Dipl.-Soz. Ralph Schilling MPH
Dr. Thomas Stein
Prof. Dr. phil. Adelheid Kuhlmey
Dr. rer. pol. Stefan Blüher

Anwendung von Text Mining zur Auswertung von Begutachtungsdaten des MDK am Beispiel sozialer Einflussfaktoren von Pflegebedürftigkeit

Zusammenfassung

Die quantitative Analyse gesundheitsbezogener Routinedaten liefert oft nur begrenzte Erkenntnisse, wenn wichtige Informationen in Form von Freitexten vorliegen. Sogenannte Text Mining-Verfahren bieten hier – häufig ungenutzte – Möglichkeiten für zusätzlichen Erkenntnisgewinn. Im vorliegenden Beitrag soll dies am Beispiel von Freitextangaben aus Routinedaten des Medizinischen Dienstes der Krankenversicherung (MDK Berlin-Brandenburg) gezeigt werden. Es wurden Erstbegutachtungsdaten des MDK zur Feststellung einer Pflegebedürftigkeit aus dem Jahre 2017 (72.680 Antragstellende im Alter von 50-99 Jahren, etwa 80% dieser Personen erhielten eine Einstufungsempfehlung in einen Pflegegrad) mit dem Ziel ausgewertet, bedeutsame Einflussfaktoren für die Entstehung einer Pflegebedürftigkeit zu identifizieren. Hierbei spielen neben krankheitsbezogenen Ursachen auch soziale Faktoren, wie enge (familiale) Beziehungen mit hohem Unterstützungspotenzial, eine wichtige Rolle. Der Beitrag beschreibt zum einen das methodische Vorgehen beim Text Mining-Verfahren und präsentiert zum anderen Zusammenhänge von sozialer Unterstützung und der Einstufung in einen Pflegegrad, die ohne die Anwendung des Text Mining nicht analysierbar gewesen wären. Text Mining-Verfahren sollten daher zur Ausschöpfung des Informationsgehalts, gerade von Routinedaten, viel stärker genutzt und methodisch fortentwickelt werden.

Schlüsselwörter

Text Mining, Freitextanalyse, Routinedaten, Pflegebedürftigkeit, Soziale Unterstützung

1. Hintergrund: quantitative Analysen und Text Mining

>> Die Analyse von umfangreichen Routinedatensätzen, wie sie etwa durch Sozialversicherungsträger erhoben werden, liefert zu gesundheitswissenschaftlichen Fragestellungen häufig nur begrenzte Erkenntnisse. Dies liegt zum einen daran, dass die Daten meistens nicht – oder nur zu einem geringen Teil – für einen wissenschaftlichen Erkenntniszweck erhoben werden, zum anderen aber auch daran, dass die angewendeten quantitativen Analysemethoden Informationen nicht in dem Maße zugänglich machen, wie sie eigentlich im Datenmaterial enthalten sind. Ein Beispiel hierfür sind Freitextpassagen, die in sehr großen Datensätzen in entsprechend hoher Zahl zu finden sind, die aber mit den üblichen quantitativen Auswertungsmethoden nicht erfasst werden können und damit häufig als Datenquellen verloren gehen. So wächst der Bedarf an Verfahren, mit denen die in digitalen Datenquellen, wie Texten, Literaturdatenbanken, Webseiten oder E-Mails enthaltenen Informationen gewonnen und verarbeitet werden können. In Analogie zum Begriff des „Schürfens" von Bodenschätzen werden solche Verfahren als Data Mining oder Text Mining bezeichnet.

Während sich der allgemeinere Begriff des *Data Mining* oder der *Knowledge Discovery in Databases* (KDD) dabei eher auf algorithmus-basierte Verfahren zur Identifikation von Bedeutungsmustern (patterns) aus stark strukturierten Daten bezieht, wie sie bspw. im Zuge von routinemäßigen Abfragen in wirtschaftlichen Unternehmen oder staatlichen Verwaltungen anfallen, umfasst der Begriff des *Text Mining* eine Vielzahl von Verfahren zur Gewinnung und Verarbeitung von Informationen aus eher schwach- oder unstrukturierten textbasierten Datenbanken (Feldman/Dagan 1995). Text Mining oder *Knowledge Discovery in Textual Databases* (KDT), das als Begriff erstmals im Jahr 1995 von (Feldman/Dagan 1995) in die Forschungsterminologie eingeführt wurde, ist ein weitgehend automatisierter Prozess der Wissensentdeckung in textuellen Daten, der eine effektive und effiziente Nutzung verfügbarer Textarchive ermöglichen soll (Mehler/Wolff 2005). Joachims und Leopold (2002: 4) bezeichnen das Text Mining als „eine Menge von Methoden zur (halb-)automatischen Auswertung großer Mengen natürlichsprachlicher Texte".

Die Ziele der Anwendung von Text Mining-Verfahren sind nach Tiedemann (2019):

- Die Auswertung von Textdaten, die so umfangreich sind, dass sie nicht im Einzelnen von Menschen verarbeitet werden können,
- die Identifikation von Mustern und Beziehungen von Informationen, die in Texten repräsentiert sind, sowie
- die Extraktion von Wissen, das in großen Mengen von Textdaten implizit enthalten ist.

Abhängig von der Perspektive für die Anwendung von Text Mining können an die o.g. Definition anschließend verschiedene und zum Teil auch kombinierte Techniken verstanden werden, mittels derer „nützliches Wissen" (Fayyad et al. 1996; Fayyad et al. 1996; Kodratoff 2005) aus textbasierten Datenbanken extrahiert werden soll, um es mit linguistischen und statistischen Methoden zu erschließen (Hotho et al. 2005). Insgesamt ist die Entwicklung von Techniken innerhalb der interdisziplinären und noch recht jungen Disziplin der Wissensgenese aus großen Datenmengen weiter stark im Fluss und eine eindeutige methodische Abgrenzung deshalb schwierig. Eine detaillierte Darstellung verschiedener anwendungsorientierter Techniken des Text Mining findet sich aber bei (Hotho et al. 2005).

