Kapitel 6 – Digitale Daten

Margrit Schreier

Im Folgenden werden zwei Arten digitaler Daten genauer beschrieben: Webseiten und Blogs sowie Daten aus sozialen Medien. Zu weiteren Formen digitaler Daten wie etwa Online-Spielen, Sprachassistenten wie Siri oder Alexa oder den sog. Wearables wie Fitness-Armbändern liegen in der qualitativen Sozialforschung noch nicht genügend Erfahrungen vor.

Webseiten und Blogs

Sowohl Webseiten als auch Blogs können als **kommunikative Dokumente** gelten. **Webseiten** sind öffentlich zugänglich, tendenziell statisch aufgebaut und multimodal. Je nach Zweck der Webseite stehen jeweils mehr oder weniger Interaktionsmöglichkeiten zur Verfügung: Webseiten, die in erster Linie der Selbstdarstellung dienen, bieten – ebenso wie Webseiten aus der Frühzeit des Internet generell – eher wenig Kommunikationsmöglichkeiten. Webseiten von Organisationen sowie insbesondere kommerzielle Seiten bieten dagegen eine Reihe von Möglichkeiten der Kommunikation und Interaktion, wie beispielsweise die Nutzung von Kontaktformularen, Bestellung, Kauf- und Bezahloptionen sowie Kommentarfunktionen (etwa die Seiten von Zeitungen, Fernseh- oder Rundfunkanstalten oder von Webshops).

Im Gegensatz zu Webseiten sind **Blogs** tendenziell stärker textbasiert. Weiterhin befinden sie sich an der Schnittstelle von privaten und öffentlichen Inhalten: Je nachdem, ob sie in erster Linie von einer Privatperson zu privaten Zwecken oder von einer öffentlichen Person zum Zweck der öffentlichen Nutzung verfasst sind, können Blogs eher privaten oder eher öffentlichen Charakter haben. Blogs weisen außerdem in höherem Maß als Webseiten eine Kommentarfunktion auf, und innerhalb der Kommentare können sich zwischen Nutzer*innen eigenständige Diskussionen entwickeln. Bei der Nutzung von Daten aus Blogs im Rahmen der qualitativen Sozialforschung ist entsprechend darauf zu achten, dass die Bloginhalte und die Kommentare je separate Datenformen darstellen, so dass bei der Auswahl der Daten ggf. auch je unterschiedliche Kriterien anzulegen sind.

Die Form der **Datenerhebung** bei der Nutzung von Webseiten und Blogs richtet sich nach der Menge der zu erhebenden Daten. Bei einem umfangreichen Datencorpus wird die Datenerhebung üblicherweise automatisiert; dabei kommen sog. **Web Crawler** (zum Aufbau einer Liste relevanter Webseiten und deren Indexierung) sowie **Web Scraper** (zur automatisierten Extraktion der Inhalte) zur Anwendung (Ignatow & Mihalcea, 2017, Kap. 6). Wenn ein kleineres Datencorpus angestrebt wird, kann die Datenerhebung auch per Hand erfolgen. Auch bei einer Erhebung per Hand ist allerdings darauf zu achten, das Material regelmäßig zu sichern (aufgrund der Unbeständigkeit und Veränderlichkeit des Materials) und

Aus: Schreier, M., Echterhoff, G., Bauer, J. F., Weydmann, N. & Hussy, W. (2023). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (3. Aufl.). Springer.

in übersichtlicher Weise zu archivieren.

Bei der **Auswertung** kommen Verfahren zur Anwendung, wie sie auch zur Auswertung anderer verbaler und visueller Daten verwendet werden. Darüber hinaus ist bei der Auswertung – je nach Beschaffenheit der Daten – darauf zu achten, dem kommunikativen und dem multimodalen Charakter des Materials gerecht zu werden. Zur Berücksichtigung kommunikativer Aspekte eignen sich beispielsweise die Konversationsanalyse, sequenzanalytische oder argumentationsanalytische Verfahren. Für die Analyse multimodalen Materials sind in den letzten Jahren eigenständige Verfahren entwickelt worden.

Daten aus sozialen Medien

Daten aus sozialen Medien stellen den bisher größten Anteil der Nutzung digitaler Daten in der qualitativen Sozialforschung dar, wobei in der Mehrzahl der Studien der Schwerpunkt auf Facebook oder Twitter liegt.

