

Quantitative Analyse von Twitter und anderer usergenerierter Kommunikation

90

Jochen Mayerl und Thorsten Faas

90.1 Warum Twitter und Co. in der empirischen Sozialforschung?¹

Die Nutzung von sozialen Medien („Social Media“) ist ein wichtiger Bestandteil der alltäglichen Kommunikation von Millionen von Menschen. Mitunter werden Beiträge in diesen Medien selbst zu tagespolitischen Ereignissen. Letzteres galt in besonderem Maße für die Reichweite und das Echo auf die täglichen Tweets des US-amerikanischen Präsidenten Donald Trump. All das hat nicht nur Folgen für die Strukturen moderner Öffentlichkeit, sondern auch für die empirische Sozialforschung, denn in diesen Netzwerken liegt ein riesiger Datenpool für die empirische Sozialforschung.

Was zeichnet nun diese Netzwerke und ihre Daten aus? Ging es im Zeitalter des Web 1.0 in erster Linie darum, Internet-Angebote im Sinne einseitiger Informationsflüsse zu verwenden, also Informationen von einem Sender zu Empfängern zu transportieren, lebt das Web 2.0 von der Interaktion. Von einem Nutzer ausgesendete Informationen können somit von anderen nicht nur empfangen, sondern auch kommentiert, ergänzt und verändert werden. Das Internet ist so zu einer Plattform geworden, auf der Inhalte und Anwendungen nicht länger von einer Person kreiert und veröffentlicht werden, sondern häufig von vielen Nutzern in einem kollaborativen Akt kontinuierlich verändert werden können. Mit diesem Wandel wurde die Basis für Social Media-Anwendungen gelegt.

Unter dem Begriff „Social Media“ werden zunächst einmal alle internetbasierten Anwendungen gefasst, die es Nutzern erlauben, Inhalte zu kreieren und auszutauschen, also etwa die Online-Enzyklopädie Wikipedia, die große Bandbreite von Blogs (Schmidt, Kapitel 89 in diesem Band), Foren (Schünzel/Traue und Nam, Kapitel 88 und 91 in diesem Band) ebenso wie soziale Online-Netzwerke wie Telegram, Insta-

1 Unser Dank geht an Dominic Nyhuis, Petra Lipski, Antön Könecke und Jonathan Kohl für die kritische Durchsicht einer früheren Version dieses Beitrags.

gram, Facebook (Schrape/Siri, Kapitel 92 in diesem Band) und Twitter. Abstrakt gesprochen sind soziale Online-Netzwerke thematisch nicht begrenzte Kommunikationsplattformen, in denen Nutzer eindeutig identifizierbare Profile anlegen. Diese Profile umfassen Inhalte, die der Inhaber des Profils selbst generiert, sowie Verbindungen zu anderen Nutzern der Plattform. Jeder Nutzer kann sowohl Inhalte produzieren als auch Inhalte von anderen konsumieren oder in Interaktion mit anderen Nutzern treten. Über Facebook und Twitter hinaus ist dabei auch an Rezensionen auf Bewertungsportalen zu Hotels oder Bewertungen bei Internetauktionen zu denken.

Jedes soziale Online-Netzwerk hat bestimmte charakteristische Züge – bei Twitter etwa die Begrenzung der Textlänge auf maximal 280 Zeichen (bis November 2017 waren es noch 140 Zeichen). Weltweit gibt es ca. 217 Millionen monatlich aktive Nutzer auf Twitter (Stand: 4. Quartal 2021), zwei Milliarden Nutzer auf Instagram (Stand: Dezember 2021) und 2,9 Milliarden Nutzer von Facebook (Stand: 4. Quartal 2021).

Der Charme von Twitter besteht darin, dass Tweets mit Metamarkierungen versehen sind, die eine schnelle Grob-Kategorisierung sowie Verweise auf andere Tweets und Nutzer ermöglichen. Zu diesen Metamarkierungen gehören sog. Hashtags (#), mit deren Hilfe man sich alle Beiträge zu einem bestimmten Thema anzeigen lassen kann, Verweise auf andere Nutzerprofile (@) sowie wiedergegebene Retweets (RT). Ein Retweet ist schlicht die wiederholte Wiedergabe eines Tweets anderer Nutzer über das eigene Nutzerprofil. Abbildung 90.1 zeigt beispielhaft einen Screenshot eines Tweets. Die Nachricht wurde über das Nutzerprofil „@CIA“ abgesendet (das blaue Häkchen hinter dem Namen zeigt an, dass es sich tatsächlich um den offiziellen Account der US-amerikanischen Central Intelligence Agency CIA handelt), es findet



Abb. 90.1 Beispiel-Tweet vom 7. Juni 2014

sich im Text ein Verweis auf ein Thema (Hashtag) „#unclassified“ sowie einen anderen Nutzer (@Twitter), die Meldung wurde bis zum Zeitpunkt des Screenshots 13 215 Mal von anderen Nutzerprofilen weiterverbreitet (Retweets), 16 082 Mal mit „Gefällt mir“ markiert, 2 800 Mal wurde der Tweet von anderen Nutzern kommentiert.

