Die Arbeit mit unstrukturierten textbasierten Daten

Eine Reflexion zu Mixed-Methods-Ansätzen für Textanalysen

Franziska Hein-Pensel

Beitrag zur Veranstaltung »'Mixed Methods' zwischen Methodenintegration und Methodenpluralismus« der Sektion Methoden der qualitativen Sozialforschung

Einleitung

Aufgrund der steigenden Verfügbarkeit großer Textdaten sind Text Mining und insbesondere Topic Modeling relevante Methoden, um sich Forschungsfragen in verschiedenen Fachbereichen zu nähern (Roberts et al. 2019). Topic Modeling kann dabei als eine explorative Technik beschrieben werden, um Informationen aus Textdaten in großem Maßstab zu gewinnen (DiMaggio et al. 2013). Dies führt dazu, dass das Interesse an Topic Modeling im letzten Jahrzehnt deutlich gewachsen ist und sich von der Informatik in andere Disziplinen, wie der Soziologie (z.B. Apishev et al. 2016; Bohr, Dunlap 2018) oder den Wirtschaftswissenschaften (z.B. Wang et al. 2017; Schmiedel et al. 2019), verzweigt hat.

Mit Hilfe einer methodischen Kombination von Topic Modeling und qualitativer Kodierung können Wissenschaftler*innen Informationen aus einem Datenkorpus gewinnen, die von Hand nicht vollständig auswertbar gewesen wären (z.B. Shimizu 2017; Croidieu, Kim 2018). Dieser methodengemischte Ansatz erfordert eine konstante Zirkulation zwischen der Interpretation des Outputs und der Datenaufbereitung für die quantitative Analyse. Aufgrund der Komplexität dieses Prozesses ist sowohl Zeit als auch Sorgfalt gefordert.

Trotz der wachsenden Popularität von Topic Modeling in den Sozialwissenschaften fehlt es nach wie vor an gemeinsamen Qualitätsrichtlinien für Wissenschaftler*innen, um die Transparenz ihrer Arbeit zu gewährleisten (Antons et al. 2020). Im Vergleich dazu gehört es in der Informatik zur gängigen Praxis, den Leser*innen einen detaillierten technischen Bericht anzubieten, der alle Informationen zur Erstellung der präsentierten Ergebnisse enthält. Hierdurch wird die Nachvollziehbarkeit der Analyseschritte gewährleistet.

Der vorliegende Beitrag plädiert dafür, dass bei der Implementierung eines Text-Mining-Algorithmus aus der Informatik in die Sozialwissenschaft die erzeugten Ergebnisse mit den gleichen Standards wie in ihrer Ursprungsdisziplin behandelt werden sollten und zeigt hierfür Möglichkeiten auf. In diesem Beitrag wird die Verwendung von Topic Modeling und induktiver Kodierung sowie das Zusammenspiel beider Methoden diskutiert. Der Mehrwert dieser Studie besteht darin, Qualitätsleitli-

nien für den Umgang mit unstrukturierten Textdaten zur Gewährleistung von Transparenz vorzustellen.

Die Kombination von qualitativen und quantitativen Textanalysen erfordert ein Verständnis beider Methoden und Wissen über ihre Stärken und Schwächen, um sie optimal miteinander zu verknüpfen. Das nächste Kapitel gibt einen kurzen Überblick über beide methodischen Bereiche, ihre Stärken sowie ihrer kritischen Diskussionen. Nachdem sie getrennt vorgestellt wurden, geht es im Folgenden um die Kombination beider Methoden. Dabei werden nicht nur die neuen Möglichkeiten, sondern auch die Limitationen beleuchtet, die solche Mixed-Methods-Ansätze für Textanalyseverfahren mit sich bringen. Deshalb werden im letzten Kapitel Qualitätsleitlinien vorgestellt, die die Transparenz und Glaubwürdigkeit von Forschungsergebnissen sicherstellen sollen.

Grundprinzipien der qualitativen und quantitativen Textanalyse

Qualitative Textanalyse

Der klassische Umgang mit textbasierten Daten – wie Interviews, Web-Inhalte oder Geschäftsberichte – ist die qualitative Textanalyse (z.B. Gioia, Thomas 1996; Kuckartz, Sharp 2011; Kohlbacher 2006). Je nach Forschungsziel können sowohl die Datenerhebung als auch der methodische Ansatz variieren. Um beispielsweise ein Untersuchungsziel zu analysieren, das bisher nicht im Fokus der Forschung stand, müssen Wissenschaftler*innen einen offenen Ansatz wählen, um dem Forschungsgegenstand gegenüber unvoreingenommen zu sein. Wenn die Datenquelle aus ethnografischen oder semi-strukturierten Interviews oder semi- beziehungsweise unstrukturierten Textdaten aus digitalen oder Printmedien stammt, beinhaltet eine qualitative Datenanalyse wahrscheinlich einen Bottom-Up-Prozess, bei dem das Kodierungsschema aus den Daten selbst abgeleitet wird. Im Gegensatz zu diesem induktiven Kodierungsverfahren steht die Untersuchung eines gut spezifizierten Forschungsziels, welches mittels strukturierter Interviews oder halb- und strukturierter Textdaten aus digitalen und gedruckten Medien untersucht wird. Der Inhalt dieser Textdaten wird in einem Top-Down-Prozess kodiert. Hierfür wird ein vorformuliertes Kodierungsschema auf der Grundlage des Forschungsziels und der verwendeten Literatur (deduktives Verfahren) gefordert.