Die Anwendung des Text Mining im Rahmen der vorliegenden Arbeit lehnt sich an die *Technik der Informationsextraktion* (IE) an. Nach Grishman (2004: 545) kann IE als das automatisierte Erkennen von bestimmten Informationen bezeichnet werden: "The automatic

identification of selected types of entities, relations, or events in free text", bei dem es u.a. darum geht, diese Informationen bspw. einer Häufigkeitsanalyse zuzuführen.

Der Informationsgewinn, der durch die Anwendung des Text Mining erzielt werden kann, soll im Folgenden auf Basis der Daten des Medizinischen Dienstes der Krankenversicherung (MDK) Berlin-Brandenburg demonstriert werden. Dabei soll anhand von extrahierten Informationen aus textbasierten Angaben zu sozialer Unterstützung gezeigt werden, dass nicht nur routinemäßig erhobene Daten wie pflegebegründende Diagnosen oder soziodemographische Merkmale wie das Alter, sondern auch soziale Beziehungen und Netzwerke einen bedeutenden Einfluss auf den Eintritt von Pflegebedürftigkeit haben. Diese Zusammenhänge lassen sich mit Hilfe der Extraktion von Informationen aus Freitexten und ihrer Verarbeitung im Rahmen quantitativer statistischer Analysen finden.

2. Datensatzbeschreibung und Untersuchungsmethoden

2.1. Datensatzbeschreibung

Die Basis für die hier beschriebenen Analysen bilden die Pflegegutachten des Medizinischen Dienstes der Krankenversicherung Berlin-Brandenburg (MDK BB). Es handelt sich um Erstbegutachtungen des Jahres 2017 unter Nutzung des Begutachtungsinstruments nach dem neuen Pflegestärkungsgesetz II und der damit einhergehenden Umstellung von drei Pflegestufen auf fünf Pflegegrade. Diese Pfle-

Charakteristika des verwendeten Datensatzes

	ohne Einstufungs-empfehlung	Pflegegrad					
kein Pflegegrad/ Pflegegradeinstufungsempfehlung		1	2	3	4	5	Summe
n	15.108	18.798	24.836	10.287	2.979	672	72.680
Geschlecht							
Frauen, n (%)	9.138 (60,5)	11.997 (63,8)	14.596 (58,8)	5.305 (51,6)	1.428 (47,9)	294 (43,8)	42.758 (58,8)
Männer, n (%)	5.970 (39,5)	6.801 (36,2)	10.240 (41,2)	4.982 (48,4)	1.551 (52,1)	378 (56,3)	29.922 (41,2)
Alter in Jahren, Mittelwert (SD)	75,4 (10,9)	77,6 (10,3)	78,6 (10,1)	78,3 (10,1)	77,5 (10,6)	75,9 (10,8)	78,2 (10,2)
Haushaltszusammensetzung/Unterstützung							
zu Hause alleinlebend, n (%)	10.020 (67,7)	11.027 (60,2)	10.757 (47,9)	2.289 (33,5)	211 (21,1)	27 (13,2)	34.331 (54,0)
zu Hause mit weiterer Person zusammenlebend, n (%)	4.786 (32,3)	7.304 (39,8)	11.702 (52,1)	4.542 (66,5)	791 (78,9)	177 (86,8)	29.302 (46,0)
Verteilung Altersgruppen							
50-74, n (%)	6133 (40,6)	6.094 (32,4)	7.185 (28,9)	3.257 (31,7)	1.139 (38,2)	280 (41,7)	24.088 (33,1)
75-89, n (%)	8.265 (54,7)	11.313 (60,2)	15.243 (61,4)	6.135 (59,6)	1.640 (55,1)	356 (53,0)	42.952 (59,1)
90-99, n (%)	710 (4,7)	1.391 (7,4)	2.408 (9,7)	895 (8,7)	200 (6,7)	36 (5,4)	5.640 (7,8)

Tab. 1: Die Tabelle beschreibt die im Datensatz abgebildeten Personen nach Einstufungsempfehlungen in die verschiedene Pflegegrade sowie die Personen, die keine Einstufungsempfehlung erhielten. Weitere abgebildete Charakteristika sind Geschlecht, Durchschnittsalter, die Haushaltszusammensetzung in Kombination mit dem Vorhandensein oder Fehlen sozialer Unterstützung sowie einer Verteilung der Altersgruppen. Quelle: Anonymisierte Pflegebegutachtungen des Medizinischen Dienstes der Krankenversicherung 2017, eigene Berechnungen.

gebegutachtungen führen in den meisten Fällen Pflegefachkräfte in der Häuslichkeit der versicherten Person durch.

Der Datensatz umfasst 72.680 Anträge auf Leistungen aus der Pflegeversicherung in Berlin und Brandenburg und enthält neben soziodemographischen Daten auch Angaben über die pflegerelevante Vorgeschichte und aktuelle Versorgungssituation, Fremdbefunde, den gutachterlichen Befund und die Beschreibung von Wohnformen, Haushaltszusammensetzungen sowie Informationen zum Unterstützungspotenzial, zu Partnerschaft und sozialen Netzwerken. Diese Informationen sowie die im Begutachtungsinstrument angegebenen pflegebegründenden Erst- und Zweitdiagnosen führen letztlich zur gutachterlichen Einschätzung der Fähigkeit der antragstellenden Person, bestimmte Aufgaben und Anforderungen des Alltags selbstständig durchzuführen. Diese Einschätzung erfolgt differenziert nach sechs definierten Modulen mit entsprechender Gewichtung:

1. Mobilität (Gewichtung 10%), 2. kognitive und kommunikative Fähigkeiten, 3. Verhaltensweisen und psychische Problemlagen (15%, höchster Punktwert aus Modul 2 und 3 wird verwendet), 4. Selbstversorgung (40%), 5. Bewältigung von und selbstständiger Umgang mit krankheits- und therapiebedingten Anforderungen und Belastungen (20%) sowie 6. Gestaltung des Alltagslebens und sozialer Kontakte (15%).

Durch die Bewertung der Selbstständigkeit der antragstellenden Person nach den erwähnten Modulen wird letztlich die Einstufungsempfehlung in einen bestimmten Pflegegrad vorgenommen bzw. auch keine Pflegebedürftigkeit festgestellt.

Insgesamt stellten 42.758 (58,8%) Frauen und 29.922 (41,2%) Männer einen Antrag auf die Feststellung der Pflegebedürftigkeit (Tabelle 1).

