SSOAR

Open Access Repository
www.ssoar.info

Garbage in - Garbage out? Datenqualität im Umgang mit digitalen Verhaltensdaten

Fröhling, Leon; Birkenmaier, Lukas; Daikeler, Jessica

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Fröhling, L., Birkenmaier, L., & Daikeler, J. (2023). Garbage in - Garbage out? Datenqualität im Umgang mit digitalen Verhaltensdaten. *easy_social_sciences*, 68, 21-30. https://doi.org/10.15464/easy.2023.03

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

doi: 10.15464/easy.2023.03

Garbage in – Garbage out?

Datenqualität im Umgang mit digitalen Verhaltensdaten

Leon Fröhling, Lukas Birkenmaier & Jessica Daikeler

Während in den quantitativen Sozialwissenschaften Umfragedaten seit jeher das Herzstück der Informationsgewinnung bilden, spielten Beobachtungsdaten und andere Datenquellen eine eher untergeordnete Rolle. Soziale Medien und mobile Endgeräte lassen nun digitale Verhaltensdaten immer mehr in den Mittelpunkt sozialwissenschaftlicher Forschung rücken. Doch selbst die innovativsten und umfangreichsten Datenmengen sind unzureichend, wenn sie nicht von hoher Qualität sind. Dieser Artikel diskutiert anhand eingängiger Beispiele die grundlegenden Herausforderungen bei der Analyse digitaler Verhaltensdaten und präsentiert einen zentralen Ansatz zur Evaluation ihrer Qualität.

While survey data has always been at the heart of information gathering in the quantitative social sciences, observational data and other data sources have played a rather subordinate role. Social media and mobile devices are now making new forms of digital behavioral data increasingly central to social science research. However, even the most innovative and comprehensive data sets are insufficient if they are not of high quality. This article discusses the basic challenges of analyzing digital behavioral data using several use cases. Ultimately, it presents one central framework to evaluate the applicability of digital behavioral data for social science research.

Keywords: Beobachtungsdaten, digitale Verhaltensdaten, Repräsentativität, Validität, Datenqualität

Digitale Verhaltensdaten für die Sozialwissenschaften

Eine Vielzahl wirtschaftlicher, sozialer und gesellschaftlicher Vorgänge baut mehr und mehr auf digitalen Technologien und Onlineplattformen auf. Von der Kommunikation mit Freund*innen und Verwandten, über die Nutzung von digitalen Medieninhalten, bis hin zum Kauf und Vertrieb von Produkten und Dienstleistungen, prägen digitale Formate und Werkzeuge zunehmend unseren Alltag und führen zu großen Mengen digitaler Verhaltensdaten. Digitale Verhaltensdaten sind digitale Spuren menschlichen Verhaltens, wie sie beispielsweise von Online-Plattformen, smarten Geräten und speziellen Forschungssensoren erfasst werden. In Verbindung mit neuen, computergestützten Auswertungsmethoden bieten diese digitalen Verhaltensdaten neuartige Potentiale zur Beschreibung sozialer und politischer Prozesse – von globaler Vernetzung über politische Polarisierung hin zur Beschreibung von Interaktionsmustern im digitalen Raum (King, 2011).

Der Mehrwert einer systematischen Auswertung dieser Daten ist enorm. In der Wirtschaft etwa passen Internetkonzerne wie Google, Facebook oder Amazon ihre Produktempfehlungen und Werbeanzeigen mit

digitalen Daten individuell an die einzelnen Kund*innen an, und entwickeln sich so mehr und mehr zu global dominierenden Konzernen.

In der sozialwissenschaftlichen Forschung finden digitale Verhaltensdaten ebenso regelmäßig Anwendung. An der Schnittstelle zwischen den Sozialwissenschaften (*Social Science*) und der Informatik (*Computer Science*) entsteht das neue Forschungsfeld der *Computational Social Science*, welches einerseits klassisch sozialwissenschaftliche Phänomene mit neuen Daten und Methoden untersucht, aber auch neue, durch die zunehmende Digitalisierung erst auftretende Phänomene erforscht.