Beide, die induktive und die deduktive Kodierung, können im Prozess der Datenanalyse kombiniert werden, die sich als komplementäre oder abduktive Ansätze zusammenfassen lassen (Dubois, Gadde 2002). Eine solche Kombination ist für die Textanalyse sehr gebräuchlich und beschreibt einen Rechercheprozess, der sowohl von etablierten theoretischen Modellen als auch von neuen Konzepten, die aus den Daten hervorgehen, geleitet wird. Auf diese Weise ist die Forschungsphase der Datenerhebung mit der Phase der Datenanalyse verflochten, was zu einer gegenseitigen Beeinflussung führt (vgl. Tabelle 1).

Validität, Reliabilität und Transparenz

Es gibt keinen Standard für die Wahl eines Ansatzes oder einer Methode zur Generierung neuen Wissens. Jedoch werden drei Qualitätskriterien, die innerhalb der wissenschaftlichen Gemeinschaft geteilt werden, herangeführt um einen „akademischen Wert" zu sichern: dazu gehört die Validität, die Reliabilität sowie die Generalisierbarkeit (Carp, Carp 1981; Marsh et al. 2008; Leung 2015).

Um zu untersuchen, ob die Ergebnisse einer qualitativen Textanalyse zuverlässig und valide sind, ist es besonders wichtig, innerhalb des Prozesses der Datenanalyse transparent zu sein (Roller 2019; Yin 1994). Beispielsweise kann das Forschungsdesign mehrere Analytiker*innen einschließen, welche

unabhängig voneinander die gleiche Stichprobe kodieren und die hieraus resultierenden Ergebnisse vergleichen und vor allem Unterschiede hierbei diskutieren (Interkodierer-Reliabilität). Wenn die Ergebnisse konsistent sind, kann der Kodierungsprozess mit periodischen Überprüfungen auf kontinuierliche Interkodierer-Übereinstimmung fortgesetzt werden (MacQueen et al. 1998, S. 35). Zusätzlich sollte eine Reflexion über die Forschungsumgebung durchgeführt werden, um mögliche Limitationen aufzuweisen. Wichtig ist, dass die Forscher*in dem Forschungsfeld aufgeschlossen begegnen sollte und es so wenig wie möglich beeinflussen sollte.

Tabelle 1: Ansätze zur qualitativen Textanalyse

Methode	Datenanalyse	Literaturbeispiele
Induktiver Ansatz	**Bottom-Up Kodierung** Offene Kodierung, Kodierungsschema wird von den Daten abgeleitet	Gioia et al. 2013 Mayring 2000, 2015, 2019
Deduktiver Ansatz	**Top-Down Kodierung** Formulieren vordefinierten Kodierungsschema auf Grundlage von Literaturrecherche und Forschungsfrage(n)	Mayring 2015
Abduktiver Ansatz	**Verwendung einer pragmatischen Sichtweise** Ständiger Wechsel von der empirischen zur theoretischen Dimension der Analyse	Dubois, Gadde 2002 Lukka, Modell 2010

Stärken und Schwächen

Trotz der steigenden Beliebtheit der qualitativen Textanalyse zur Untersuchung theoretischer Konstrukte ist eine Reflexion ihrer Stärken und Schwächen notwendig. Insgesamt hat die qualitative Forschung innerhalb der Sozialforschung ein gemischtes Ansehen. Einerseits ermöglicht sie es den Forscher*innen, sich intensiv mit unbekannten Forschungszielen auseinanderzusetzen, wodurch die Ergebnisse anschließend in der Grundlagenforschung genutzt werden können. Andererseits betonen Kritiker*innen, dass die qualitative Forschung innerhalb der Datenerhebung und -interpretation das Gütekriterium Generalisierbarkeit nicht einhalten kann und es somit ihr an einer adäquaten Rechtfertigung ihrer Ergebnisse mangelt und es zur Überinterpretation eher spärlicher Evidenz einlädt (Gioia et al. 2013).

Insbesondere der letzte Punkt ist durch die natürliche Begrenzung der menschlichen Kapazitäten bedingt. Um diese natürliche Limitation zu überwinden, kann die Verwendung eines Mixed-Methods-Ansatzes hilfreich sein. Durch den Einsatz von quantitativen Textanalysen mit Hilfe der Verarbeitung natürlicher Sprache, kann die Untersuchung von Daten verbessert werden und die Analyse gesellschaftlicher Phänomene bereichert werden. Bevor ein standardisierter Leitfaden zur Verbindung von qualitativen und quantitativen Daten eingeführt wird, werden im nächsten Abschnitt quantitative Methoden der Textanalyse und deren Charakteristika vorgestellt.