Der Altersbereich der Antragstellerinnen und Antragsteller liegt zwischen 50 und 99 Jahren. Mit 42.952 (59,1%) Personen stellten die 75-89-Jährigen am häufigsten einen Antrag auf Feststellung der Pflegebedürftigkeit, gefolgt von 24.088 (33,1%) Personen in der Altersgruppe der 50-74-Jährigen, die 90-99-Jährigen taten dies 5.640 (7,8%) mal. Aus dem Bundesland Berlin kamen 38.819 (53,4%) aller Anträge, aus Brandenburg entsprechend 33.861 (46,6%). Die Erstbegutachtungen münden bei 57.572 (79,2%) Personen in die Einstufungsempfehlung in einen Pflegegrad, davon sind 33.620 (58,4) weiblichen, 23.952 (41,6%) männlichen Geschlechts.

2.2. Datenschutz

Um dem hohen Gut des Schutzes der Sozialdaten beim MDK BB Rechnung zu tragen, erfolgte im Vorfeld der Datennutzung mit höchster Priorität die vollständige Anonymisierung der Gutachten. Für die ge-

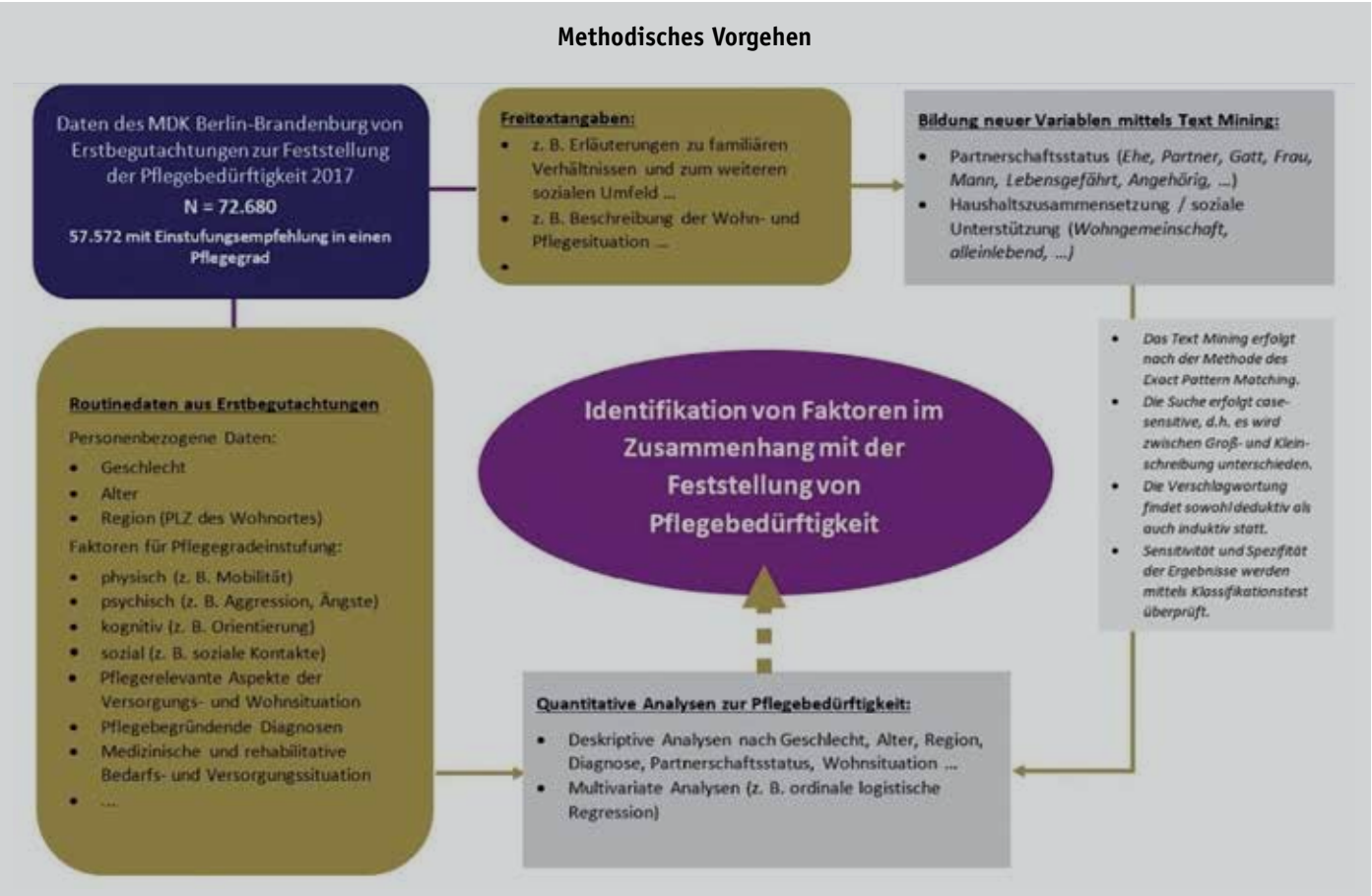

Abb. 1: Die Inhalte der Begutachtungsunterlagen erlauben zwei methodische Zugänge zur Verarbeitung dieser Daten; erstens über die enthaltenen Routinedaten wie z.B. Geschlecht, Alter, Region und zweitens über die Freitextangaben zu bspw. familiären Verhältnissen und dem sozialen Umfeld, die die Bildung neuer Variablen (z.B. Partnerschaftsstatus) mittels Text Mining erlauben. Beide Informationsquellen werden schließlich deskriptiven und multivariaten Analysen zugeführt für die Identifikation von Faktoren im Zusammenhang mit der Feststellung von Pflegebedürftigkeit. Quelle: Anonymisierte Pflegebegutachtungen des Medizinischen Dienstes der Krankenversicherung 2017, eigene Darstellung.

planten Analysen erhielt das Institut für Medizinische Soziologie und Rehabilitationswissenschaft eine einzelne, auf einem verschlüsselten Datenträger gespeicherte Datei. Zugriff auf die Daten haben ausschließlich die im Projekt arbeitenden Wissenschaftlerinnen und Wissenschaftler.