Merkmale von Daten aus sozialen Medien

Daten aus sozialen Medien zeichnen sich zunächst dadurch aus, dass sie von den Nutzer*innen selbst erzeugt sind (**user-generated data**), wobei die Rollen der (professionellen) Produktion und der Nutzung von Material zunehmend verschwimmen. Weiterhin entstehen Daten in den sozialen Medien aus der Interaktion zwischen Nutzer*innen heraus; sie sind somit **interaktiv** und **aufeinander bezogen**. Dabei sind unterschiedliche Interaktionsformen möglich, etwa das Verfassen eigener Beiträge im Rahmen der Selbstdarstellung, das Hochladen und Weiterleiten von Material, das ‚Liken' der Beiträge anderer oder auch das ‚Lurking', d.h. das Verfolgen der Beiträge anderer, ohne sich selbst explizit an der Kommunikation zu beteiligen.

Innerhalb der sozialen Medien ist zwischen selbstverfassten Beiträgen und der Kommunikation über diese Beiträge zu unterscheiden. Die Art der verfassten Beiträge richtet sich nach der Art der Plattform: So handelt es sich bei Twitter beispielsweise um eine textbasierte, bei Facebook um eine multimodale, bei Instagram um eine bild- und bei Snapchat oder YouTube um eine primär videobasierte Plattform.

Die Kommunikation zwischen den Teilnehmer*innen erfolgt in Form kurzer Beiträge, die auch als **Chats** bezeichnet werden; Beiträge innerhalb eines Chats, die sich unmittelbar aufeinander beziehen, werden Threads genannt (wie Fäden der Diskussion, die sich durch einen Chat ziehen). Chats zeichnen sich durch ihren Stellenwert zwischen mündlicher und schriftlicher Kommunikation aus. Sie sind schriftlich verfasst, weisen aber den informellen Charakter von mündlicher Kommunikation auf. Die für die mündliche Kommunikation charakteristische Gestik und Mimik – die nicht zuletzt einen wesentlichen Interpretationskontext darstellt – wird im Chat durch die Verwendung von Emoticons, Emojis, Gifs, Memes und verschiedener stilistischer Mittel wie Großschreibung, Wiederholung von Buchstaben und Wörtern etc. realisiert, die ihrerseits Gegenstand der Analyse sein können (Nam, 2020).

Aus: Schreier, M., Echterhoff, G., Bauer, J. F., Weydmann, N. & Hussy, W. (2023). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (3. Aufl.). Springer.

Die Art und Weise der Kommunikation und die Interaktionsmöglichkeiten werden wesentlich durch die **technische Infrastruktur** der verschiedenen Plattformen bestimmt. Als Beispiele seien hier etwa die Begrenzung der Zeichenzahl bei Twitter, die Begrenzung der Länge von Videos bei Snapchat, die Möglichkeiten der Nutzung von Räumen bei Facebook, oder die Erstellung von Stories bei Facebook oder Instagram genannt. Bei der Durchführung einer Studie unter Nutzung von Daten aus sozialen Medien sind daher immer auch die technische Infrastruktur und die damit verbundenen Optionen und Restriktionen zu berücksichtigen.

Auswahl, Herunterladen und Auswertung von Material aus sozialen Medien

Entsprechend komplex gestalten sich auch die Entscheidungen bei der **Materialauswahl**. Neben Fragen der Zugänglichkeit und Öffentlichkeit von Material ist beispielsweise zu fragen, ob lediglich ‚gepostete' Beiträge oder auch die Kommentare zu diesen Beiträgen in die Studie einbezogen werden sollen. Wenn auch die Kommentare einbezogen werden, so sind auch hier entsprechende Kriterien zu formulieren (etwa Auswahl der Kommentare, die ihrerseits am häufigsten kommentiert oder ‚geliked' wurden; Auswahl der ersten Kommentare, der kontroversesten Kommentare usw.). Weitere Fragen stellen sich etwa dahingehend, ob Text- und Bildmaterial gleichermaßen (und nach welchen Kriterien) erhoben werden soll, ob auch stilistische Darstellungsmittel berücksichtigt werden u.ä. Ein grundsätzliches Problem bei der Materialauswahl aus sozialen Medien stellen Fake Profiles dar, also gefälschte Profile von angeblichen Nutzer*innen, die so gar nicht existieren. Auch die sog. sozialen Bots sind inzwischen bei der Materialauswahl zum Problem geworden. Dabei handelt es sich um Computerprogramme, deren Aktionen die Handlungen menschlicher Nutzer*innen in sozialen Medien imitieren, also Beiträge posten, weiterleiten, liken oder kommentieren.