Quantitative Textanalyse

Im Jahr 1966 veröffentlichten Stone et al. ein Buch über einen computergestützten Ansatz zur Inhaltsanalyse (Stone et al. 1966). Sie argumentierten, dass es in der menschlichen Kommunikation wiederkehrende Muster gibt, die von Computern erkannt werden können. Die Einführung dieses heutzutage sehr rudimentären Systems, das in der Lage ist, Texte zu verarbeiten, nach Wörtern und Phrasen zu suchen, Vorkommen zu zählen sowie Sätze mit bestimmten Merkmalen zu finden, war der Ausgangs-

punkt für den Einsatz einer computergestützten Inhalts- und Textanalyse in verschiedenen Bereichen außerhalb der Informatik (Früh 2015).

Giegler (1992) fasst den Nutzen der computergestützten Textanalyse zusammen und unterscheidet zwischen zwei Säulen. Zum einen die wirtschaftlichen Vorteile, indem große Textdaten sehr effizient verarbeitet werden können. Zum anderen steht die zweite Säule für die Verlässlichkeit der Ergebnisse, welche darauf hindeutet, dass mit einem vordefinierten Wörterbuch derselbe Textkorpus mehrfach analysiert werden kann und die Ergebnisse jedes Mal identisch wären.

Seitdem ist die Methodenvielfalt mit der Entwicklung der Informatik und des Bereichs der Verarbeitung natürlicher Sprache gewachsen. Gut etablierte Algorithmen auf dem Gebiet der Informatik haben in der sozialen Forschungsgemeinschaft immer mehr Beachtung gefunden. Daher nehmen die Implementierungen von Algorithmen in den Forschungsdesigns der Sozialwissenschaftler*innen zur Analyse eines Textkorpus und damit zur Gewinnung von Einsichten in die latenten Themen stetig zu (Hannigan et al. 2019). Insbesondere die Zunahme der verfügbaren Daten und die damit verbundenen Probleme, diese manuell zu analysieren, haben die Nachfrage nach Text-Mining-Techniken erhöht (Antons et al. 2020). Häufig werden so genannte *unbeaufsichtigte Lernalgorithmen* verwendet, die ohne vordefinierte Kodierungsschemata oder Kennzeichnungen auskommen (Blei et al. 2003).

Obwohl es ein wachsendes Interesse an der computergestützten Sozialwissenschaft gibt, mangelt es an Verständnis dafür, was diese Algorithmen tun und wie sie funktionieren. Eine kürzlich durchgeführte Studie hat gezeigt, dass zwar die Anzahl an veröffentlichten Studien steigt, welche Text-Mining verwenden, es jedoch immer noch ein Mangel an Transparenz und Replizierbarkeit dieser Artikel gibt (Antons et al. 2020).

Validität, Reliabilität und Transparenz

Mit computergestützten Werkzeugen für quantitative Textanalysen scheint es fast so, als sei eine der größten Herausforderungen die Sicherung der Transparenz (Antons et al. 2020). Dies ist überraschend, wenn man bedenkt, dass einige dieser Algorithmen seit den neunziger Jahren oder Anfang zweitausend existieren. Transparent zu sein bedeutet im Falle der computergestützten Textanalyse, die verwendete Umgebung[1] und jeden Schritt des Analyseprozesses zu dokumentieren[2]. Darüber hinaus ist es wichtig, die Parameter nicht willkürlich zu wählen und zu benennen. Hierfür gibt es allgemeine Richtlinien und Tests, an denen sich Wissenschaftler*innen orientieren können[3]. Nichtsdestotrotz benötigt der Output eine qualitative Interpretation als Validierung.

Bei der Verwendung von *überwachten Lernalgorithmen* für Vorhersagemodelle ist ein weiterer wichtiger Aspekt darauf zu achten, dass die herangeführten Daten nicht zu sehr angepasst werden. Das bedeutet, dass alle Einstellungen für den jeweiligen Datensatz sehr einzigartig sind und dass das Modell nicht auf Daten verallgemeinert werden kann, die es noch nie zuvor gesehen hat. Ein fertiges Papier sollte eine Zusammenfassung dieser Dokumentation enthalten[4].

Stärken und Schwächen

Die Vielfalt der Text-Mining-Werkzeuge bietet Sozialwissenschaftler*innen neue Möglichkeiten zur Analyse gesellschaftlicher Phänomene. Besonders mit der Datenmenge, die durch die Digitalisierung zugänglich wird, können Wissenschaftler*innen wichtige Themen oder Muster auf sehr effiziente Wei-

[1] Zum Beispiel: R, Python oder eine Anwendung wie https://www.minemytext.com/ oder WordStat.
[2] Einschließlich jeder Datenmanipulation, jeder angepassten Parametereinstellung und jedes Tests.
[3] Um zum Beispiel ein k für Topic Modeling zu wählen, kann ein Kohärenzmodell verwendet werden.
[4] Je nach Zeitschrift kann die Verwendung eines Appendix für die Beschreibung hilfreich sein.

se identifizieren. Ein weiterer Vorteil ist, dass Interpretationsverzerrungen vermieden werden können (DiMaggio et al. 2013; Hannigan et al. 2019). Im Allgemeinen liefern Text-Mining-Techniken im Vergleich zu menschlichen kodieren sehr konsistente Ergebnisse.