Weitere Maßnahmen zum Schutz der Daten sind die Verfügbarkeit der Postleitzahlen lediglich für die ersten drei Stellen (dementsprechend auf ein größeres geographisches Aggregat bezogen) und die Zusammenfassung des Alters der Antragstellerinnen und Antragsteller zu 5-Jahres-Altersgruppen.

2.3 Methode

Grundlage für die statistischen Analysen zur Identifikation von Faktoren, die im Zusammenhang mit der Feststellung von Pflegebedürftigkeit und Einstufungsempfehlung in einen Pflegegrad stehen, bilden die zuvor beschriebenen Daten des MDK BB. Unter anderem beinhalten diese neben den pflegebegründenden Diagnosen auch routinemäßig erhobene Daten zu personenbezogenen Merkmalen wie Geschlecht, Alter, Haushaltszusammensetzung oder Wohnort (in Form der ersten drei Stellen der PLZ) sowie Informationen zu individuellen psychischen, physischen, kognitiven und sozialen Voraussetzungen für die Pflegegradeinstufung. Sofern diese Informationen bereits als numerische Daten vorliegen, können sie direkt in die statistischen Analysen zu den Einflussfaktoren für die Einstufungsempfehlung einbezogen werden. Zur Anwendung kommen dabei bspw. nach Geschlecht, Alter, Haushaltszusammensetzung und Wohnort differenzierte statistische Analysen, die Aufschluss über das Zusammenwirken verschiedener Charakteristika mit der Pflegegradeinstufung liefern können. Darüber hinaus finden sich im Zuge der Erstbegutachtungen durch den MDK auch Daten zu den familiären Verhältnissen der Antragstellenden, zum sozialen Umfeld oder Beschreibungen der Wohnbedingungen und Pflegesituation in Form von Freitextangaben. Diese Angaben sind zwar nicht obligat, erlauben aber oftmals Rückschlüsse zu sozialen Konstellationen, die im Zusammenhang mit einer Pflegegradeinstufung relevant sind.

Der beschriebene methodische Ansatz umfasst somit zwei Zugänge zur Identifizierung von Einflussfaktoren der Pflegebedürftigkeit: standardisierte Routinedaten und Text Mining, Abbildung 1 verdeutlicht beide Zugangswege.

Im Rahmen der Anwendung von Text Mining auf die vorliegenden Daten der Begutachtungen werden mithilfe einer Verschlagwortung von Schlüsselbegriffen aus den Freitextangaben weitere Merkmale der begutachteten Personen extrahiert und einer statistischen Analyse zugänglich gemacht. Das Verfahren zur Ermittlung dieser Informationen wird nach der Methode des „Exact Pattern Matching" durchgeführt. Dazu werden thematisch relevante Freitextangaben in den Datensätzen der Antragstellenden selektiert und in Zeichenketten (Strings) umgewandelt. Anschließend wird jeder dieser Strings exakt nach relevanten synonymen Schlagwörtern durchsucht. Für das Merkmal Partnerschaft wurden bspw. die Begriffe *Ehe, Partner, Gatt*[1], *Mann, Lebensgefährt, Lebensgefaehrt* ermittelt und ausgewählt. Im Falle eines Treffers wird dann einer neu gebildeten numerischen Variable Partnerschaft eine 1 für „in Partnerschaft lebend", anderenfalls eine 0 für „alleinstehend" zugeordnet. Angaben zu weiteren Dimensionen des Merkmals „soziale Unterstützung" wie *Kinder, andere Familienangehörige* oder *Kontaktpersonen* werden nach dem gleichen Vorgehen ermittelt und entsprechend in numerische Variablen transformiert. Infokasten 1 zeigt detailliert, wie bei der Informationsextraktion und der Variablenbildung zu unterstützenden Personen vorgegangen wurde.

Die Suche der Zeichenketten erfolgt case-sensitive, d. h. es wird zwischen Groß- und Kleinschreibung (z.B. *Angehörig, angehörig*) unterschieden. Durch diese Verfahrensweise wird bspw. verhindert, dass eine Enkeltochter zugleich als Tochter gewertet und damit ein und derselbe Eintrag doppelt gezählt wird. Die Verschlagwortung findet sowohl deduktiv als auch induktiv statt. Somit werden Schlagwörter einerseits anhand von theoretisch abgeleiteten bzw. in Wörterbüchern befindlichen Synonymen sowie andererseits nach empirisch ermittelten Textfeldinhalten gebildet.

In einem zweiten Schritt wird das Vorhandensein unterstützender Personen aus den ermittelten Ergebnissen in eine Variable „soziale Unterstützung" aggregiert. Sofern mindestens eine der gebildeten Variablen die Ausprägung „1" aufweist, wird die Variable „soziale Unterstützung" ebenfalls mit einer „1" kodiert, anderenfalls mit „0", siehe Infokasten 1.

Zur Validierung der Suchergebnisse werden Sensitivität und Spezifität des Verfahrens mittels Klassifikationstest überprüft. Zu diesem Zweck werden Zufallsstichproben von Gutachten gezogen

Bildung der Variablen für „soziale Unterstützung"

In einem ersten Schritt wurden die Variablen „Partnerschaft", „Kinder", „andere Familienangehörige" und „andere Kontaktpersonen" gebildet. Dies geschah über die Ermittlung, wie oft bestimmte Zeichenketten Erwähnung in den Freitexten finden:

Partnerschaft: *Partner, Lebensgefaehrt, Lebensgefährt, Ehe, Gatt, Mann* – die Zeichenkette *Frau* wurde nicht verwendet, weil es hier zu viele false positives gab (z.B. „Die Telefonnummer von **Frau** … lautet …").

Kinder: *Tochter, Töchter, Toechter, Sohn, Söhne, Soehne, Kind, Schwieger*

Andere Familienangehörige: *Familie, Angehörig, angehörig, Angehoerig, angehoerig* – die Zeichenkette familie wurde nicht verwendet, da es zu viele false positives gab (z.B. „Ein**familie**nhaus")

Andere Kontaktpersonen: *Nachbar, nachbar, Betreuer, betreuer* – die Zeichenketten *freund* und *bekannte* wurden nicht verwendet, da es zu viele false positives gab aufgrund der Verwendung beider Begriffe als Adjektive (z.B. „**freund**lich", „**bekannt**es Datum", …)

In einem zweiten Schritt wurde die Verfügbarkeit von unterstützenden Personen zur neu gebildeten Variable „soziale Unterstützung" zusammengefasst. Sofern mindestens eine der vier gebildeten Variablen die Ausprägung „1" aufweist, wurde die Variable „soziale Unterstützung" ebenfalls mit einer „1" kodiert, anderenfalls mit „0".