Nicht ohne Grund sind die Daten aus sozialen Netzwerken – Profile ebenso wie Verknüpfungen zwischen Profilen – als für die Forschung vergleichsweise leicht zugänglicher Teil der Big Data-Ära (Trübner/Mühlichen, Kapitel 10 in diesem Band) zu sehen; für die quantitative Sozialforschung (Stein, Kapitel 8 in diesem Band) kann das ein echter Schatz sein.

90.2 Mögliche Erkenntnisgewinne und Anwendungsgebiete

Usergenerierte Kommunikationsdaten ermöglichen der empirischen Sozialforschung den vergleichsweise leichten Zugang zu einer großen Menge an aktuellen Meinungsäußerungen von Nutzern, ohne diese Personen befragen zu müssen. Stattdessen handelt es sich um prozessgenerierte Daten (Salheiser, Kapitel 104 in diesem Band) in Form von digitalen Verhaltensspuren. Noch während soziopolitische Ereignisse stattfinden (z. B. Filmausstrahlungen, Protestkundgebungen oder Fernsehdebatten) können Daten zu individuellen Bewertungen, Beschreibungen, Sorgen, Hoffnungen und Argumenten erhoben und ausgewertet werden (sog. Sentiment-Analyse, vgl. Siegel/Alexa 2020). Dies ist in Echtzeit bzw. mit extrem kurzer Verzögerung möglich, aber ggf. auch nachträglich über entsprechende Archive, die mit sog. „Crawlern“ durchsucht werden können (vgl. Abschnitt 90.3.3). Auf diese Weise können erhebliche Mengen an Daten über individuelle Überzeugungen und Erfahrungsberichte auf kostengünstige Weise erfasst werden, was mit den klassischen Datenerhebungsverfahren der empirischen Sozialforschung in dieser Form nicht möglich ist.

Sozialwissenschaftliche Anwendungen zur Analyse solcher Daten, insbesondere von Twitter-Meldungen, liegen mittlerweile in einer beachtlichen Menge zu einer Vielzahl an soziopolitischen Themen vor. Zu diesen Studien zählen u. a. Arbeiten aus dem Bereich des Public Health. So zeigen Aramaki et al. (2011) in einem Methodenvergleich zur Identifikation von Influenza-Epidemien, dass Twitter-Analysen zum Teil besser abschneiden als die State-of-the-Art-Methoden, zu denen offizielle Daten des Medikamentenverkaufs in Apotheken sowie Log-File-Analysen (Schmitz/Yanenko, Kapitel 87 in diesem Band) der Suchmaschine Google gehören. Ein Methodenvergleich von Eichstaedt et al. (2015) zur Vorhersage von atherosklerotischen Herzerkrankungen zeigt, dass Regressionsanalysen mit Twitter-Daten herkömmliche Analysen mit soziodemographischen Prädiktorvariablen übertreffen. Instagram-Daten wurden u. a. zur Analyse von vorsätzlicher Selbstverletzung (Scherr et al. 2020) und Depressionen (McCosker/Gerrard 2021) eingesetzt.

Im Bereich der politischen Soziologie werden usergenerierte Kommunikationsdaten in der Wahlforschung zur Wahlvorhersage eingesetzt. Allerdings wird daran

auch deutlich, dass noch Herausforderungen zu bewältigen sind: Gayo-Avello (2013) kommt mit einer Meta-Analyse (Weiß/Wagner, Kapitel 54 in diesem Band) von Twitter-basierten Wahlprognosen zu dem Schluss, dass Twitter-Analysen begleitend zwar hilfreich sein können, aber herkömmliche Umfragemethoden der Wahlforschung derzeit noch nicht ersetzen können. Im Bereich der Bewertung von politischen TV-Duellen zeigen Nyhuis und Friederich (2017), dass man Tweets zum Zwecke von Echtzeitreaktionen von Rezipienten nutzen kann, wie man sie sonst nur unter Laborbedingungen realisieren kann. Im Themenbereich der Covid-19-Pandemie untersuchen Hohlfeld et al. (2021) die Kommunikation zu Verschwörungstheorien über Facebook und Telegram, während Machuca et al. (2021) eine Sentiment-Analyse von Tweets über den Coronavirus vorlegen.