Datenqualität – eine zentrale Herausforderung

Obwohl digitale Verhaltensdaten für Forschende eine neue und vielversprechende Datenquelle darstellen, ist die Erhebung und Auswertung dieser Daten in der Praxis oft mit großen Herausforderungen verbunden. Bereits kleine Entscheidungen in der Sammlung, Verarbeitung und Auswertung digitaler Verhaltensdaten können einen gewichtigen Einfluss auf die Qualität von Daten und damit auch die Ergebnisse ganzer Studien haben. Ein populäres Beispiel hierfür ist der „Google Flu Trends"-Algorithmus (GFT), welcher erstmals 2008 von Google veröffentlicht wurde (Ginsberg et al., 2009). Die Idee hinter der simplen Anwendung ist, auf Basis von Google-Suchanfragen zu Grippesymptomen in Echtzeit zu erkennen, wie sich die Grippe an verschiedenen Orten der Welt ausbreitet. Obwohl Grippewellen so anfangs noch mit beeindruckender Präzision vorhergesagt werden konnten, verschlechterte sich in den Jahren darauf die Vorhersagequalität teils erheblich, was schließlich zur Einstellung des Projektes im Jahr 2015 führte. Abbildung 1 zeigt den Verlauf der Grippe zusammen mit den GFT-Prognosen für die Jahre 2009 bis 2013. Während GFT in den Anfangsjahren noch eine hohe Übereinstimmung mit den Daten der nationalen Gesundheitsbehörde in den USA (CDC) hatte, nahm die Vorhersagegenauigkeit über die Jahre zunehmend ab. Das ging so weit, dass für 2013 der Anteil der vorhergesagten Grippeerkrankungen mehr als doppelt so hoch wie der Anteil der tatsächlich gemeldeten Erkrankungen war. Grund für die fehlerhaften Vorhersagen waren zum einen kleine Änderungen des Google-Suchalgorithmus sowie des Verhaltens der Nutzenden, welche einen direkten Effekt auf die Vorhersagequalität hatten (Lazer et al., 2014). Zum anderen konnten Lazer et al. (2014) zeigen, dass sich die Symptome jahreszeittypischer Erkrankungen wie Erkältungen und die der Grippe, und damit auch die jeweiligen Suchanfragen, zu stark ähneln. Da das GFT-Modell diese aber

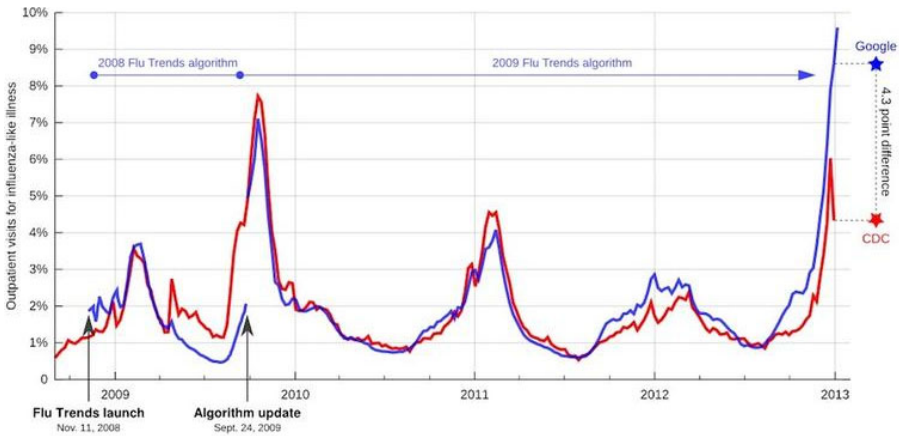

Abb 1 Vorhersage von Grippewellen mithilfe digitaler Verhaltensdaten. In Blau der Anteil der von Google Flu Trend vorhergesagten Grippeerkrankungen und in Rot der Anteil der tatsächlich registrierten Grippeerkrankungen in der Gesamtbevölkerung, über den Zeitverlauf von 2009 bis 2013. Spätestens ab Mitte 2012 ist eine deutliche Abweichung der Vorhersage von den tatsächlich gemeldeten Daten zu erkennen.
Quelle: https://www.wbur.org/news/2013/01/13/google-flu-trends-cdc (Zuletzt abgerufen am 16.12.2022).