Infokasten 1: Bildung der Variablen für „soziale Unterstützung"

1 Bei Verwendung des Schlagwortes „Gatt" würden bspw. die Begriffe „Gatte" und auch „Gattin" und weitere gefunden werden.

Technische Umsetzung des Text Mining

Zum Einlesen der txt-Datei wurde die Funktion import im Paket rio verwendet. Zum Exact-String-Matching wird die Funktion grep (base R) genutzt.

Die Syntax der Funktion lautet:

grep(pattern, x, ignore.case = FALSE, perl = FALSE,

value = FALSE,fixed = FALSE, useBytes = FALSE, invert = FALSE)

Die Funktion hat folgende Argumente:

pattern: character string containing a regular expression (or character string for fixed = TRUE) to be mat ched in the given character vector. Coerced by as.character to a character string if possible. If a character vector of length 2 or more is supplied, the first element is used with a warning. Missing values are allowed except for regexpr and gregexpr.

x, text: a character vector where matches are sought, or an object which can be coerced by as.character to a character vector. Long vectors are supported.

ignore.case: if FALSE, the pattern matching is case sensitive and if TRUE, case is ignored during matching.

perl: logical. Should Perl-compatible reg exps be used?

value: if FALSE, a vector containing the (integer) indices of the mat ches determined by grep is returned, and if TRUE, a vector containing the matching elements themselves is returned.

fixed: logical. If TRUE, pattern is a string to be matched as is. Overrides all conflicting arguments.

useBytes: logical. If TRUE the matching is done byte-by-byte rather than character-by-character. See 'Details'.

invert: logical. If TRUE return indices or values for elements that do not match.

Die Funktion gibt per default die Indizes der Felder an, die den gesuchten String enthalten. Sofern die Länge des zurückgegebenen Vektors größer gleich 1 ist, enthält mindestens eines der Felder den gesuchten Begriff. Wir wenden die Funktion mittels „apply" zeilenweise auf unseren Datensatz an. Die Dauer für den gesamten Datensatz beträgt ca. 1 Minute.

Infokasten 2: Technische Umsetzung des Text Mining

(n=100) und sowohl händisch als auch unter Anwendung des beschriebenen Verfahrens auf das Vorkommen und die Validität der Bedeutungszuschreibung der verwendeten Schlagwörter geprüft. Die Zuverlässigkeit der Methode wird bewertet, indem Zuordnungen zu „true positive", „false positive", „true negative" sowie „false negative" berechnet werden. Die händische Erfassung fungiert dabei als Referenz, die automatisierte Erfassung als Komparator. Thematisch relevante Suchbegriffe, die erst im Zuge der Validierung ermittelt werden können, werden dem Schlagwortkatalog hinzugefügt. Nach erfolgter „Sättigung" der Verschlagwortung wird die Suche auf den gesamten Datensatz angewendet. Die Extraktion wurde mit der Software R, Version 3.4.3. durchgeführt. Infokasten 2 zeigt detailliert die Vorgehensweise für die technische Umsetzung des Verfahrens.

Die beschriebene Variante des Text Mining eröffnet die Möglichkeit, Informationen bspw. zu sozialen Unterstützungspotenzialen aus den Freitexten herauszufiltern und sie quantitativen statistischen Analysen zugänglich zu machen.

3. Erkenntnisgewinn durch Text Mining am Beispiel sozialer Unterstützung

Wie vorangehend erörtert, nutzen wir für unsere Analysen der sozialen Unterstützung von Personen, die einen Antrag auf Leistungen aus der Pflegeversicherung stellen, Schlagworte, die auf soziale Beziehungen oder Netzwerke hinweisen; diese sind z.B. „Partnerin" oder „Partner", „Kinder" und andere Familienangehörige, aber auch außerfamiliäre Kontaktpersonen wie „Freund", „Freundin", „Nachbarin" oder „Nachbar". Die mittels Text Mining gebildeten Variablen können nun im Zusammenhang mit einer vorhandenen Pflegegradeinstufung ausgewertet werden.

Unsere Beispielergebnisse zeigen die neu gebildete Variable zu sozialer Unterstützung in Verbindung mit der Haushaltszusammensetzung (alleinlebend/ nicht alleinlebend) und dem Zusammenhang zur Feststellung der Pflegebedürftigkeit. Die Kombinationsvariable Haushaltszusammensetzung beinhaltet drei Kategorien: *alleinlebend ohne Unterstützung, alleinlebend mit Unterstützung* und *nicht alleinlebend*, wobei bei letzterer Kategorie davon ausgegangen wird, dass diese Haushaltszusammensetzung mit dem höchsten sozialen Unterstützungspotenzial einhergeht.

Folgende Ergebnisse erläutern beispielhaft den Zusammenhang zwischen der Konstellation Haushaltszusammensetzung/soziale Unterstützung und der Höhe des empfohlenen Pflegegrades.

Abbildung 2 verdeutlicht, dass nicht alleinlebende Personen im Vergleich der drei Kategorien die höchsten Pflegegradempfehlungen erhalten; Mit Ausnahme von Pflegegrad 1 liegt deren Anteil stets über 50%. Die geringsten Anteile zeigen sich über alle Pflegegrade für Alleinlebende, die von keiner Unterstützung durch andere Personen berichten (von 20% in Pflegegrad 1 bis 2% in Pflegegrad 5). Insgesamt werden für Alleinlebende unabhängig von der Verfügbarkeit sozialer Unterstützung durchschnittlich niedrigere Pflegegrade empfohlen als für Nicht-Alleinlebende (Pflegescore[2] 1,66 zu 1,97).

4. Diskussion

Der vorliegende Artikel beschreibt die Anwendung eines Text Mining-Verfahrens zur Identifikation von textbasierten Informationen am Beispiel sozialer Einflussfaktoren auf die Feststellung und Einstufung von Pflegebedürftigkeit. Durch diese Methode ist es möglich, Angaben in Freitexten des Begutachtungsinstruments des Medizinischen Dienstes der Krankenversicherung einer quantitati-

2 Der Pflegescore wird aus dem arithmetischen Mittel der jeweiligen Häufigkeiten für jeden einzelnen Pflegegrad gebildet.