Ein weiteres Anwendungsgebiet, das die Vorteile der Echtzeit von Twitter-Daten hervorhebt, zeigen Middleton et al. (2014). Sie nutzen Twitter-Daten zum „Real-Time Crisis Mapping“ bei Naturkatastrophen. Grundsätzlich bezeichnet „Crisis Mapping“ die Kartographierung von (Natur-)Katastrophen oder Ereignissen von sozialen bzw. politischen Konflikten, was sowohl zeitlich versetzt als auch in Echtzeit erfolgen kann. Soziale Medien werden dabei zu einem kollektiven virtuellen Sensor für Krisensituationen, der unterfüttert mit Geodaten (Lakes, Kapitel 118 in diesem Band) ein besseres Krisenmanagement ermöglichen kann. Fownes et al. (2018) beschäftigen sich mit der Frage, wie auf Twitter über den Klimawandel diskutiert wird, und kommen zu dem Schluss, dass sich das Netzwerk zur Analyse von Einstellungsänderungen zum Klimawandel eignen kann. Im Bereich der Kriminologie zeigt Gerber (2014), dass Twitter-Daten mit Geoinformationen die präzise Vorhersage von Kriminalität bedeutsam erhöhen können. Li et al. (2021) zeigen, wie illegaler Drogenhandel über Instagram detektiert werden kann. Die Arbeit von Mitchell et al. (2013) stammt aus dem Bereich der Forschung zum persönlichen Wohlbefinden. Mitchell und Kollegen zeigen dabei am Beispiel der USA, wie auf Aggregatebene (z. B. in Städten) (Graeff, Kapitel 102 in diesem Band) geodatenfundierte Twitter-Beiträge als nützliche Erweiterungen zu Zensus-Daten (Hartmann/Lengerer, Kapitel 106 in diesem Band) beitragen können, um das Ausmaß und Veränderungen von populationsbezogenen Maßen des Wohlbefindens in Echtzeit zu ermitteln und diese geographisch zu verorten (sog. Sentiment-Analyse). Diekmann et al. (2014) haben Nutzer-Bewertungen bei Online-Auktionen zur Analyse von Reputationseffekten und Kooperationsverhalten in anonymen digitalen Märkten eingesetzt. Daneben eröffnen soziale Netzwerke wie Twitter naturgemäß innovative Möglichkeiten, Netzwerkanalysen (Baur, Kapitel 99 in diesem Band) durchzuführen.

90.3 Erhebung von usergenerierten Daten

Die empirische Analyse von usergenerierten Daten erfordert in der Regel zunächst sechs Schritte zur Datenerhebung, die in den folgenden Unterkapiteln erläutert werden. Sie entsprechen in weiten Teilen den Verfahrensschritten, die bei allen quantitativen Inhaltsanalysen (Klein, Kapitel 108 in diesem Band) erforderlich sind, sodass wir uns hier auf spezifische Punkte der Analyse usergenerierter Social-Media-Daten konzentrieren. Um es schematisch vorwegzunehmen, geht es um folgende Fragen:

- Welche Variante(n) usergenerierter Kommunikation, welches soziale Netzwerk sollen in die Untersuchung einbezogen werden – Twitter, Facebook, Instagram, Chats?
- Sind die einzelnen Beiträge von Interesse oder eher die Netzwerkstrukturen zwischen Profilen und Beiträgen?
- Wessen Beiträge stehen im Zentrum des Interesses, wer oder was gehört zur Grundgesamtheit?
- Wie werden genau die Informationen aus den (vielen!) Beiträgen extrahiert, die im Zentrum des Interesses stehen?
- Erfüllt das alles letztlich etablierte Gütekriterien der empirischen Sozialforschung (Krebs/Menold, Kapitel 35 in diesem Band)?

90.3.1 Bestimmung der Datenquelle(n): Sachliche Eingrenzung der Zielpopulation

In einem ersten Schritt ist die Datenquelle zu bestimmen, was primär durch die inhaltliche Fragestellung bedingt sein sollte, gerade im Bereich sozialer Netzwerke aber auch durch Verfügbarkeitsaspekte bestimmt wird: Twitter-Daten etwa sind einfacher zugänglich als Facebook-Daten. Die nötige *sachliche Eingrenzung* erfolgt dabei in mindestens zwei Schritten:

1. *Auswahl des Netzwerks/Portals*: In einem ersten Schritt wird man sich typischerweise für einige wenige (oder gar nur ein einziges) soziale(s) Netzwerk(e) oder Portal(e) entscheiden (zur Fallauswahl bei Einzelfallanalysen siehe Hering/Jungmann, Kapitel 42 in diesem Band). Potenziell infrage kommen hier tausende Anbieter von Kommunikationsdiensten, Chat-Kanälen, Diskussionsforen oder auch User-Bewertungen (z. B. auf Amazon, Partner- oder Urlaubsportalen). Konkret: Wer sich für Chats auf Partnerbörsen interessiert, kann eine, wenige oder alle Börsen einbeziehen wollen. Die Entscheidung ist nicht ohne Folgen, da sich die Nutzer zwischen verschiedenen Partnerbörsen erheblich unterscheiden (sollten).
2. *Auswahl der einzelnen Beiträge*: Nach dieser Entscheidung folgt auf der zweiten Ebene innerhalb der ausgewählten Netzwerke oder Portale die Auswahl einzel-

ner Beiträge (oder Beziehungen zwischen ihnen) nach jeweils einschlägigen Kriterien.

Die resultierende Datenstruktur hat damit Züge einer Mehrebenenanalyse (Pötschke, Kapitel 53 in diesem Band): Zu Merkmalen auf der Ebene der Netzwerke oder Portale kommen Merkmale auf der Ebene einzelner Beiträge. Für beide Ebenen stellen sich dabei zuvorderst Fragen der *Auswahl(logik)*:

1. Auf der *Ebene der Netzwerke/Portale* erleichtert etwa ein Fokus auf Twitter alleine den Zugang zu den Daten, impliziert aber eine nicht-zufällige Auswahl von bestimmten Nutzern. Eine Einbeziehung verschiedener sozialer Netzwerke wirft hingegen Fragen der Verfügbarkeit und Vergleichbarkeit der Daten auf.
2. Auf der *Ebene einzelner Beiträge* stellt sich die Frage, ob eine Vollerhebung – ggf. auch über verschiedene Anbieter hinweg – möglich ist. Ist das nicht der Fall, sind geeignete Verfahren der Ziehung von (Zufalls-)Stichproben (Häder/Häder, Kapitel 27 in diesem Band) zu prüfen.