Trotz dieser neuen Möglichkeiten tauchen auch Limitationen auf, die berücksichtigt werden müssen. Das Wissen über die Methoden innerhalb der sozialwissenschaftlichen Gemeinschaft ist sehr heterogen. Einige Wissenschaftler*innen definieren sie immer noch als kleine Black-Box. Diese Einschätzung trifft für maschinelles Lernen mittels neuronalen Netzwerken zu, passt aber nicht allgemein zu Text-Mining-Werkzeugen. So basieren beispielsweise Topic Modeling, K-Means oder Sentiment-Analysen, auf mathematischen Algorithmen, die zur Einsicht offenstehen. Ein weiteres Problem ist die langsame Verbreitung neuer Erkenntnisse, Tests und Algorithmen innerhalb der Computational Social Science in den Sozialwissenschaften.

Insgesamt ist es schwierig, die quantitative Textanalyse allein zur Analyse eines Phänomens zu verwenden, da die Kontextualisierung fehlt. Dies kann durch die Verbindung der Ergebnisse mit qualitativer Textanalyse ergänzt werden. Solche Mixed-Methods-Ansätze sind für das Gebiet der Sozialwissenschaften nicht neu (Croidieu, Kim 2018; Shimizu 2017; Wilden et al. 2018). Allerdings fehlen ihnen ein gemeinsames Verständnis und Qualitätsrichtlinien. Das folgende Kapitel stellt die Möglichkeiten und Einschränkungen vor, die Mixed-Methods-Ansätze für Textanalyse bieten und fasst grundlegende Leitlinien zusammen.

Möglichkeiten und Einschränkungen von Mixed-Methods-Ansätzen bei der Textanalyse

Mixed-Methods-Ansätze und ihre Anwendungen variieren von der Datenerhebung bis zur Interpretation und Validierung. Der vorliegende Beitrag diskutiert die Umsetzung eines gemischten Methodenansatzes bei der Dateninterpretation von textuellen un- oder semistrukturierten Daten.

Zuvor ist es wichtig zu wissen, dass es im Rahmen der quantitativen Textanalyse bei der Aufbereitung der Daten unvermeidlich ist, die maschinellen Ausgaben zu interpretieren. Die Verbindung der quantitativen Textanalyse mit einer qualitativen Untersuchung ist nur eine Erweiterung der bereits notwendigen Schritte der menschlichen Interpretation. Antons et al. (2020, S. 342) beschreiben Text Mining als eine Brücke zwischen qualitativen und quantitativen Forschungstraditionen.

Chancen – Möglichkeiten

Bei der Nutzung eines Mixed-Methods-Ansatzes ergeben sich für Sozialwissenschaftler*innen Möglichkeiten, Phänomene in einem neuen Umfang zu beobachten. Es stehen große Datensätze zur Verfügung, die mehrere Forschungsfragen innerhalb verschiedener Disziplinen beantworten können. Beispielsweise bieten Twitter und Facebook *Application Programming Interfaces* (APIs, deutsch: Programmierschnittstelle) für den Zugriff auf Daten an, es können Jahres-, Quartals- oder Monatsberichte über große Zeiträume analysiert oder mit anderen Unternehmen oder anderen Informationsquellen verglichen werden. Ein weiterer Datenpool können die Web-Inhalte von Unternehmen oder Zeitungsartikel sein. Insgesamt wachsen die potenziellen Quellen für Textdaten stetig weiter. Es ist möglich, Metadaten mit diesen Textdaten in Beziehung zu setzen, um die Ergebnisse zu kontextualisieren. Durch die Neustrukturierung der Daten können neue Informationen sichtbar gemacht werden, die zu besseren Interpretationen führen. Bei bereits qualitativ analysierten Daten kann eine zusätzliche

quantitative Analyse neue Einsichten bieten oder die qualitative Interpretation stärken oder in Frage stellen (Tran et al. 2013).

Neben den vielfältigen Möglichkeiten, verschiedene Datenquellen für eine groß angelegte Analyse miteinander zu verbinden, können nun auch Forschungsfragen untersucht werden, die in der Datenstruktur verborgen waren oder die aufgrund der menschlichen kognitiven Begrenzungen übersehen wurden. Beispielsweise können sich wiederholende Muster der institutionellen Reproduktion von Ungleichheiten, wie Sexismus, Rassismus oder Diskriminierung aufgrund des sozialen Status, branchenübergreifend oder über lange Zeiträume hinweg durch Geschäftsberichte, Medienpräsentationen und Archive analysiert werden.

Ein weiterer Forschungszweig mit vielen Möglichkeiten ist das Zusammenspiel zwischen Medien, Politik und anderen Mitgliedern der Gesellschaft. Solche und weiter Forschungsbereiche können mit bestehenden Theorien in empirischen Studien untersucht werden.