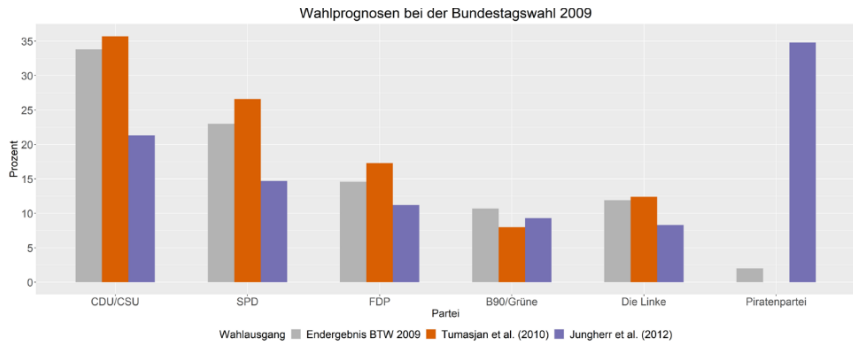

Abb. 2 Vergleich der tatsächlichen Stimmanteile ausgewählter Parteien bei der Bundestagswahl 2009 mit den Vorhersagen von Tumasjan et al. (2010) und Jungherr et al. (2012).

Abbildung 2 vergleicht die Ergebnisse der Bundestagswahl 2009 mit den Vorhersagen von Tumasjan et al. (2010) und Jungherr et al. (2012). Obwohl die nur von Jungherr et al. (2012) berücksichtigte Piratenpartei am Ende nur 2 % der Wähler*innenstimmen bekam, hätte sie nach dem Prognosemodell mit knapp 35 % aller Stimmen den Wahlsieg holen müssen. Diese hohe Diskrepanz erregt Zweifel, ob die zugrundliegende Methodik in der Lage ist, verlässliche und belastbare Wahlprognosen zu generieren.

nicht unterscheiden konnte, funktionierte es als Vorhersagemodell jahreszeittypischer Erkrankungen (“winter detector”), aber nicht spezifisch für die Grippe.

Der Versuch, Wahlergebnisse durch Social Media Daten vorherzusagen, dient als weiteres Beispiel für die Herausforderungen in der Arbeit mit digitalen Verhaltensdaten. Ein solches Modell, das einzig anhand der Häufigkeit der Erwähnungen von politischen Parteien auf der Plattform Twitter Stimmanteile bei politischen Wahlen prognostizieren soll, stellen Tumasjan et al. (2010) vor. Ihr Modell, so die Autor*innen, hätte den Ausgang der Bundestagswahl 2009 mit nur minimaler Abweichung vorhergesagt. In einer Replik zeigen Jungherr et al. (2012) aber, dass die Plausibilität der Ergebnisse erheblich von den ohne weitere Begründungen getroffenen Entscheidungen der Forschenden abhängt. Eine dieser Entscheidungen stellt der Verzicht auf Parteien, die sich 2009 nicht im Bundestag befanden, dar. Insbesondere die Piratenpartei, die 2009 als aufstrebende politische Kraft stark im digitalen Raum präsent war, wurde ohne Begründung von Tumasjan et al. (2010) nicht berücksichtigt. Ebenfalls unbegründet und in ihren Auswirkungen unreflektiert blieb die Entscheidung, die Datenerhebung acht Tage vor der Bundestagswahl zu beenden. Jungherr et al. (2012) zeigen, dass die Einbeziehung der acht nicht-berücksichtigten Tage bis zur Bundestagswahl zu einer erheblich schlechteren Vorhersagequalität geführt hätte.