90.3.2 Räumliche und zeitliche Abgrenzung der Grundgesamtheit

Kommen wir zum zweiten Schritt: Facebook wurde 2004 gegründet, Twitter 2006. Beides sind soziale Netzwerke mit weltweiter Reichweite. Das wirft die Frage auf: Wie kann und sollte die Eingrenzung auf die Zielpopulation, den Untersuchungszeitraum und die regionale/sprachliche Abdeckung erfolgen? Denkt man an die Wahlforschung etwa vor einer Bundestagswahl, so würde man bei einer repräsentativen Umfrage mit Zufallsauswahl die wahlberechtigte Bevölkerung Deutschlands als Grundgesamtheit ansteuern. Wie aber soll das auf Twitter gelingen? Deutschsprachige Tweets könnten aus Österreich kommen, englischsprachige Tweets auch von Deutschen. Dieser Problematik kann zwar teilweise mit der Berücksichtigung von Geodaten, die über die Herkunft der Tweets informieren, begegnet werden. Aber auch dann ist eine genaue Definition der Zielpopulation nicht möglich, da noch immer die Staatsangehörigkeit und damit die Wahlberechtigung der Nutzer unklar bleiben und Wahlberechtigte auch im Ausland wohnhaft sein können. Auch an eine annähernde Bevölkerungsrepräsentativität ist noch nicht zu denken (vgl. Abschnitt 90.4.2). Dennoch sollte angestrebt werden, die räumliche und zeitliche Abgrenzung der Grundgesamtheit so genau wie möglich zu bestimmen.

90.3.3 (Halb-)automatische Extraktion von Informationseinheiten (z. B. Tweets)

Der dritte Schritt ist die Extraktion der Tweets nach bestimmten Kriterien mittels sog. „Crawlern“, durch „Scraping“ (d. h. das automatisierte Auslesen von Texten) oder via „API“ (Application Programming Interface). Im Detail unterscheidet sich das Vorgehen erheblich voneinander, allerdings finden sich gute „How-to“-Erläuterungen in der Literatur (siehe etwa Munzert et al. 2014: Kapitel 14).

Allgemein gesprochen geht es mit Blick auf Beiträge von Nutzern darum, durch die Angabe von bestimmten Wörtern oder Wortkombinationen (ggf. mit den Operatoren „AND“, „OR“ oder mit Hilfe der Trunkierung „*“) einen Pool an Informationseinheiten zu bestimmen, der inhaltlich untersucht werden soll. So haben beispielsweise Aramaki et al. (2011) mit dem Begriff „influenza“ die Anzahl von 300 Millionen Tweets in ihrem Untersuchungszeitraum auf 400 000 Tweets reduziert.

Wie eingangs beschrieben, sind Twitter-Meldungen mit Metamarkierungen wie Verweisen auf andere Nutzerprofile (@), Retweets (RT) und Hashtags (#) versehen, die eine Recherche von Twitter-Meldungen zu bestimmten Themen im Schneeballverfahren ermöglichen. Natürlich sollte dies nicht das allein ausschlaggebende Kriterium dafür sein, Twitter als Grundlage für empirische Analysen von sozialen Medien auszuwählen. An Daten aus Facebook kommt man jedoch bei weitem nicht so leicht und automatisierbar heran, weswegen hier mehr Nacharbeit nötig ist.

90.3.4 Manueller Ausschluss ungeeigneter Informationseinheiten (optional)

Die automatisiert ermittelte Menge an Tweets enthält natürlich häufig Informationseinheiten, die nichts mit dem Untersuchungsthema zu tun haben und die in vielen Fällen, wenn auch nicht immer, im Zuge der Datenaufbereitung (Lück/Landrock, Kapitel 33 in diesem Band) durch Kodierer in einem vierten Schritt manuell aussortiert werden müssen.

90.3.5 Differenzierung durch manuelle (oder halb-automatische) Kodierung

Der fünfte Schritt bedeutet häufig die zeitintensivste Arbeit durch Kodierer. Je nachdem, ob ein Kodierschema bereits vorliegt oder aber aus den Tweets Kategoriensysteme explorativ zu entwickeln sind, sind hier Iterationen zwischen den Schritten 3 bis 5 notwendig.

Bei sogenannten *Valenzanalysen* bewerten Kodierer etwa die Tweets entlang von Kategorien wie „positiv“, „negativ“ oder „neutral“. Häufig wird, nachdem ca. 10 % der

Informationseinheiten kodiert wurden, das bis dahin gesammelte Material beurteilt und gegebenenfalls zu Schritt 3 zurückgekehrt. Je nach Anzahl der zu kodierenden Tweets ist es gerade im Hinblick auf die Big Data-Problematik (Trübner/Mühlichen, Kapitel 10 in diesem Band) sowie der schier Menge an Informationseinheiten möglich, dass dieser Schritt nicht mehr komplett manuell erfolgen kann, sodass eine Kombination aus automatischen und manuellen Selektions- und Kategorisierungsschritten notwendig wird (Lewis et al. 2013) – wobei der Reiz der „Big Data“ oft genau darin besteht, die Masse an Daten automatisiert zu bewältigen.