Einschränkungen – Begrenzungen

Diese neuen Möglichkeiten haben auch Einschränkungen, die es zu berücksichtigen gilt. Erstens ist es wichtig, die Quelle der Daten in Frage zu stellen. Wenn zum Beispiel eine API verwendet werden kann, sollte hinterfragt werden, wer diese anbietet? Die gesammelten Daten innerhalb der APIs sind vorselektiert und können ein bestimmtes Bild ergeben, das wahrgenommen werden soll. Zweitens hat nicht jeder Zugang zum Erlernen dieser neuen Algorithmen. Letztendlich ist es die Aufgabe von Sozialwissenschaftler*innen, gesellschaftliche Phänomene zu analysieren und nicht Informatiker*in zu werden. Dies ist neben vielen anderen Gründen eine weitere Bestärkung, die interdisziplinäre Arbeit auszuweiten und offen für andere Disziplinen zu sein.

Ein weiterer wichtiger Punkt, den es zu berücksichtigen gilt, ist die Datenaufbereitung für eine quantitative Textanalyse. Der Datenbereinigungsprozess ist ein notwendiger Schritt, um brauchbare Ergebnisse zu erhalten. Nichtsdestotrotz besteht der Kern der Reduktion der Daten darin, eine abstraktere Ebene zu erhalten. Je mehr Datenmanipulation vorgenommen wird, desto abstrakter sind die Daten am Ende. Je nach Forschungsagenda und Fragestellung kann das Niveau variieren. Es gibt keine Best-Practices. Das beste Vorgehen für ein Forschungsprojekt hängt von dem angestrebten Ziel ab.

Ein weiteres mögliches Problem der quantitativen Textanalyse kann der Mangel an Dokumentation sein, der zu nicht reproduzierbaren Ergebnissen führen kann. Deshalb ist es so wichtig, die Parameter der Methoden zu kennen, bevor diese angewandt werden. Ein weiterer Grund für nicht reproduzierbare Ergebnisse ist, dass es keine gemeinsamen Qualitätsrichtlinien für gemischte Methodenansätze bei Textanalysen gibt. Dieser Mangel an Richtlinien und die oft fehlenden Dokumentationen erschweren es den Forscher*innen, solche Arbeiten zu validieren und zu verstehen. Darüber hinaus ist es für Forscher*innen schwieriger, solche Forschungsmethoden für ihre Projekte zu adaptieren, die neue Ergebnisse erzeugen und die Theorien und empirischen Studien in verschiedenen Forschungsströmungen bereichern würden.

Aufgrund der großen Popularität und der neuen Möglichkeiten dieses Ansatzes ist es wichtig, einen gemeinsamen Leitfaden zu entwickeln, der Transparenz und Verlässlichkeit garantiert, aber dennoch flexibel genug ist, um für verschiedene Forschungsbereiche und Forschungsfragen angepasst werden zu können.

Qualitätsleitlinien

Der vielversprechende Charakter von Mixed-Methods-Ansätzen liegt in der Möglichkeit für die Forschenden, ihre Untersuchung an die ständig wachsende Datenmenge anpassen zu können. Dennoch verlangt die Einführung mathematischer Methoden in die qualitative Forschung ein Nachdenken darüber ab, ob durch die Verknüpfung der Methoden einen Mehrwehrt für die Untersuchung generiert wird. Forscher*innen sollten in der Lage sein, die Frage zu beantworten, warum es notwendig ist, die spezifischen Methoden zu kombinieren und wie diese bei ihrer Untersuchung helfen.

Darüber hinaus stellten Gioia et al. (2013, S.19) fest, dass „gute" qualitative Forschung von einer gut spezifizierten oder eher allgemeinen Forschungsfrage zu Beginn der Analyse sowie von der Einbettung mehrerer Datenquellen in die Studie abhängt. Dieser Standard lässt sich auf die Textanalyse mittels Mixed-Methods-Ansätzen übertragen. Des Weiteren ist es bei qualitativer sowie bei quantitativer Forschung gängige Praxis, den Kodierungsprozess transparent zu gestalten und den Datensatz zu beschreiben. Bei Studien, welche quantitative Textanalyse verwenden, mangelt es jedoch häufig noch an Transparenz bei der Darstellung der Datenvorverarbeitung und -bearbeitung (Antons et al. 2020, S. 340f.). Um das Problem der unzureichenden Transparenz zu lösen, führen Antons et al. (2020) sechs Aspekte ein, denen Studien, die quantitatives Text Mining verwenden, berücksichtigen müssen, um transparent zu sein und so zukünftige Replikationsstudien und die kumulative Entwicklung des Wissens in unserem Bereich zu ermöglichen (Antons et al. 2020, S. 342). Mit diesem Beitrag möchte ich den von Antons et al. (2020) vorgestellten Leitfaden zur Transparenz von Text-Mining-Ansätzen zu einem allgemeinen Leitfaden für einen gemischten Methodenansatz erweitern, der qualitative und quantitative Textanalyse verbindet(vgl. Tabelle 2).