Neben den unten noch näher erläuterten Problemen in der Fallstudie existieren eine Vielzahl weiterer Herausforderungen, die das Arbeiten mit digitalen Verhaltensdaten erschweren. So ist bekannt, dass die Nutzenden von Social-Media-Plattformen wie Twitter oftmals jünger, eher männlich und gebildeter als der Durchschnitt der deutschen Bevölkerung sind (Blank, 2017; Sloan, 2017). Außerdem ist unklar, welcher Anteil der ausgewerteten Tweets von automatisierten Accounts (*Bots*) oder von anderweitig nicht-wahlberechtigten Accounts (z.B. Unternehmen oder Medienanstalten) verfasst wurde. Dies verdeutlicht, dass sich auf Basis von Twitter-Daten nur schwer Aussagen über das allgemeine Wahlverhalten machen lassen.

> ***Es fehlt, im Gegensatz zur Umfrageforschung, an allgemein gültigen Verfahren zur Einschätzung der Datenqualität.***

Ebenfalls zeigen die Beispiele, dass Forschende in der Arbeit mit digitalen Verhaltensdaten eine Vielzahl an Entscheidungen treffen müssen, welche die Ergebnisse ihrer Forschung maßgeblich beeinflussen. Diese reichen von der Sammlung und Aufbereitung der Daten, über die Auswahl und Spezifizierung

der Auswertungsmethode, bis hin zur kritischen Auseinandersetzung und der Interpretation der eigenen Ergebnisse. Insbesondere weil digitale Verhaltensdaten in erster Linie nicht für Forschungszwecke erstellt, sondern lediglich für diese nachgenutzt werden, fehlen in vielen Fällen oft wichtige Hintergrundinformationen zu ihrer Entstehung und ihrem Kontext. Zudem macht die Größe vieler Datensätze mit häufig Millionen von Datenpunkten eine manuelle Überprüfung und Verarbeitung der Daten meist praktisch unmöglich. Während Forschende bei einer kleinen Stichprobe die Möglichkeit haben, beispielsweise die gesammelten Tweets selbst zu überprüfen, kann dies bei großen Datensätzen nur automatisiert durchgeführt werden. Schließlich fehlt es, im Gegensatz zur Umfrageforschung, an allgemein gültigen Verfahren zur Einschätzung der Datenqualität. Es haben sich noch keine Standards etabliert, um alle Faktoren, die für die Qualität der Daten, die Interpretation der Ergebnisse und deren Reproduzierbarkeit durch andere Forschende, einheitlich zu dokumentieren und zu bewerten. Dies ist besonders problematisch, da wissenschaftliche Forschung solche Standards benötigt, damit Forschende die Zuverlässigkeit und Transparenz ihrer Forschung sicherstellen können.

Error Frameworks zur Qualitätsprüfung

Zur Lösung der umrissenen Probleme können sogenannte *Error Frameworks* hilfreich sein. Die Idee der *Error Frameworks* stammt aus den Sozialwissenschaften, insbesondere der Umfrageforschung, wo diese seit den 2000er Jahren dazu verwendet werden, den Forschungsprozess systematisch auf potenzielle Fehlerquellen zu durchleuchten (Groves & Lyberg, 2010). Ziel eines *Error Frameworks* in der Umfrageforschung ist es, für jede Phase im Forschungsprozess, von der Stichprobenziehung bis zur Datenerhebung und -auswertung, mögliche Fehler (*Errors*) und deren Quellen zu identifizieren. Der Begriff Fehler wird hier nicht in den Dimensionen „richtig“ und „falsch“ verwendet, sondern ist als *Verzerrung* zu verstehen. Eine Verzerrung ist eine systematische Beeinflussung der Daten, die von den Forschenden weder beabsichtigt noch kontrolliert ist, und sich damit potenziell in einem verzerrten Ergebnis der Analyse niederschlägt. *Error Frameworks* ermöglichen, die Fehler über den gesamten Forschungsprozess zu identifizieren und zu aggregieren. Dadurch können Aussagen über die Qualität der Forschung sowie über die Aussagekraft ihrer Ergebnisse gemacht werden.