90.3.6 Evaluation und Gütekriterien

Der sechste und letzte Schritt der Datenerhebung betrifft die Evaluation der Datenqualität (Krebs/Menold, Kapitel 35 in diesem Band). Wie bei anderen quantitativen inhaltsanalytischen Verfahren üblich, sollte auch hier die Intra- oder Interkodierer-Reliabilität geprüft werden. Ersteres erfordert das wiederholte Kodieren einer Auswahl an Tweets oder anderen Informationseinheiten durch denselben Kodierer, während für letzteres mehrere Kodierer denselben Pool an Tweets kodieren (auch hier zumeist wieder eine Auswahl an Tweets, um den Arbeitsaufwand zu verringern), um so eine hohe Reliabilität der gewonnenen Daten zu gewährleisten.

90.4 Methodologische Herausforderungen und Einschränkungen

Die Anwendung von usergenerierten Daten aus dem Web 2.0 für die empirische Sozialforschung ist mit einer Reihe von methodischen Problemen konfrontiert, die je nach Art der Datenquelle unterschiedlich stark ausgeprägt sind.

90.4.1 Güte

Zunächst stellt sich bei Analysen usergenerierter Kommunikationsdaten die Frage, ob und wie wissenschaftliche Standards wie Gültigkeit (Validität), Zuverlässigkeit (Reliabilität), Wertfreiheit und intersubjektive Nachprüfbarkeit (Krebs/Menold und Flick, Kapitel 35 und 34 in diesem Band) eingehalten werden können:

- Die analysierten Daten wurden nicht nach Qualitätskriterien der empirischen Sozialforschung erhoben, – wie z.B. bei Befragungen (Reinecke, Kapitel 62 in diesem Band) oder Beobachtungsstudien (Thierbach/Petschick, Kapitel 109 in diesem Band) –, sondern sind als prozessproduzierte Rohdaten (Salheiser, Kapitel 104 in diesem Band) ideologisch geprägt, oft stark verkürzt und zumindest zum Teil unwissentlich oder wissentlich falsch (sog. „Fake News“).

- Da Nutzer mehrere Profile anlegen können, ist zudem nicht gesichert, in welchem Ausmaß sich einzelne Personen mehrfach unter unterschiedlichen Profilnamen beteiligen, d. h. die Unabhängigkeit der Beobachtungseinheiten ist nicht sichergestellt, sie ist auch kaum überprüfbar.
- Hinzu kommt, dass mit Social Bots in großem Ausmaß „Meinungs-Roboter“ eingesetzt werden (Hegelich 2016; Pfaffenberger 2021), sodass unklar bleibt, inwieweit hinter jeder Informationseinheit ein menschlicher Akteur oder eine lernfähige Maschine steht. Möglichkeiten und Verfahren zur Identifikation von Bots werden derzeit kontrovers diskutiert, wobei als besondere Herausforderung natürlich hinzukommt, dass die Macher von Bots ihr Vorgehen in dem Moment auch wieder anpassen können, sobald die aktuell eingesetzten Kriterien zur Bot-Identifikation publik werden. Inhaltlich entscheidend ist die Frage, ob letztlich Bot-generierte Daten in Analysen einbezogen werden sollen oder nicht. Im Falle von Effektstudien etwa sollte man Bots eher einbeziehen.
- Nach Hsieh/Murphy (2017) sind bei Twitter-Analysen insbesondere drei Fehlerquellen zu unterscheiden, die die Stichprobenqualität und mögliche Messfehler betreffen: Coverage-Fehler, (d. h. die Zielpopulation unterscheidet sich von den Twitternutzern, vgl. Abschnitt 90.4.2), Fehler bei der Extraktion von Informationen, (d. h. es fehlen wichtige Schlagworte oder es werden solche verwendet, die irrelevante Tweets einbinden oder relevante Tweets ausschließen, vgl. Abschnitt 90.3.3), und Interpretationsfehler, (d. h. Fehler der Valenzzuweisung durch Kodierer oder durch (halb-)automatische Algorithmen, vgl. Abschnitt 90.3.5)

Im Zuge der Replizierbarkeit und Nachprüfbarkeit der Ergebnisse sind die gesammelten Informationseinheiten zu dokumentieren und zu archivieren (Mochmann, Kapitel 17 in diesem Band). Das ist besonders dann problematisch, wenn Beiträge auf Homepages (Schünzel/Traue, Kapitel 88 in diesem Band) früher oder später gelöscht werden oder ganze Dienste/Web-Auftritte nicht mehr online sind. Diese Temporalität der Datenquellen führt teilweise zu hohem Archivierungsaufwand, wenn z. B. mehrere Millionen Tweets oder Hunderte von Chat-Protokollen zu archivieren sind. Erst recht gilt dies, wenn die Extraktionslogik im Prozess geändert oder verfeinert wurde, ggf. sogar durch händisches Zutun.

Viele weitere methodische Herausforderungen sind deckungsgleich mit denjenigen der quantitativen Inhaltsanalyse, sodass diese hier nicht gesondert in aller Ausführlichkeit diskutiert werden müssen. Wichtig ist dabei die Reliabilität der manuellen Kodierungen, was durch Verfahren der Inter- und Intra-Kodierer-Reliabilität geprüft werden kann.

