair	home	participant	serious games	detection
gesture	technology	resident	HCI	monitor
navigation	ageing	assistive devices	player	detecting
human	AAL	older adults	learning	fall detection
agent	ICT	paro	digital	abnormal
robotics	innovation	caregiver	participatory design	pattern
interface	project	disability	gaming	alert

TABELLE IV. SUBTOPICS VON TOPIC #33: VERTEILUNG

	primäres Topic		**Anzahl**
Topic	***Anteil***	***Anzahl***	**P > 20 %**[a]
#33.1	16,0%	1.180	2.157
#33.2	30,4%	2.246	3.604
#33.3	19,9%	1.467	2.556
#33.4	17,4%	1.286	2.242
#33.5	16,3%	1.204	2.161

[a] Anzahl Dokumente mit einem Anteil von mehr als 20 % des jeweiligen Topics

Insbesondere die Topics #33.1 (Robotik) und #33.5 (Monitoring/Sensorik) präsentieren dabei Themen und Begriffe, die bereits im ersten Topic Model (siehe Tabellen I & II) als eigene Cluster identifiziert wurden. Das Topic #33.1 weist eine starke Ähnlichkeit zu Topic #26 auf, während in #33.5 Begriffe sowohl aus den Topics #1 als auch #27 vorhanden sind.

Um die Inhalte der so identifizierten Publikationen zu verdeutlichen, stellt Tabelle V zehn zufällig ausgewählte Titel aus allen Dokumenten mit dem primären Topic #33.1 dar, sowie den Anteil des Textes, der den Topics #26 (Robotik), #33 (Senioren, Pflegeheime, Demenz) und #33.1 zugeordnet wird. Hier zeigt sich insbesondere, dass die Mehrzahl der Dokumente aus Topic #33.1 auch bedeutsame Anteile aus Topic #26 aufweisen, obgleich diese trotz im Titel benannten Robotik-Bezug lediglich bei etwa 20 % oder weniger liegen (Titel Nr. 1-4, 6, 7, 9). Sechs der zehn Titel weisen an Hand des Titels potenziell einen Pflegebezug auf (Titel Nr. 1, 2, 4, 6, 7, 9). Es finden sich jedoch auch Titel, die keinen offenkundigen Bezug zu Robotik oder Pflege erkennen lassen (insb. Titel Nr. 8 & 10). Diese weisen auch insgesamt geringere Anteile an allen drei dargestellten Topics aus. In der Gesamtschau spiegeln sich die durch das Topic Modeling identifizierten Themenanteile überwiegend in den Titeln wieder.

TABELLE V. ZUFÄLLIG AUSGEWÄHLTE TITEL AUS TOPIC #33.1

Nr.	Titel	Anteil #26	Anteil #33	Anteil #33.1
1	Light weight autonomous climbing robot for elderly and disabled persons' services	24,1%	39,4%	98,9%
2	Ambient intelligence and rehabilitation robots – A necessary symbiosis for robust operation in unstructured environments	17,9%	26,4%	91,7%
3	Task-specific assessment of robot effectiveness: enhancing the independence of quadriplegics in the workplace	19,6%	48,7%	90,5%
4	'Teach Me-Show Me'-End-User Personalization of a Smart Home and Companion Robot	13,1%	49,2%	84,3%
5	Healthcare system design focusing on emotional aspects using augmented reality – Relaxed service design	6,1%	35,1%	81,1%
6	User responses to social robots - experimental insights and psychophysiological measures	23,5%	28,4%	68,6%
7	Investigation of practical use of humanoid robots in elderly care centres	20,0%	47,5%	67,3%
8	Design of a cognitive user-support system for skin progress analysis using a smart phone	0,0%	26,0%	62,0%
9	Development of a network distributed telecare robotic system using CORBA	5,4%	22,3%	56,6%
10	On-Demand Virtual Health Counselor for Delivering Behavior-Change Health Interventions	0,0%	35,2%	51,9%