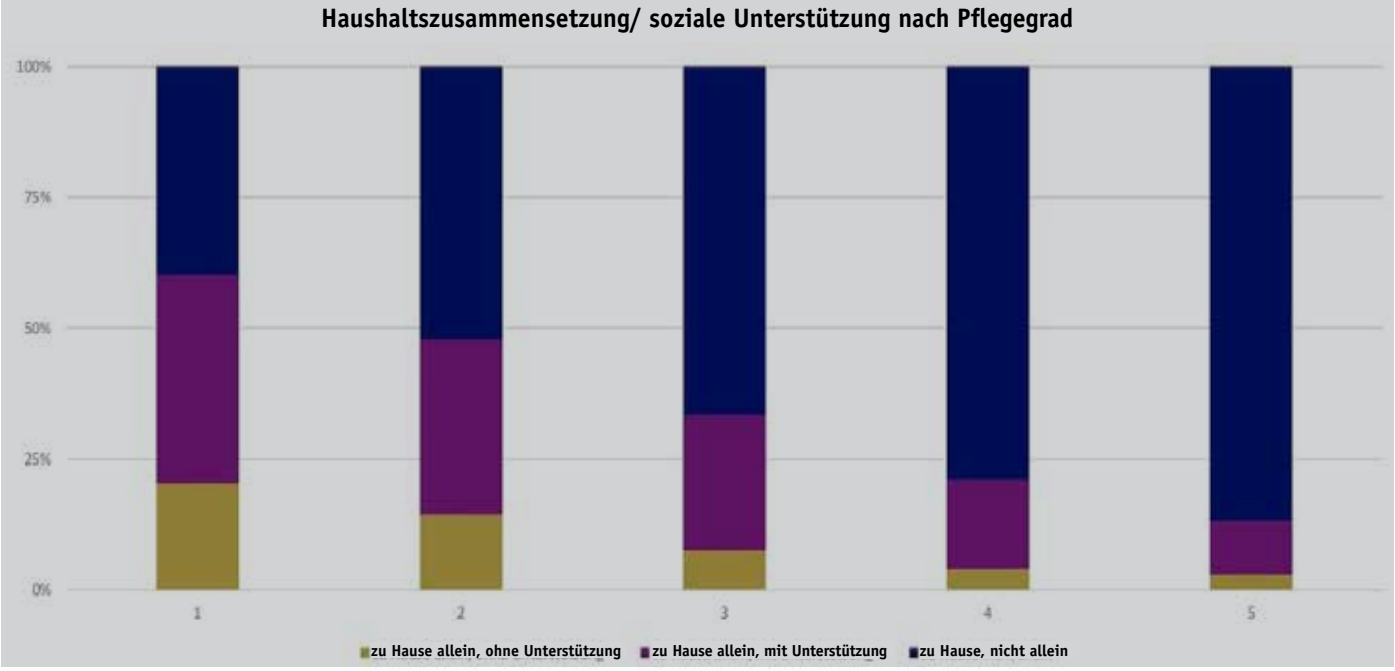

Abb. 2: Gezeigt wird hier der prozentuale Anteil der in der eigenen Häuslichkeit nicht alleinlebender und alleinlebender Antragsteller und Antragstellerinnen in Verbindung zur Pflegegradempfehlung. Alleinlebende Antragsteller und Antragstellerinnen werden nach dem Vorhandensein oder der Abwesenheit sozialer Unterstützung differenziert. Quelle: Anonymisierte Pflegebegutachtungen des Medizinischen Dienstes der Krankenversicherung 2017, eigene Berechnungen.

ven statistischen Analyse zuzuführen und beispielsweise Befunde zur Bedeutung von Partnerschaft sowie sozialen Netzwerken für die Einstufung in einen Pflegegrad zu generieren.

Methodisch ist das gewählte Vorgehen an verschiedene Studien (Karystianis et al. 2018; Ananiadou et al. 2006; Kayser/Blind 2017) angelehnt und folgt dem aktuellen Forschungsstand. In Bezug auf die Größe des Trainingssamples orientierten wir uns an der Studie von Karystianis et al. (2018), in der ein vergleichbares Verfahren eingesetzt wurde und ein Trainingssample von n=100 bei einer deutlich höheren Fallzahlbasis (492.393) als in vorliegender Studie zur Anwendung kommt. Dabei konnten äußerst präzise Ergebnisse erzielt werden.

Limitationen des angewendeten Verfahrens zum Text Mining liegen hingegen bspw. in der Identifikation von zwar relevanten, aber orthografisch fehlerhaft übertragenen Textbestandteilen im Datensatz. So würde bspw. der Begriff „Prtner" nicht als Partner oder Partnerin identifiziert werden können, wenn er nicht zufällig in den ermittelten Ergebnissen oder den gezogenen Zufallsstichproben auftaucht und auf diese Weise Eingang in die Verschlagwortung findet. Bei der Bewertung der Ergebnisse muss außerdem berücksichtigt werden, dass die Validierung der Suchergebnisse mittels Klassifikationstest auf der Basis von jeweils n=100 Zufallsstichproben durchgeführt wurde. Inwiefern eine Erhöhung der Anzahl an Stichproben die Zuverlässigkeit der Ergebnisse verbessern kann, sollte in nachfolgenden Studien weiter erprobt werden. Prinzipiell bleibt die Anzahl der für die Validierung gezogenen Stichproben aus pragmatischen Gründen aber begrenzt.

Die Analysen beziehen sich außerdem auf *berichtete* Partnerschaften sowie soziale Netzwerke und damit verbundene Unterstützungspotenziale. Unabhängig von der beschriebenen Validierung der Suchergebnisse können somit Fälle auftreten, in denen z.B. von einer Partnerschaft berichtet wird, es sich dabei aber um eine/einen verstorbene*n Partner*in handelt. Oder es können Fälle auftreten, in denen Personen zwar keine Partnerschaft berichten, aber dennoch in einer solchen leben und auf Unterstützung zurückgreifen können. Demzufolge besteht die Möglichkeit einer Über- oder Unterschätzung der auf Basis des hier beschriebenen Text Mining-Verfahrens ausgewiesenen Prävalenzen.