90.4.2 Stichprobenqualität und Verallgemeinerbarkeit

Weitere Probleme betreffen die Qualität der Stichprobenverfahren (Häder/Häder, Kapitel 27 in diesem Band). Häufig ist die Grundgesamtheit nicht oder nur vage zu bestimmen – dies gilt sowohl für die Auswahl der Anbieter als auch die Auswahl der Beiträge, sodass es prinzipiell schwierig wird, inferenzstatistische Schlüsse zu ziehen. Vor diesem Hintergrund ist auch Vorsicht geboten, wenn mitunter mit „Vollerhebungen“ argumentiert wird. „Wir haben alle Tweets mit dem Hashtag #tvduell gesammelt“ ist ein großer Satz. Aber was sagt er uns letztlich über das Duell und die Wahl, wenn nicht bekannt ist, wer überhaupt in Deutschland twittert und wer speziell an diesem Abend zum TV-Duell getwittert hat?

Allgemein gilt: Twitter, Chat- und Social Media-Nutzer sind eben nur ein bestimmter, zudem nicht klar bestimmbarer, aber sicherlich verzerrter Ausschnitt aus der Bevölkerung, weswegen die Anwendung von Verfahren der schließenden Statistik auf eine bevölkerungsbezogene Grundgesamtheit vermieden werden sollte. Grundsätzlich treffen hierbei die Probleme von Web-Surveys (Wagner-Schelewsky/Hering, Kapitel 70 in diesem Band) auch auf die Analyse von usergenerierten Kommunikationsdaten zu, insbesondere, dass weiterhin nicht alle Personen in Deutschland Internet-Zugang haben. Laut der ARD/ZDF-Online-Studie aus dem Frühjahr 2021 verfügen 94 % der deutschsprachigen Wohnbevölkerung ab 14 Jahre über einen Internetzugang und 76 % nutzen das Internet täglich (vgl. Beisch/Koch 2021). Dabei sind insbesondere ältere Personen ab 70 Jahre mit einer Quote von 77 % Internetnutzern weiterhin unterrepräsentiert. Wenn es um die mindestens wöchentliche Nutzung von sozialen Medien geht, so nutzen derzeit nur 28 % der deutschsprachigen Bevölkerung mindestens wöchentlich Facebook, und nur 4 % mindestens wöchentlich Twitter. Unter den jüngeren Alterskohorten (14–29 Jahre) sind es bei Facebook 35 % und bei Twitter 9 %, während Personen ab 70 Jahren Nutzungsquoten von lediglich 8 % bei Facebook und lediglich 1 % bei Twitter aufweisen. Entsprechend sind derzeit grundsätzlich keine annähernd repräsentativen Bevölkerungsanalysen mit solchen Daten möglich.

90.4.3 Datenschutz

Mit der Analyse von Social Media-Daten sind auch Probleme von Verletzungen des *Datenschutzes* betroffen. Sie werfen ethische Fragen auf (Meeder et al. 2010, vgl. hierzu auch Friedrichs, Mühlichen und Dreier, Kapitel 21, 23 und 25 in diesem Band), die teilweise identisch mit anderen Verfahren und Ansätzen, teils aber auch spezifischer Natur sind. Zudem besteht in umgekehrter Richtung das Problem, dass *Datenverknüpfungen* (Cielebak/Rässler, Kapitel 31 in diesem Band) häufig nicht möglich sind, wenn die Nutzer über anonyme Profile kommunizieren, sodass keine soziodemographischen Daten (Hoffmeyer-Zlotnik/Warner, Kapitel 77 in diesem Band) oder andere

Daten mit den Sendern der Nachrichten verknüpft werden können. Über die Analyse einer Vielzahl von Tweets eines Profils kann versucht werden, demographische Daten zu schätzen (Sloan et al. 2013), zudem existieren *Algorithmen zur De-Anonymisierung* (z. B. Narayanan/Shmatikov 2009). Alleine die Möglichkeit ihres Einsatzes wirft unmittelbar Datenschutz- und damit ethische Fragen auf. Entsprechend sollten solche Studien durch *Ethikkommissionen*, die zunehmend an Universitäten als Gremien verankert sind, evaluiert und abgesichert werden.

90.4.4 Operationalisierung von latenten Konstrukten

Eine weitere methodische Einschränkung ist die häufig problematische Operationalisierung von latenten Konstrukten. Im Unterschied zu klassischen Primärerhebungen (z. B. Surveys), bei denen alle möglichen individuellen Wahrnehmungen, Ängste, Sorgen, Verhaltensabsichten, Einstellungen, Erwartungen etc. in standardisierten Skalen (Latcheva/Davidov und Blasius, Kapitel 78 und 49 in diesem Band) mittels multipler Indikatoren (Burzan und Huinink, Kapitel 47 und 48 in diesem Band) erhoben werden, bleibt bei der Analyse von usergenerierten Kommunikationsdaten nur die indirekte Messung mittels geeigneter Kategoriensysteme. Ob sich die Analyse usergenerierter Kommunikationsdaten für eine Studie eignet, hängt demnach vor allem von der inhaltlichen Fragestellung ab. Interessant ist dabei auch die mögliche Verknüpfung von Twitter- und anderen Social Media-Daten mit klassischen Befragungsmethoden.