In den letzten Jahren wurden *Error Frameworks* speziell für die Anforderungen der Forschung mit digitalen Verhaltensdaten angepasst (Amaya et al., 2020; Hsieh and Murphy, 2017; Sen et al., 2021). Abbildung 3 gibt einen Überblick über das *Total Error Framework for Digital Traces of Human Behavior on Online Platforms* (*TED-On*) von Sen et al. (2021). Dieses ist explizit für die Arbeit mit digitalen Verhaltensdaten von Online-Plattformen entwickelt worden. Das TED-On Framework orientiert sich an fünf Phasen des Forschungsprozesses mit digitalen Verhaltensdaten (Definition des Konstrukts, Auswahl der Plattform, Sammlung, Aufbereitung und Analyse der Daten). Es unterscheidet zwischen zwei verschiedenen Kategorien von Fehlern: Mess- und Repräsentationsfehlern. Messfehler beziehen sich darauf, wie das Konstrukt (z.B. die Präferenz für eine politische Partei) aus den Daten gemessen wird (z.B. durch die Anzahl ihrer Erwähnungen auf Twitter), während Repräsentationsfehler sich auf Fehler in der Erfassung der für die Studie relevanten Personengruppe beziehen (ob etwa die Zusammensetzung der analysierten Twitter-Nutzenden die gesamte Wahlbevölkerung gut repräsentiert). In den folgenden Absätzen wird nun anhand eines Fallbeispiels das TED-On mitsamt seiner verschiedenen Fehlerkategorien angewandt. Es wird aufgezeigt, wie die Verwendung des *Error Frameworks* Forschende bei der kritischen Reflektion ihrer Arbeit mit

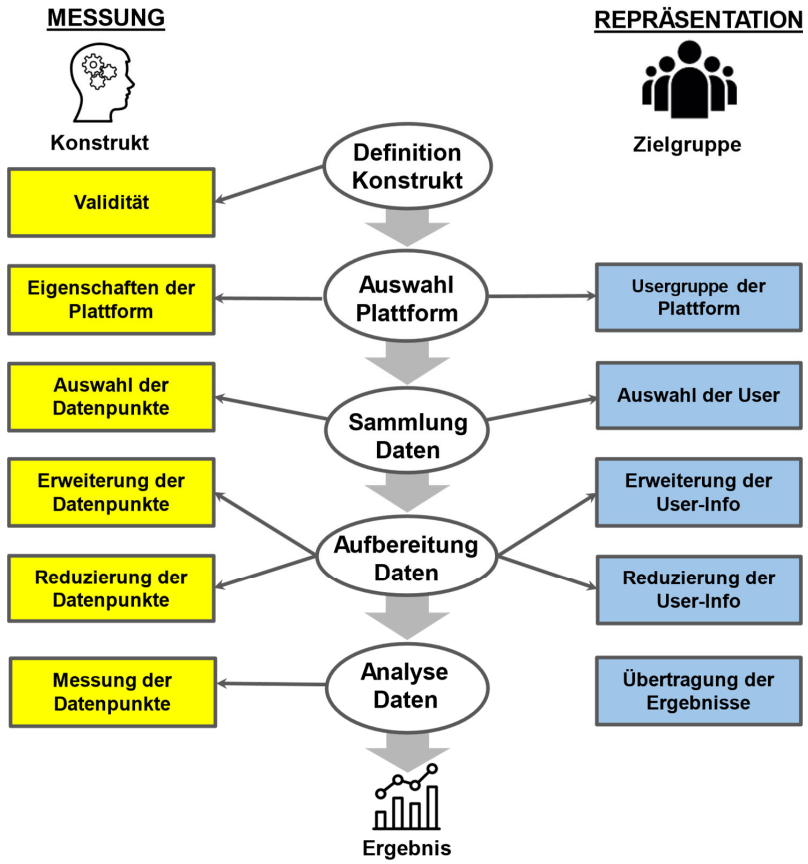

Abb.3 TED-On Framework (übersetzte Abbildung aus Sen et al., 2021).

digitalen Verhaltensdaten unterstützen kann. Als Fallbeispiel wird die in Abschnitt 2 bereits eingeführte Studie von Tumasjan et al. (2010) zur Vorhersage des Bundestagswahlergebnisses anhand von Twitter-Daten betrachtet.