Schlussendlich erfolgt die Auswahl von Material aus sozialen Medien meist mittels der Formulierung von **Suchanfragen**, die den zuvor festgelegten Kriterien in Bezug auf Erhebungszeiträume, Lokalisation der Nutzer*innen und Inhalte entsprechen. Dabei sind in der Regel einige Vorarbeiten erforderlich, bevor Suchanfragen formuliert werden können, die der jeweiligen Fragestellung und den Auswahlkriterien auch tatsächlich gerecht werden. Bei der Suche auf Plattformen wie Twitter, die sog. Hashtags verwenden, ist außerdem zu berücksichtigen, ob und in welchem Umfang auch Hashtags für die Suche genutzt werden sollen. Hashtags vermitteln einerseits gezielten Zugang zu relevanten Inhalten. Andererseits werden sie nicht von allen Nutzer*innen verwendet.

Der eigentliche Zugang zu den Daten kann entweder per Hand oder automatisiert erfolgen. Beim Zugang per Hand können Forscher*innen Daten aus ihrem eigenen Netzwerk herunterladen – die Zustimmung der entsprechenden Personen natürlich vorausgesetzt. Oder sie können Personen ansprechen, ihr Forschungsvorhaben erläutern und um Zugang zu den entsprechenden Daten bitten, etwa indem die Moderator*innen von Gruppen einwilligen, dass Forscher*innen dort ihr Vorhaben vorstellen. Die Sicherung der Daten erfolgt dann durch Copy and Paste oder auch durch sog. Screen Capturing Software.

Alternativ können Forscher*innen automatisiert auf Daten zugreifen und diese herunterladen. Dabei kommen APIs, sog. **Application Programming Interfaces**, zur Anwendung (Janetzko,

Aus: Schreier, M., Echterhoff, G., Bauer, J. F., Weydmann, N. & Hussy, W. (2023). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (3. Aufl.). Springer.

2017), über die der Zugang zu einer ausgewählten Plattform hergestellt wird. Deren Nutzung erfordert allerdings entsprechende programmiertechnische Kenntnisse. Außerdem werden die APIs von den Betreiber*innen der jeweiligen Plattformen kontrolliert. Auch bieten sie nicht Zugang zu sämtlichen Daten, die einer je konkreten Suchanfrage entsprechen, sondern nur zu einer algorithmisch bestimmten Auswahl (bei Twitter entspricht dies 1% der potenziell verfügbaren Daten). Die Algorithmen sind allerdings nur den Betreiber*innen der Plattformen bekannt; es ist also für die Forscher*innen nicht ersichtlich, nach welchen Kriterien die Daten zur Verfügung gestellt werden (oder auch nicht). Außerdem können die Ergebnisse einer Suchanfrage über API je nutzer- und profilspezifisch sein, wie wir das beispielsweise auch von Suchanfragen in Google kennen.

Weiterhin ist bei der Nutzung von Daten aus sozialen Medien zu beachten, dass jede Plattform in ihren **Nutzungsbedingungen** Vorgaben für die Nutzung von Daten zu Forschungszwecken macht, die im konkreten Fall zu beachten sind. Diese Bedingungen werden ständig überarbeitet, sollten daher regelmäßig überprüft werden. Auch stehen selbst über API nicht sämtliche Daten potenziell zur Verfügung: Twitter stellt beispielsweise keine historischen Daten bereit. Dieser Begrenzungen und der daraus resultierenden Einschränkungen bei der Beantwortung der Forschungsfrage sollten Forscher*innen sich bei der Planung und Durchführung ihrer Studie mit Daten aus sozialen Medien stets bewusst sein.

Für die **Auswertung** der Daten aus sozialen Medien steht – je nach deren Modalität – die gesamte Bandbreite an Verfahren zur Auswertung qualitativer Daten zur Verfügung. Je umfangreicher der Datencorpus ist, desto eher kommen darüber hinaus automatisierte Auswertungsverfahren in Frage, auch in Kombination mit vertiefenden qualitativen Auswertungen. Automatisierte Auswertungsverfahren, für die jeweils unterschiedliche Softwarelösungen existieren, umfassen unter anderem die Analyse von Worthäufigkeiten; von Häufigkeiten von Worten in ihrem jeweiligen Kontext (KWIC: keywords in context); die Sentiment Analysis zur Bestimmung der vorherrschenden Valenzen (positiv, negativ, neutral); die Topic Analysis, bei der auf der Grundlage von Worthäufigkeiten und Ko-Okkurrenzen von Worten Themenfelder gebildet werden, sowie Formen der Netzwerkanalyse (Ignatow & Mihalcea, 2017).