Weitere limitierende Aspekte ergeben sich aus der vorliegenden Datenbasis. So handelt es sich zunächst um Querschnittsdaten, so dass evtl. vorliegende Progredienzen der Pflegebedürftigkeit nicht über den Zeitverlauf analysiert werden können. Darüber hinaus sind für die vorliegenden Analysen ausschließlich die Begutachtungsunterlagen des MDK für die Bundesländer Berlin und Brandenburg genutzt worden – jedes Bundesland und damit auch Berlin und Brandenburg zeichnet sich durch diverse Spezifika aus (z.B. Altersstruktur, Geschlechterverhältnis, Krankheitsgeschehen) – und stellt demnach kein repräsentatives Sample für die gesamte Pflegesituation in Deutschland dar. Im Hinblick auf unsere Beispielergebnisse zeigen jedoch auch andere Studien (Borchert/Rothgang 2008), dass eine Partnerschaft häufig mit einer verschobenen Beantragung von Leistungen aus der Pflegeversicherung in Zusammenhang steht. Ebenso wurde der protektive Charakter einer Partnerschaft bereits in anderen Studien hervorgehoben (Schneider et al. 2020; Hajek/König 2016); unsere Befunde decken sich mit diesen Ergebnissen und bestätigen damit die Bedeutung einer Partnerschaft als wichtigem sozialen Einflussfaktor im Zusammenhang mit der Entstehung einer Pflegebedürftigkeit.

Über den Informationsgewinn hinaus, der sich aus der Anwendung des beschriebenen Text Mining-Verfahrens ergibt, ist eine noch detailreichere Freitext-Analyse wünschenswert: Inwieweit ergeben sich bspw. durch Unterstützungsarrangements einerseits familiärer Art und andererseits nicht-familiärer Art (z.B. Nachbarn/Bekannte) Synergieeffekte oder Reibungsverluste? Evtl. ist es nicht ausreichend, pauschal das Unterstützungspotenzial zu betrachten – so können trotz Unterstützung durch zahlreiche Nachbarn andere Bedarfe bestehen bleiben (z.B. Unterstützung bei der Körperpflege, wenn Nachbarn ausschließlich bei Besorgungen helfen). Diese Analysen haben im Rahmen der Studie bereits begonnen und werden das Wissen um spezifische Unterstützungspotenziale erweitern.

5. Schlussfolgerungen

Das Ziel unserer Analysen unter Anwendung von Text Mining ist es, neben soziodemographischen Charakteristika und pflegebegründenden Diagnosen auch soziale Parameter als Einflussfaktoren auf einen Pflegebedarf zu untersuchen. Zu diesen sozialen Parametern zählen wir vor allem (i) die Eingebundenheit in familiäre Beziehungen (Partnerschaft, Kinder) und andere soziale Netzwerke (Bekannte, Nachbarn und Freunde), (ii) die aktuelle Haushaltszusammensetzung/soziale Unterstützung und (iii) das Vorhandensein von Barrieren in der eigenen Wohnumgebung von Antragstellerinnen und Antragstellern. Es konnte gezeigt werden, dass mittels Text Mining-Verfahren deutliche Informationsgewinne aus Routinedaten des Medizinischen Dienstes der Krankenversicherung erzielt werden können. Dies sollte Anlass sein, den Einsatz von Text Mining in der Freitextanalyse weiter zu etablieren und durch die Modifikation des eingesetzten Validierungsverfahrens methodisch fortzuentwickeln. Aktuell wird im Rahmen unserer Studie ein weiterer Beitrag erarbeitet, der sich auf die hier beschriebenen Verfahren stützt und sich thematisch auf die von Antragstellerinnen und Antragstellern berichteten Barrieren in der eigenen Wohnumgebung konzentriert. <<

Literatur

Ananiadou, S./Kell, D.B./Tsujii, J. (2006): Text mining and its potential applications in systems biology. In: Trends in biotechnology 24, 12,: 571-579

Borchert, L./Rothgang, H. (2008): Soziale Einflüsse auf das Risiko der Pflegebedürftigkeit älterer Männer. In: Bauer, U./Büscher, A. (Hrsg.) (2008): Soziale Ungleichheit und Pflege: Beiträge sozialwissenschaftlich orientierter Pflegeforschung 2008: 215-237

Fayyad, U./Piatetsky-Shapiro, G./Smyth, P. (1996): The kdd process for extracting useful knowledge from volumes of data. In: Communications of the ACM 1996, 39, 11: 27–34

Fayyad, UM./Piatetsky-Shapiro, G./Smyth, P. (1996): From data mining to knowledge discovery: An overview. In: Fayyad, UM./Piatetsky-Shapiro, G./Smyth, P./Uthurusamy, R (Hrsg.) (1996): 1-34

Feldman, R./Dagan, I. (1995): Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD) 112–117. https://www.aaai.org/ Papers/KDD/1995/KDD95-012.pdf. (abgerufen am 15.10.2020)

Grishman, R. (2004): Information Extraction. In Mitkov, R. et al. (Hrsg.) (2004): 545-559. Zitiert nach Schmolz, H. (2015): Anaphora Resolution and Text Retrieval. In: Imo, W./Spieß, C. (Hrsg.) Empirische Linguistik / Empirical Linguistics, Band 3 Walter de Gruyter, Berlin/Boston. file:///C:/Users/User/AppData/Local/Temp/[9783110416756%20-%20Anaphora%20Res olution%20and%20Text%20Retrieval]%20Anaphora%20Resolution%20and%20Text%20Retrieval.pdf. (abgerufen am 30.10.2020)

Hajek, A./König, H.H. (2016): Longitudinal Predictors of Functional Impairment in Older Adults in Europe- Evidence from the Survey of Health, Ageing and Retirement in Europe. In: PLoS One 2016, 11, 1: e0146967. doi:10.1371/journal. pone.0146967

Hotho, A./Nürnberger, A./Paaß, G. (2005): A Brief Survey of Text Mining. In: Zeitschrift für Computerlinguistik und Sprachtechnologie 2005, 20, 1. http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf (PDF). (abgerufen am 15.10.2020)

Joachims, T./Leopold, E. (2002): Themenheft: Text-Mining. Vorwort der Herausgeber. Künstliche Intelligenz 2(4). Zitiert nach: Mehler, A./Wolff, C. (2005): Einleitung: Perspektiven und Positionen des Text Mining. In: Zeitschrift für Computerlinguistik und Sprachtechnologie 2005, 20, 1. https://web.archive.org/web/20150402143908/ http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=1773642& fileOId=2311304 (abgerufen am 15.10.2020)

Karystianis, G. et al. (2018): Automatic extraction of mental health disorders from domestic violence police narratives: text mining study. In: Journal of medical internet research 2018, 20, 9: e11548.