Tabelle 2: Qualitätsleitlinien

Forschungsprozess	Beschreibung	Leitfragen
1. Forschungsfrage	• Herausarbeitung einer spezifischen Forschungsfrage • Literaturübersicht verschaffen	*Was ist der aktuelle Forschungsstand?* *Welche Zielsetzung und Methode passt am besten zu dieser Forschungsfrage?* *Welche Daten werden benötigt?*
2. Datensammlung ***(Beispielvariante I)***	*Interviews:* • Sich mit dem Forschungsfeld vertraut machen • Transparent über den Ansatz sein • Vorgespräche (offene Fragen klären) • Interviewprotokoll: o Forschungsfrage(n) müssen verständlich sein und keine Suggestivfragen	*Sind die Daten ausreichend?* *Welche Limitationen gibt es?* *Welche Rolle nimmt der/die Interviewer*in ein?* *Sind die Fragen offen genug?* *Wie wirkt sich der Kontext (Ort, Kommunikation im Vorfeld, Kultur, …) auf die Interviewbedingungen und -ergebnisse aus?*
(Beispielvariante II)	*Website:* • Nutzungsbedingungen lesen • Automatisiertes oder Manuelles erheben der Daten (automatisiert: API Schnittstellen oder html codes analysieren/nutzen) • Reduktion des Datenmaterials auf eine Sprache	*Sind die Daten ausreichend?* *Welche Limitationen gibt es?* *Was wird nicht benötigt zu erheben?* *Manuelle oder automatische Erfassung?* *Was ist das beste Tool dafür?* *Wie ist die Website kodiert?* *Wie groß ist die Fehlerquote?*

(weitere Varianten)	• Andere verfügbare Textdaten wie: Zeitungen, Geschäftsberichte, oder Literatur aller Art (Gedichte, Artikel, Bücher, Songtexte, …)	…
3. Auswahl der Methodik (Quantitative und qualitative)	• Methoden wählen, die am besten zum Forschungsobjekt und den Daten passen • Tools kombinieren, um so viele Erkenntnisse wie möglich über die Daten zu gewinnen	*Welche Bibliotheken gibt es für die Programmiersprache?* *Welche Ressourcen für Expertenunterstützung sind vorhanden?* *Wie etabliert ist die Methode?* *Gibt es Werkzeuge für diese Methode und wenn ja, sind diese Open Source? (externe Validierung)*
4. Datenvorbereitung für die quantitative Textanalyse	• Mit den Daten vertraut machen • Korpus definieren und in strukturierten Text konvertieren • Basierend auf der Forschungsagenda den Abstraktionsgrad wählen: o Entfernen von Ziffern und Stoppwörtern (Präpositionen, Eigennamen, selbst definierte Stoppwörter) o Entfernen der obersten x% der Wörter (Indikator für Präpositionen oder Lückenwörter) und die unteren x% der Wörter (repräsentieren nicht die Daten) oder Wörter, die weniger als x und/oder mehr als x Mal vorkommen o Lemmatisierung o Stemming	*Welche Informationen (features) werden für die Analyse benötigt?* *Welche Software oder welche Tools sind am besten geeignet? (Python, R, MineMyText, https://voyant-tools.org/, …)* *Was ist die Analyseeinheit? (Wortstämme oder Lemmata, Wörter, Wortsegmente, sprachliche Merkmale, …)*
5. Quantitative Textanalyse	• Datenvisualisierung • Text in eine quantitative Matrix umwandeln (bag of words, tf-idf, …) • Definieren von Parametern … Beispiel für LDA (Topic Modeling): o Alpha: Dokument-Thema-Verteilung (wenn NULL: Anzahl der Topics sind gleichmäßig auf die Dokumente verteilt) o Beta: Wörter-Themen-Verteilung (genism library = eta) (default 0,1) o Anzahl der Themen (k) (Testthemenanzahl, z.B. Ellenbogentest, Kohärenzmodell, …) o Seed setzen (Ergebnisse sind reproduzierbar) o Maximale Anzahl von Iterationen wählen	*Wie sind die Daten verteilt?* *Wie viele Daten sind bei der Datenaufbereitung verloren gegangen?* *Sind die Dokumente thematisch kohärent oder weitgehend unstrukturiert? (Alpha definieren)* *Wie ausgeprägt ist die Sprache im Datensatz? (beta definieren)* *Wie viele Dokumente umfasst der Datensatz? (Rahmen für Tests zur Bestimmung von k setzen)*
6. Qualitative Textanalyse – Teil I	• Untersuchen der Ausgaben und Zurückgehen auf die Dokumente, wenn nötig • Aufgeschlossen bleiben und die Analyse validieren	*Sind Fehler vorhanden? (Rechtschreibfehler, Entitätsnamen, falsche Parameter)* *Sind die Ausgaben unterscheidbar (interpretierbar)? (Wenn*