Anwendung des TED-On Frameworks auf ein Fallbeispiel

Definition und Operationalisierung des Konstrukts

Als ersten Schritt im Forschungsprozess mit digitalen Verhaltensdaten benennt das TED-On Framework die Definition des zu untersuchenden Konstrukts. Ein Konstrukt ist das Phänomen, welches in der Studie gemessen und untersucht werden soll. Da das zu untersuchende Konstrukt zumeist nicht direkt beobachtbar ist (beispielhaft etwa die Konstrukte politische Beeinflussung oder Ideologie), greifen Forschende auf sogenannte Operationalisierungen zurück. Eine Operationalisierung stellt dabei die Verbindung zwischen dem Konstrukt und den Daten her. Sie beschreibt also die Methode, wie das Konstrukt aus den Daten zu messen ist. Dieser Zusammenhang zwischen Konstrukt und Daten ist im besten Fall theoretisch herleitbar und empirisch (z.B. durch andere Studien) belegbar. Erfolgt die Definition des Konstrukts gar nicht oder nicht ausreichend präzise, oder wird durch die gewählte Operationalisierung das Konstrukt nicht genau gemessen, entstehen Probleme mit der Validität des Forschungsdesigns. Validität kann hier verstanden werden als „Angemessenheit", ob und wie weit also das Forschungsdesign geeignet ist, die Forschungsfrage zu beantworten.

Im Fallbeispiel ist das Konstrukt schnell gefunden: Es sollen die Stimmanteile der Parteien in der Bundestagswahl 2009 anhand von Twitter-Daten vorhergesagt werden. Die verwendete Operationalisierung ist unmittelbar nachvollziehbar, so wird der Anteil der Erwähnungen der betrachteten Parteien in den Tweets, verfasst in einem bestimmten Zeitraum, verwendet, um den Stimmenanteil bei der Wahl zu prognostizieren. Der in der Operationalisierung unterstellte Zusammenhang und damit die Validität der Studie kann zumindest angezweifelt werden. Wenn etwa Tweets, gerichtet an den Account einer bestimmten Partei, ausschließlich Ausdruck negativer Gefühle gegenüber der Partei sind, etwa im Kontext eines *Shitstorms*, dann ist nicht unmittelbar ersichtlich, wieso sich dies ausgerechnet in einem besseren Wahlergebnis der

betroffenen Partei widerspiegeln sollte. Die in der Studie gewählte Form der Operationalisierung würde diesen Wirkzusammenhang aber direkt implizieren.

Auswahl einer Plattform zur Sammlung der Daten

Nachdem das Konstrukt definiert und die Messung dieses Konstrukts aus den noch zu sammelnden Daten festgelegt ist, sieht der zweite Schritt im Forschungsprozess die Wahl einer passenden Plattform als Datenquelle vor. Der Begriff Plattform meint an dieser Stelle hauptsächlich Soziale Medien als Orte sozialer Online-Interaktionen. Bei der Auswahl ist zu beachten, dass Eigenschaften der Plattform häufig einen direkten Einfluss auf die verfügbaren Daten haben. Es ist daher wichtig zu unterscheiden, welches in den Daten zu beobachtende Verhalten unabhängig von der Plattform Rückschlüsse auf die Person hinter dem Verhalten zulässt, und welches Verhalten von den Eigenschaften der Plattform bestimmt wird. Ferner haben verschiedene Plattformen unterschiedliche User*innengruppen, wodurch bereits die Wahl der Plattform vorgibt, welche demographischen Gruppen in der Studie überhaupt betrachtet werden können.