V. DISKUSSION UND AUSBLICK

In der vorliegenden Untersuchung wurde Topic Modeling mittels Latent Dirichlet Allocation zur inhaltlichen Analyse von Titeln und Abstracts einer Literaturrecherche mit über 100.000 Treffern eingesetzt. Mit der verwendeten Suchstrategie dominierten in den Ergebnisse solche Themen, die auch bereits in anderen Literaturrecherchen zum Einsatz von Technik in der Pflege beschrieben werden [10, 11]. Hierzu zählen insbesondere die elektronische Dokumentation, Monitoring von Personen und Vitalzeichen einschließlich Sensorik und Sturzerkennung, Kommunikationstechnik für Telemedizin und Telecare, sowie Robotik. Besonders elektronische Dokumentation und Sensorik waren häufige Themen in den vorliegenden Artikeln. Mit großem Abstand und auffallend geringen Anzahlen konnten Publikationen im Bereich der Robotik mit deutlich unter 1.000 Dokumenten und damit weniger als 1 % aller Ergebnisse identifiziert werden. Im Rückblick auf die Forschungsfrage lässt sich feststellen, dass mit den hier dargestellten Analyseschritten bisher in erster Linie Erkenntnisse zu Kategorien von Technologien und kaum zu konkreten technischen Lösungen oder Produkten in der Pflege gewonnen werden konnten.

Neben den einschlägigen Technologien wurde auch eine Vielzahl von Clustern identifiziert, die vor allem durch die Rahmenbedingungen, Einsatzgebiete oder auch Krankheitsbilder beschrieben wurden. Exemplarisch wurde für ein für die Forschungsfrage besonders relevantes Cluster eine Untermenge von Dokumenten eines zunächst wenig technikspezifischen Clusters zu älteren Menschen analysiert. Inhaltlich bestätigten sich dabei die bereits zuvor benannten Themengebiete. Hierbei zeigte sich, dass durch differenzierte Begutachtung der relevanten Begriffe der Sub-Cluster ein deutlicherer Einblick gewonnen und auch einzelne Technologien identifiziert werden können. Vor allem die durch *pyLDAvis* [9] zur Verfügung gestellte Relevanz-Metrik stellte sich hierfür als wichtiges Instrument dar. Trotzdem müssen Begriffe häufig vorkommen, damit diese in LDA eine entsprechende Relevanz erhalten. Der soziale Roboter *Paro* ist der einzige Produktname, der eine ausreichend hohe Häufigkeit und Relevanz hatte, um in den dargestellten Begriffen zu erscheinen.

Eine Limitation der hier präsentierten Arbeit besteht in der verwendeten Suchstrategie. Zur Erfassung eines möglichst breiten Spektrums von Technologien wurden neben allgemeinen Begriffen auch spezifischere Technologien benannt. Dies kann zu einer Überrepräsentation dieser Themen in den genutzten Daten führen. Da jedoch eine Vielzahl von weiteren Begriffen identifiziert wurde und ebenso nicht alle Suchbegriffe auch in den Modellen als relevant bewertet wurden, kann die Bedeutung für die hier vorliegende Forschungsfrage als nachrangig betrachtet werden. Ferner ist es durch die bereits im Suchterm definierten Ausschlusskriterien möglich, dass Publikationen vorzeitig ausgeschlossen wurden, obwohl sie für die Fragestellung relevant gewesen wären. Diesem Problem wurde begegnet, indem lediglich sehr spezifische Begriffe als Ausschlussbegriffe verwendet wurden. Beide Aspekte führen jedoch insbesondere dazu, dass Anteilswerte von Themen am gesamten Datensatz zurückhaltend zu interpretieren sind. Während die Dokumente eines Topics im Zähler eine hohe Vollständigkeit der verfügbaren Literatur aufweisen sollten, ist die Gesamtzahl der Dokumente im Nenner maßgeblich durch die Suchstrategie beeinflusst. Aus diesem Grund wurden vornehmlich absolute Anzahlen dargestellt.