		nicht, sollten die Parameter angepasst werden, um die Lesbarkeit der Ergebnisse zu verbessern.) *Ist die gewählte quantitative Methode für Ihre Daten und Ihre Fragestellung geeignet?* *Können die Topics weiter eingegrenzt werden?*
7. Zirkuläre Interpretation und Datenauswertung	• Einstellungen verbessern, Fehler beheben, Schritte und gewählte Parameter dokumentieren • Zirkulieren zwischen Schritt 5 und 6, bis ein stabiles Ergebnis erreicht ist	
8. Qualitative Textanalyse – Teil II	• Qualitative Interpretation der Ausgaben • Induktives kodieren der Ergebnisse oder bilden von Kategorien • Muster und Strukturen der kodierten Ergebnisse und der quantitativen Analyse untersuchen und gegebenenfalls generalisieren/ abstrahieren	*Bieten die Befunde sinnvolle Beiträge zum Forschungsziel?* *Sind die Bedeutungen der Befunde kohärent zueinander?* *Welche Muster lassen sich erkennen?* *Abhängig von der Methodik,* *Beispiel Topic Modeling:* *Was sind dominanten Topics?* *Wie können die Topics mit dem Datensatz in Verbindung gebracht werden?* *Wie repräsentativ sind die Topics für meinen Datensatz?*
9. Ergebnisse zusammenfassen	• Visualisierung von quantitativen Ergebnissen in Plots und Tabellen • Vernetzen der Ergebnisse mit der Literatur in dem jeweiligen Fachgebiet (und interdisziplinär) • Transparenz der Forschungsmethode	*Welche Grafik oder welches Schema beschreiben die Erkenntnisse am besten?* *Was sind die Beiträge?*

Diskussion

Der vorliegende Beitrag bietet eine detaillierte Orientierungshilfe, um ein einheitliches Verständnis für die verschiedenen Anforderungen zu schaffen, die an eine Studie mit Mixed-Methods-Ansätzen zur Textanalyse gestellt werden. Gerade in der heutigen global und medial vernetzten Zeit, in der über Social-Media-Plattformen massenhaft neue Informationen in Sekundenschnelle verbreitet werden, gewinnen solche Methoden an Attraktivität. Die methodische Relevanz ergibt sich jedoch nicht nur für Social Media, sondern auch für die Textanalyse über politische Reden, wissenschaftliche Publikationen, Geschäfts- oder Nachrichtenberichte. Die erwähnten Textdaten können im Zentrum einer solchen Analyse stehen und erlauben eine Reduktion der Datenmasse, ohne die Analyseeinheit auf eine Teilstichprobe zu reduzieren.

Diesen Möglichkeiten stehen jedoch auch Grenzen gegenüber. Auch wenn diese Methoden eine effizientere Art, Textdaten in großem Umfang zu analysieren, versprechen, erfordern sie viel Zeit und Geduld. Es ist wichtig, bei Bedarf Rücksprache mit Expert*innen zu halten, anstatt einfach einen vor-

handenen Code zu kopieren und einzufügen. Es empfiehlt sich auch immer Open-Source-Programme anstatt „One-fits-all"-Softwaretools zu verwenden. Open-Source-Programme oder -Bibliotheken erlauben Nutzer*innen Parameter zu ändern und diese auch einzusehen. Außerdem sind sie peer-reviewed und werden von einer Online-Community unterstützt, die sich gegenseitig hilft.

Der Fokus dieser Arbeit liegt auf der Darstellung eines Qualitätsleitfadens für Mixed-Methods-Ansätze in der Textanalyse. Die vorgestellten Methoden sind jedoch nur ein kleiner Ausschnitt der Möglichkeiten, die Text Mining und induktive Kodierung bieten. Es gibt noch viele weitere Möglichkeiten, textuelle Daten zu analysieren.

Eine weitere Einschränkung diese Arbeit ist, dass die moralischen Fragen, denen sich Forschende stellen müssen, nicht beleuchtet werden. Nur weil eine neue Datenmasse zur Analyse verfügbar ist, muss dies nicht immer ethisch oder moralisch vertretbar sein. Das Recht auf Anonymität ist auch hier wichtig zu sichern und zu gewährleisten.

Literatur

Antons, David, Eduard Grünwald, Patrick Cichy und Torsten O. Salge. 2020. The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Management* 50:329–351.

Apishev, Murat, Sergei Koltcov, Olessia Koltsova, Sergey Nikolenko und Konstantin Vorontsov. 2017. Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. In *Advances in computational intelligence. 15th Mexican International Conference on Artificial Intelligence, MICAI 2016 Cancun, Mexico, October 23-28, 2016 Proceedings, Part I.* Lecture Notes in Computer Science, Hrsg. Grigori Sidorov und Oscar Herrera-Alcántara, 169–184. Cham: Springer.

Blei, David, Andrew Y. Ng und Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Bohr, Jeremiah, und Riley E. Dunlap. 2018. Key Topics in environmental sociology, 1990–2014: Results from a computational text analysis. *Environmental Sociology* 4:181–195.

Carp, Frances M., und Abraham Carp. 1981. The validity, reliability and generalizability of diary data. *Experimental aging research* 7:281–296.

Croidieu, Grégoire, und Phillip H. Kim. 2018. Labor of Love: Amateurs and Lay-expertise Legitimation in the Early U.S. Radio Field. *Administrative Science Quarterly* 63:1–42.