In dem Fallbeispiel wird der Kurznachrichtendienst Twitter als Plattform gewählt. Neben der bislang guten Zugänglichkeit der Daten über die von Twitter bereitgestellte API for Academic Research spricht für Twitter auch der Ruf, eine besonders *politische* Kommunikationsplattform zu sein. Dort tauschen sich vornehmlich Akteur*innen aus dem Journalismus, den Medien, der Wissenschaft und eben der Politik aus, häufig auch über (tages-)politische Entwicklungen. Während zumindest der inhaltliche Diskurs auf der Plattform thematisch passend erscheint, bleibt die Frage, ob die User*innengruppe repräsentativ für den wahlberechtigten Teil der deutschen Gesellschaft ist, der für das gewählte Konstrukt (Stimmanteil bei der Bundestagswahl 2009) maßgeblich sein sollte, was für Twitter-Nutzende nicht zutrifft. Darüber hinaus reflektieren auch schon die Autor*innen des Fallbeispiels, inwiefern sich das auf Twitter geltende Zeichenlimit (zur Zeit der Studie durfte ein Tweet nicht mehr als 140 Zeichen umfassen, aktuell liegt das zulässige Zeichenlimit bei 280) auf den Informationsgehalt der Tweets auswirkt – ein klassisches Beispiel für beobachtetes Verhalten, das direkt durch die Eigenschaften der Plattform beeinflusst wird.

Sammlung der Daten

Als nächstes ist im Forschungsprozess die Festlegung einer geeigneten Methode zum Sammeln der benötigten Daten vorgesehen. Mögliche Probleme und Verzerrungen sind hier in zwei gegensätzliche Richtungen denkbar. Zum einen kann es sein, dass Daten gesammelt und für die Analyse berücksichtigt werden, die für die Messung des Konstrukts und damit die gesamte Studie nicht relevant sind. Zum anderen kann es passieren, dass durch die gewählte Methodik der Datensammlung nicht alle relevanten Beobachtungen erfasst und in den Datensatz aufgenommen werden. Sowohl im Falle der zu liberal als auch im Falle der zu restriktiv gewählten Kriterien für die Inklusion von Daten führt dies zu verzerrten Ergebnissen.

Eindrucksvoll aufgezeigt wird diese Problematik wie oben bereits eingeführt in der Replikationsstudie von Jungherr et al. (2012) zu unserem Fallbeispiel. Nach Jungherr et al. (2012) verzerren die Auswahl der für die Datensammlung berücksichtigten Parteien und der festgelegte zeitliche Rahmen, für den Daten gesammelt wurden, das Ergebnis der Studie. Durch die Nichtberücksichtigung der bei Twitter-Nutzenden damals populären Piratenpartei bei Tumasjan et al. (2010) für die Prognose der Stimmanteile verschieben diese sich deutlich zu Gunsten der übrigen Parteien und nähern sich somit dem tatsächlichen Wahlausgang an (vgl. Abbildung 2). Jungherr et al. (2012) zeigen also, wie eine andere, ebenfalls plausible Entscheidung in der Auswahl der berücksichtigten Parteien und in der Festlegung des Zeitraums der Datensammlung für die Qualität und Aussa-

gekraft der Daten entscheidend sind. Jungherr et al. (2012) verdeutlichen, wie die erwarteten Stimmanteile einzelner Parteien um bis zu 6,5 Prozentpunkte schwanken, wenn der für die Auswertung betrachtete Zeitraum an Twitter-Aktivität um wenige Tage verschoben wird.

Aufbereitung der Daten

In einer typischen Studie mit digitalen Verhaltensdaten schließt sich an den Schritt der Datensammlung die Aufbereitung der Daten für die abschließende Analyse an. Zweck dieses Schrittes ist es, die gesammelten „rohen" Daten in das für die Auswertung benötigte Format zu bringen. Die Art der Aufbereitung hängt dabei sowohl von den Eigenschaften der gesammelten Daten als auch von den Anforderungen der Analysemethode ab. Bei erweiternden Methoden der Aufbereitung werden die gesammelten rohen Daten mit zusätzlichen Informationen versehen, und bei reduzierenden Methoden werden einzelne Datenpunkte oder Informationen entfernt. Wenn es bei der Erweiterung oder der Reduzierung zu systematischen Fehlern kommt, werden dadurch auch die Daten und damit potenziell auch die Ergebnisse systematisch verzerrt.