Beispiel

Die Rolle von Twitter nach dem Terrorangriff in Norwegen 2011

Am 22. Juli 2011 zündete Anders Breivik zunächst im Osloer Regierungsviertel eine Bombe. Anschließend erschoss er auf der Insel Utoya 69 Jugendliche, die an einem Ferienlager teilgenommen hatten. Insgesamt starben bei dem Terrorangriff 77 Menschen. In ihrer Studie zur Rolle sozialer Medien bei der Bewältigung eines solchen kollektiven Traumas untersuchte Moa Eriksson knapp 5.000 Tweets (2016). Dabei kombinierte sie eine breit gefächerte Inhaltsanalyse mit einer tiefergehenden Diskursanalyse einzelner relevanter Tweets.

Es wurden Tweets in die Studie aufgenommen, die in den zwei Wochen nach dem

Aus: Schreier, M., Echterhoff, G., Bauer, J. F., Weydmann, N. & Hussy, W. (2023). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (3. Aufl.). Springer.

Terroranschlag verfasst worden waren; weiterhin mussten die Tweets mindestens einen der folgenden Hashtags enthalten, die durch Vorarbeiten als einschlägig ermittelt wurden: #oslo, #oslexpl, #osloexplosion, #explosion, #osloblast, #utøya, #utoya, #utöya, #utoeya, #breivik, #prayfornorway and/or #terrorist. Zum Herunterladen der Tweets wurde das API von Sifter verwendet, das inzwischen allerdings nicht mehr zur Verfügung steht. Mittels Sifter konnten 445,786 Posts heruntergeladen werden, aus denen eine geschichtete Zufallsstichprobe gezogen wurde; Kriterium für die Schichtung war die proportionale Anzahl von Tweets pro Tag nach dem Anschlag, die im Laufe der ersten Woche schnell abnahm. Nach der Entfernung von Doubletten (‚doppelten' Tweets) verblieben noch 4282 Tweets für die weitere Analyse.

Die Inhaltsanalyse ergab vier zentrale Themen: Unmittelbar nach den Anschlägen standen die norwegische Nation und Solidaritätsbekundungen im Vordergrund; im Laufe der Zeit nahmen Erklärungsversuche zunehmend mehr Raum ein. Eine weitere, kleinere Gruppe von Tweets befasste sich mit den Details der Anschläge. Die anschließende Diskursanalyse konzentrierte sich auf die Erklärungsversuche, und hier vor allem auf die beiden größten Gruppen: Tweets zur Rolle der traditionellen Medien und zum Rechtsextremismus. Eine genauere Analyse der Tweets zur Rolle der Medien zeigt anschaulich die Relation zwischen Twitter als ‚Backchannel' und permanentem Kommentar insbesondere zu den Printmedien. Hier finden sich vielfach sowohl Ironisierungen der Berichterstattung als auch detaillierte Sprachkritik, etwa an der New York Times. Diese bezeichnete Breivik zunächst als „terrorist". Nachdem sich herausgestellt hatte, dass es sich bei Breivik nicht um einen Mann mit muslimischen, sondern mit einem rechtsextremen Hintergrund handelte, schrieb die Times plötzlich nicht mehr vom Terroristen, sondern vom ‚gunman', also dem bewaffneten Mann. Hierzu lautet ein Kommentar beispielsweise: „Media calls him a 'gunman' but if its a muslim, he's a #terrorist along with all other muslims! #hypocrisy #oslo". Die Studie ist somit ein Beispiel dafür, wie bei der Verwendung digitaler Daten eine eher oberflächliche breite mit einer tiefer gehenden qualitativen Analyse eines Teils des Materials kombiniert werden kann.

Aus: Schreier, M., Echterhoff, G., Bauer, J. F., Weydmann, N. & Hussy, W. (2023). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (3. Aufl.). Springer.

Literatur

Eriksson, M. (2016). Managing collective trauma on social media: the role of Twitter after the 2011 Norway attacks. *Media, Culture, & Society, 38*(3), 365–380.

Ignatow, G. & Mihalcea, R. F. (2017). *An introduction to text mining: Research design, data collection, and analysis*. Sage.

Janetzko, D. (2017). The role of APIs in data sampling from social media. In L. Sloan & A. Quan-Haase (Eds.), *The Sage handbook of social media research methods* (pp. 146–160). Sage.

Nam, S.-H. (2020). Qualitative Analyse von Chats und anderer usergenerierter Kommunikation.. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (2. Aufl., Bd. 2, S. 1041–1052). Springer VS.

Aus: Schreier, M., Echterhoff, G., Bauer, J. F., Weydmann, N. & Hussy, W. (2023). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (3. Aufl.). Springer.