DiMaggio, Paul, Manish Nag und David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* 41:570–606.

Dubois, Anna, und Lars-Erik Gadde. 2002. Systematic combining: An abductive approach to case research. *Journal of Business Research* 55:553–560.

Früh, Werner. 2015. *Inhaltsanalyse. Theorie und Praxis*, Bd. 2501. 8., überarbeitete Auflage. Konstanz, München: UVK Verlagsgesellschaft mbH; UVK / Lucius.

Giegler, Helmut. 1992. Zur computerunterstützten Analyse sozialwissenschaftlicher Textdaten: Quantitative und qualitative Strategien. In *Analyse verbaler Daten*, Hrsg. Jürgen H. P. Hoffmeyer-Zlotnik, 335–388. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gioia, Dennis A., Kevin G. Corley und Aimee L. Hamilton. 2013. Seeking Qualitative Rigor in Inductive Research. *Organizational Research Methods* 16:15–31.

Gioia, Dennis A., und James B. Thomas. 1996. Identity, Image, and Issue Interpretation: Sensemaking During Strategic Change in Academia. *Administrative Science Quarterly* 41:370–403.

Hannigan, Timothy R., Richard F. J. Haans, Keyvan Vakili, Hovig Tchalian, Vern L. Glaser, Milo S. Wang, Sarah Kaplan und P. D. Jennings. 2019. Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals* 13:586–632.

Kohlbacher, Florian. 2006. The Use of Qualitative Content Analysis in Case Study Research. *Forum Qualitative Sozialforschung* 7. https://doi.org/10.17169/fqs-7.1.75

Kuckartz, Anne, und Michael J. Sharp. 2011. Responsibility: A Key Category for Understanding the Discourse on the Financial Crisis—Analyzing the KWALON Data Set with MAXQDA 10. *Forum Qualitative Sozialforschung* 12.

Leung, Lawrence. 2015. Validity, reliability, and generalizability in qualitative research. *Journal of Family Medicine and Primary Care* 4:324–327.

Lukka, Kari, und Sven Modell. 2010. Validation in interpretive management accounting research. *Accounting, Organizations and Society* 35:462–477.

MacQueen, Kathleen M., Eleanor McLellan, Kelly Kay und Bobby Milstein. 1998. Codebook Development for Team-Based Qualitative Analysis. *Cultural Anthropology Methods* 10:31–36.

Marsh, Herbert W., Upali W. Jayasinghe und Nigel W. Bond. 2008. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *The American Psychologist* 63:160–168.

Mayring, Philipp. 2000. Qualitative Inhaltsanalyse. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 1(2), Art. 20, http://nbn-resolving.de/urn:nbn:de:0114-fqs0002204

Mayring, Philipp. 2015. *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. 12., überarb. Aufl. Weinheim: Beltz.

Mayring, Philipp. 2019. Qualitative Content Analysis. Demarcation, Varieties, Developments. *Forum Qualitative Sozialforschung* 20. https://doi.org/10.17169/fqs-1.2.1089

Roberts, Margaret E., Brandon M. Stewart und Dustin Tingley. 2019. stm : An R Package for Structural Topic Models. *Journal of Statistical Software* 91:1–40.

Roller, Margaret R. 2019. A Quality Approach to Qualitative Content Analysis. Similarities and Differences Compared to Other Qualitative Methods. *Forum Qualitative Sozialforschung* 20(3). https://doi.org/10.17169/fqs-20.3.3385

Schmiedel, Theresa, Oliver Müller und Jan vom Brocke. 2019. Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture. *Organizational Research Methods* 22:941–968.

Shimizu, Takumi. 2017. Material-Discursive Practices in Technology Standards Development: A Topic-Modeling Approach to Technology Evolution. *Twenty-third Americas Conference on Information Systems, Boston*. https://core.ac.uk/download/pdf/301371959.pdf

Stone, Philip J., Dexter Dunphy, Marshall Smith S. und Daniel M. Ogilvie. 1966. *The General Inquirer. A Computer Approach to Content Analysis*. Cambridge, Mass., & London: M.I.T. Press.

Tran, Nam K., Sergej Zerr, Kerstin Bischoff, Claudia Niederée und Ralf Krestel. 2013. „Gute Arbeit". Topic Exploration and Analysis Challenges for Corpora of German Qualitative Studies. *Proceedings of ENRICH 2013 – SIGIR 2013 Workshop*, 15–22.

Wang, Yinying, Alex J. Bowers und David J. Fikis. 2017. Automated Text Data Mining Analysis of Five Decades of Educational Leadership Research Literature: Probabilistic Topic Modeling of EAQ Articles from 1965 to 2014. *Educational Administration Quarterly* 53:289–323.

Wilden, Ralf, Jan Hohberger, Timothy M. Devinney und Dovev Lavie. 2018. Revisiting James March (1991): Whither exploration and exploitation? *Strategic Organization* 16:352–369.

Yin, Robert K. 1994. Discovering the Future of the Case Study. Method in Evaluation Research. *American Journal of Evaluation* 15:283–290.