Kayser, V./Blind, K. (2017): Extending the knowledge base of foresight: The contribution of text mining. In: Technological Forecasting and Social Change 2017, 116: 208-215

Kodratoff, Y. (2005): Knowledge discovery in texts: A definition and applications. In: Rás, Z.W./Skowron, A. Proceedings of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS '99) Zitiert nach: Mehler, A./Wollf, C. (2005): Einleitung: Perspektiven und Positionen des Text Mining. Zeitschrift für Computerlinguistik und Sprachtechnologie 2005, 20, 1. https://web.archive.org/web/20150402143908/http://pub.uni-bielefeld.de/luur/ download?func=downloadFile&recordOId=1773642&fileOId=2311304. (abgerufen am 15.10.2020)

Mehler, A./Wolff, C. (2005): Einleitung: Perspektiven und Positionen des Text Mining. In: Zeitschrift für Computerlinguistik und Sprachtechnologie 2005, 20, 1. https://web.archive.org/web/20150402143908/ http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=1773642&fileOId=2311304 (abgerufen am 15.10.2020)

Schneider, A./Blüher, S./Grittner, U. et al. (2020): Is there an Association between Social Determinants and Care Dependency Risk? A multistate model analysis of a longitudinal study. In: Research in Nursing & Health 2020, 43, 3: 230-240

Tiedemann, M. (2019): Text Mining – Grundlagen, Methoden und Anwendungsfälle. News-Blog der Alexander Thamm GmbH zum Thema: Künstliche Intelligenz. https://www.alexanderthamm.com/de/blog/text-mining-grundlagen-methoden-und-anwendungsfaelle/ (abgerufen am 09.11.2020)

Autorenerklärung

Ethische Richtlinien wurden bei der Durchführung und Auswertung dieser Studie eingehalten. R. Schilling, T. Stein, A. Kuhlmey und S. Blüher geben an, dass keine Interessenkonflikte besteht. Die Lieferung der anonymisierten Daten erfolgte durch den Medizinischen Dienst der Krankenversicherung.

Using text mining as a tool to analyze assessment data provided by the Health Insurance Medical Service (MDK) based on the example of social predictors of care dependency

Statistical analysis of routine health data often yields limited findings when key information is available only in unstructured text form. Text-mining methodologies allow this type of information to be processed. This study uses data from care assessments conducted by the Berlin-Brandenburg Health Insurance Medical Service (Medizinischer Dienst der Krankenversicherung Berlin-Brandenburg) to demonstrate the benefits of using text mining to obtain information by extracting data from free-form text passages on applications for social healthcare support. It uses initial assessment documents from 72,680 applicants (age range: 50-99 years) from 2017 to identify determinants influencing care dependency. Around 80% of these applicants were assigned to a care-level category. In addition to illness-related causes, social environmental factors such as family circumstances play a major role in determining which level of care the applicant requires. This article describes the methodology of text mining and examines correlations between an individual's social environment and their categorisation into a care level which would not have been analysable without the use of text mining. The findings suggest that text mining should be more widely used and methodically developed to capture relevant information, especially within sets of routine data.

Keywords

Text Mining, Free Text Analysis, Routine Data, Care Dependency, Social Support

Dipl.-Soz. Ralph Schilling
ist Diplom-Soziologe und MSc in Public Health. Nach dem Studium an der Freien Universität Berlin war er zunächst u.a. wissenschaftlicher Mitarbeiter am Robert Koch-Institut sowie am Institut für Biometrie und Klinische Epidemiologie der Charité-Universitätsmedizin Berlin. Zurzeit ist er Wissenschaftler am Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie der Charité-Universitätsmedizin Berlin. Seine Arbeitsfelder liegen im Bereich Methoden sowie Prävention und psychosozialer Gesundheitsforschung.
Kontakt: ralph.schilling@charite.de

Dr. Thomas Stein
studierte im Master Demographie an der Universität Rostock und der Autonomen Universität Barcelona. Er befasste sich bereits im Rahmen seiner Masterarbeit an der Universität Rostock mit dem Thema der Demenz als bedeutsamer Determinante für Pflegebedürftigkeit in Deutschland mit ihren Komorbiditäten. Seit April 2017 ist Herr Stein als Wissenschaftlicher Mitarbeiter am Institut für Medizinische Soziologie und Rehabilitationswissenschaft an der Charité – Universitätsmedizin Berlin.
Kontakt: thomas.stein@charite.de

Prof. Dr. phil. Adelheid Kuhlmey
ist Direktorin des Instituts für Medizinische Soziologie und Rehabilitationswissenschaft sowie wissenschaftliche Direktorin des Centrums für Human- und Gesundheitswissenschaften der Charité – Universitätsmedizin Berlin. Sie war Mitglied der 3., 4. und 5. Altenberichtskommission der Bundesregierung. 2016 bis 2020 wurde sie in den Deutschen Ethikrat berufen. Zum 01.01.2020 hat sie den Vorsitz des unabhängigen Beirats für die Vereinbarkeit von Pflege und Beruf beim BMFSFJ übernommen.
Kontakt: adelheid.kuhlmey@charite.de

Dr. rer. pol. Stefan Blüher
ist Soziologe, wissenschaftlicher Mitarbeiter und leitender Wissenschaftler am Institut für Medizinische Soziologie und Rehabilitationswissenschaft, Charité-Universitätsmedizin Berlin. Seine Arbeitsfelder sind unter anderem: Gesundheit und Prävention im höheren Lebensalter, Technisierung und Digitalisierung in Medizin und Pflege.
Kontakt: stefan.blueher@charite.de