90.5 Usergenerierte Kommunikation und Big Data

Die Zukunft der Analyse usergenerierter Kommunikation ist zweifellos im Kontext von Big Data (Trübner/Mühlichen, Kapitel 10 in diesem Band) zu sehen. Dann werden diese Daten in Verbindung mit allen möglichen digitalen Fußspuren, wie z. B. Geodaten (Lakes, Kapitel 118 in diesem Band), Informationen aus Social Media-Kanälen, aber auch prinzipiell allen möglichen Daten aus Verkehr, Konsum, Kreditkarteninformationen, Mobilfunk, Verwaltungsakten und aus anderen Quellen sowie allen möglichen elektronischen Geräten, die mit ihrer Umgebung kommunizieren können, ganz neue Anwendungsmöglichkeiten zulassen – aber auch neue ethische und methodische sowie datenverarbeitungsbezogene Fragen aufwerfen (Mayerl 2015). Viele der genannten methodischen Probleme sind Herausforderungen für die Zukunft. Usergenerierte Kommunikationsdaten sind als Teil der Big Data-Ära zwar mit all deren Problemen verbunden, die häufig nur in interdisziplinären Forschungskontexten zu lösen sind. Gleichzeitig besitzen sie das große, zukunftsweisende Potenzial einer modernen empirischen Sozialforschung. Gerade die enge Zusammenarbeit zwischen Informatik bzw. Informationswissenschaften und Sozialwissenschaften ist

hier entscheidend, um die Kompetenzen aus Informationstechnologie und Big Data einerseits sowie sozialwissenschaftlicher Theorie und sozialwissenschaftlicher Datensensibilität andererseits zu verbinden.

Literatur

- Aramaki, Eiji/Maskawa, Sachiko/Morita, Mizuki (2011): Twitter Catches the Flu: Detecting Influenza Epidemics using Twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 27–31 July 2011. Edinburgh, UK. URL: <http://sociocom.jp/PAPER/2011/EMNLP.pdf>
- Beisch, Natalie/Koch, Wolfgang (2021): 25 Jahre ARD/ZDF-Onlinestudie: Unterwegsnutzung steigt wieder und Streaming/Mediatheken sind weiterhin Treiber des medialen Internets. *Media Perspektiven* 10: 486–503. URL: https://www.ard-media.de/fileadmin/user_upload/media-perspektiven/pdf/2021/2110_Beisch_Koch.pdf
- Diekmann, Andreas/Jann, Benn/Przepiorka, Wojtek/Wehrli, Stefan (2014): Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets. In: *American Sociological Review* 79 (1): 65–85
- Eichstaedt, Johannes C./Schwartz, Hansen A./Kern, Margaret L./Park, Gregory/Labarthe, Darwin R./Merchant, Raina M./Jha, Sneha/Agrawal, Megha/Dziurzynski, Lukasz A./Sap, Maarten/Weeg, Christopher/Larson, Emily E./Ungar, Lyle H./Seligman, Martin E. P. (2015): Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. In: *Psychological Science* 26 (2): 159–169
- Fownes, Jennifer R./Yu, Chao/Margolin, Drew B. (2018): Twitter and Climate Change. *Sociology Compass*. 2018;12:e12587. DOI: 10.1111/soc4.12587
- Gayo-Avello, Daniel (2013): A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter Data. In: *Social Science Computer Review* 31 (6): 649–679
- Gerber, Matthew (2014): Predicting Crime Using Twitter and Kernel Density Estimation. In: *Decision Support Systems* 61: 115–125
- Hegelich, Simon (2016): Social Bots. Invasion der Meinungsroboter. In: *Analysen & Argumente* 221. Konrad Adenauer Stiftung. URL: http://www.kas.de/wf/doc/kas_46486-544-1-30.pdf?161222122757/
- Hohlfeld, Ralf/Bauerfeind, Franziska/Braglia, Franziska et al. (2021): Communicating COVID-19 against the Backdrop of Conspiracy Ideologies: How Public Figures Discuss the Matter on Facebook and Telegram. Disinformation Research Lab, University of Passau, Working Paper 1/2021. DOI: 10.13140/RG.2.2.36822.78406
- Hsieh, Yuli P./Murphy, Joe (2017): Total Twitted Error. Decomposing Public Opinion Measurement on Twitter from a Total Survey Error Perspective. In: Biemer, Paul P./de Leeuw, Edith/Eckman, Stephanie/Edwards, Brad/Kreuter, Frauke/Lyberg, Lars E./Tucker, Clyde N./West, Brady T. (2017): *Total Survey Error in Practice*. Hoboken: John Wiley & Sons, 23–45
- Lewis, Seth C./Zamith, Rodrigo/Hermida, Alfred (2013): Content Analysis in an Era of Big Data. In: *Journal of Broadcasting & Electronic Media* 57 (1): 34–52