Über die hier dargestellten Auswertungen hinaus bedarf die Anwendung des Topic Modeling zur Literaturanalyse in Public Health und Pflegewissenschaft weiterer Aufmerksamkeit um sinnvolle Anwendungsszenarien zu bestimmen. Es handelt sich um eine Methode, mit der mit verhältnismäßig geringem wiederkehrenden Aufwand eine große Anzahl von Dokumenten inhaltlich analysiert werden kann. Zugleich kann jedoch mit den unmittelbaren Ergebnissen lediglich ein grober Überblick über die Themengebiete der Dokumente erlangt werden. Themen, die nur selten vorkommen, können mit hoher Wahrscheinlichkeit auf diese Weise nicht identifiziert werden.

Die hier vorgestellten Ergebnisse zeigen ein Potenzial auf, um die händische Durchführung von Literaturrecherchen durch Forschende zu unterstützen und den zeitlichen Aufwand im Sichtungsprozess zu verringern. So können beispielsweise Dokumente aus nicht-relevanten Clustern ausgeschlossen werden. Abhängig von der Wahl der Grenzwerte des Anteils des Dokumentes am jeweiligen (Ausschluss-)Thema ist davon auszugehen, dass hiermit eine höhere Sensitivität und Spezifität erreicht werden kann als durch einen Ausschluss basierend auf einzelnen Worten in Titel oder Abstract. Dabei sollte auch geprüft werden, ob durch die Nutzung von Schnittmengen von Topics spezifischere Ergebnisse erzeugt werden können. Hierzu wäre es zunächst notwendig, Validierungsstudien

durchzuführen. Als nächster Arbeitsschritt besteht die Möglichkeit, ähnliche Dokumente aus Clustern zu extrahieren bzw. diese nach dem Anteil eines interessierenden Themas zu priorisieren. Schließlich stehen noch fortgeschrittenere Methoden zur Verfügung, welche explizit auf eine hierarchische Darstellung der Cluster abzielen und somit auch Zusammenhänge zwischen den Topics verdeutlichen können [12].

LITERATUR

[1] Shuyo, N. Language Detection Library for Java. 2010; Available from: http://code.google.com/p/language-detection/.

[2] Bird, S., E. Klein, und E. Loper, Natural Language Processing with Python. 2009: O'Reilly Media, Inc. 512.

[3] Řehůřek, R. und P. Sojka. Software Framework for Topic Modelling with Large Corpora. in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010. Valletta, Malta: ELRA.

[4] Blei, D.M., A.Y. Ng, und M.I. Jordan, Latent dirichlet allocation. J. Mach. Learn. Res., 2003. **3**: p. 993-1022.

[5] Konrad, M. tmtoolkit. Text Mining and Topic Modeling Toolkit for Python. 2018 22.04.2018]; Available from: https://github.com/WZBSocialScienceCenter/tmtoolkit.

[6] Mimno, D., et al., Optimizing semantic coherence in topic models, in Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011, Association for Computational Linguistics: Edinburgh, United Kingdom. S. 262-272.

[7] Röder, M., 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. 2015, ACM: Shanghai, China. S. 399-408.

[8] Cao, J., et al., A density-based method for adaptive LDA model selection. Neurocomputing, 2009. **72**(7): p. 1775-1781.

[9] Sievert, C. und K.E. Shirley. LDAvis: A method for visualizing and interpreting topics. 2014.

[10] Hielscher, V., Technikeinsatz und Arbeit in der_Altenpflege. Ergebnisse einer internationalen Literaturrecherche, in iso-Report Nr. 1. Berichte aus Forschung und Praxis, I.f.S.u.S.i. e.V., Hrsg. 2014: Saarbrücken.

[11] Merda, M., K. Schmidt, und B. Kähler, Pflege 4.0 – Einsatz moderner Technologien aus der Sicht professionell Pflegender. Forschungsbericht. 2017, Berufsgenossenschaft für Gesundheitsdienst und Wohlfahrtspflege (BGW): Hamburg.

[12] Blei, D.M., et al., Hierarchical topic models and the nested chinese restaurant process, in Proceedings of the 16th International Conference on Neural Information Processing Systems. 2003, MIT Press: Whistler, British Columbia, Canada. S. 17-24.

Dieser Beitrag entstand im Rahmen des durch das Bundesministerium für Bildung und Forschung geförderte Projekt Pflegeinnovationszentrum, Förderkennzeichen 16SV7821.