- Li, Jiawei/Xu, Qing/Shah, Neal/Mackey, Tim K. (2019): A Machine Learning Approach for the Detection and Characterization of Illicit Drug Dealers on Instagram: Model Evaluation Study. *Journal of Medical Internet Research* 21(6):e13803 doi: 10.2196/13803
- Machuca, Cristian R./Gallardo, Cristian/Toasa, Renato M. (2020): Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach. *Journal of Physics: Conference Series* 1828 (2021) 012104, doi:10.1088/1742-6596/1828/1/012104
- Mayerl, Jochen (2015): Bedeutet ‚Big Data‘ das Ende der sozialwissenschaftlichen Methodenforschung? In: *Soziopolis*. URL: <https://soziopolis.de/daten/kalender/blaetter/beobachten/wissenschaft/artikel/bedeutet-big-data-das-ende-der-sozialwissenschaftlichen-methodenforschung/>
- McKosker, Anthony/Gerrard, Ysabel (2021): Hashtagging Depression on Instagram: Towards a more Inclusive Mental Health Research Methodology. *New Media & Society* 2021, Vol. 23(7) 1899–1919
- Meeder, Brendan/Tam, Jennifer/Kelley, Patrick G./Cranor, Lorrie (2010): RT@ IWant-Privacy: Widespread Violation of Privacy Settings in the Twitter Social Network. URL: www.cs.cmu.edu/~jdtam/Documents/Meeder-SNSP2010.pdf
- Middleton, Stuart E./Middleton, Lee/Modafferi, Stefano (2014): Real-Time Crisis Mapping of Natural Disasters Using Social Media. In: *IEEE Intelligent Systems* 29(2): 9–17
- Mitchell, Lewis/Frank, Morgan R./Harris, Kameron D./Dodds, Peter S./Danforth, Christopher M. (2013): The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. In: *PLoS ONE* 8 (5). URL: <https://doi.org/10.1371/journal.pone.0064417>
- Munzert, Simon/Rubba, Christian/Meißner, Peter/Nyhuis, Dominic (2014): *Automated Data Collection with R. A practical Guide to Web Scraping and Text Mining*. West Sussex: John Wiley & Sons, 371–379
- Narayanan, Arvind/Shmatikov, Vitaly (2009): De-anonymizing Social Networks. In: *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*. URL: <http://randomwalker.info/social-networks/>
- Nyhuis, Dominic/Friederich, J. (2017): Begleitung des TV-Duells auf Twitter. In: Faas, Thorsten/Maier, Jürgen/Maier, Michaela (Hg.): *Merkel gegen Steinbrück*. Wiesbaden: Springer VS, 157–172
- Pfaffenberger, Fabien (2021): *Social Bots in der politischen Twittersphäre – Identifikation und Relevanz*. Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg. urn:nbn:de:bvb:29-opus4-169911
- Scherr, Sebastian/Arendt, Florian/Frissen, Thomas/Oramas M., José (2021): Detecting Intentional Self-Harm on Instagram: Development, Testing, and Validation of an Automatic Image-Recognition Algorithm to Discover Cutting-Related Posts. *Social Science Computer Review* 2020, Vol. 38(6) 673–685

- Siegel, Melanie/Alexa, Melpomeni (2020): Sentiment-Analyse deutschsprachiger Meinungsäußerungen. Grundlagen, Methoden und praktische Umsetzung. Wiesbaden: Springer
- Sloan, Luke/Morgan, Jeffrey/Housley, William/Williams, Matthew/Edwards, Adam/Burnap, Pete/Rana, Omer (2013): Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. In: Sociological Research Online 18 (3). URL: <http://www.socresonline.org.uk/18/3/7.html>

Thorsten Faas ist Professor für Politikwissenschaft im Bereich „Politische Soziologie der Bundesrepublik Deutschland“ an der Freien Universität Berlin. *Ausgewählte Publikationen:* The Impact of Online versus Offline Campaign Information on Citizens' Knowledge, Attitudes and Political Behaviour: Comparing the German Federal Elections of 2005 and 2009, in: German Politics 24 (zusammen mit Julia Partheymüller, 2014); The Campaign and its Dynamics at the 2005 German General Election, in: German Politics 15, 4 (zusammen mit Rüdiger Schmitt-Beck, 2006); To Defect or not to Defect? National, Institutional and Party Group Pressures on MEPs and their Consequences for Party Group Cohesion in the European Parliament, in: European Journal of Political Research 42, 6 (2003). *Webseite:* <http://www.thorstenfaas.de>. *Kontaktadresse:* Thorsten.Faas@fu-berlin.de.

Jochen Mayerl ist Professor für Soziologie mit Schwerpunkt Empirische Sozialforschung an der Technischen Universität Chemnitz. *Ausgewählte Publikationen:* Early days of the pandemic – The association of economic and socio-political country characteristics with the development of the COVID-19 death toll. PLoS ONE 16(8): e0256736, 2021 (mit Holz, M.); The dynamics of ethnocentrism in Europe. Comparing enduring and emerging determinants of solidarity towards immigrants. European Societies 21(5): 672–703, 2019 (mit Aschauer, W.); Responding to socially desirable and undesirable topics: Different types of response behaviour? methods, data, analyses 13(1): 7–35, 2019 (mit Andersen, H.). *Webseite:* https://www.tu-chemnitz.de/hsw/soziologie/Professuren/Empirische_Sozialforschung/. *Kontaktadresse:* Jochen.Mayerl@soziologie.tu-chemnitz.de.