Im Fallbeispiel kommen weder erweiternde noch reduzierende Aufbereitungsschritte zum Einsatz, da die Messung des Konstrukts direkt auf den gesammelten rohen Daten aufbaut. Häufig werden bei Twitter-Studien Tweets, die als Spam eingeordnet werden, aus den gesammelten Daten entfernt. Dies ist in vielen Fällen sinnvoll, da diese Tweets keine Auseinandersetzung mit dem jeweiligen Thema beinhalten, sondern ausschließlich Aufmerksamkeit erregen sollen. Bei der Entscheidung, ob es sich bei einem Tweet um Spam handelt oder nicht, besteht die Gefahr, dass Tweets fälschlicherweise als Spam klassifiziert („Erweiterung" der Daten um die Information, ob es sich um Spam handelt oder nicht) und aus den Daten entfernt werden („Reduzierung" der Daten durch das Herausfiltern bestimmter Beobachtungen). Wenn dies systematisch geschieht, etwa weil die Tweets bestimmter User*innengruppen ähnliche Eigenschaften wie Spam-Tweets aufweisen, führt dies zu einer systematischen Nichtberücksichtigung dieser User*innengruppen für die Auswertung. Im Fallbeispiel wäre das gleichzusetzen damit, dass die Stimmen bestimmter Nutzender für die Berechnung der Wahlergebnisse nicht gezählt werden.

Analyse der Daten

Abschließender Schritt des idealtypischen Forschungsprozesses ist die tatsächliche Messung des Konstrukts aus den gesammelten und aufbereiteten Daten. Zu treffende Entscheidungen für die statistische Analyse sind etwa, auf welcher Ebene die Daten aggregiert werden, und wie aus den verschiedenen Aggregationsebenen das finale Resultat berechnet wird.

Im Fallbeispiel ist das zu messende Konstrukt die Stimmanteile der Parteien in der Bundestagswahl 2009. Tumasjan et al. (2010) berechnen die Stimmanteile für die berücksichtigten Parteien, in dem sie die Häufigkeit ihrer Erwähnungen in Tweets durch die Gesamtzahl der gesammelten Tweets teilen. In einer alternativ denkbaren Form der Aggregierung könnten Tweets auf der Ebene der einzelnen Tweet-Autor*innen aggregiert und in Stimmen für eine Partei ausgewertet werden. So würden etwa Nutzende, die mit mehreren Tweets im Datensatz vertreten sind, nicht mehrfach für die Berechnung der Stimmanteile berücksichtigt.

Während viele der Entscheidungen, die Tumasjan et al. (2010) in ihrem Forschungsdesign getroffen haben, von Jungherr et al. (2012) kritisch hinterfragt und hinsichtlich ihrer Auswirkungen auf die Resultate der Studie untersucht wurden (vgl. Übersicht der hier erläuterten Fehlerpotentiale in Tabelle 1) ist dies vermutlich nicht die größte Schwäche dieser Studie. Kritischer einzuordnen ist vielmehr die fehlende Dokumentation und Erläuterung der Prozesse der Entscheidungsfindung, da so Entscheidungen beliebig erscheinen und die Datenqualität nicht unmittelbar ersichtlich

Phase des Forschungsprozesses	Fehler-Kategorie	Identifiziertes Fehlerpotential
Definition Konstrukt	Validität	Passt die Operationalisierung (Nennung der Partei) zuverlässig zum Konstrukt (Stimmanteil)?
Auswahl Plattform	Eigenschaften der Plattform	Können kurze Tweets den nötigen politischen Informationsgehalt haben?
Auswahl Plattform	User*innengruppe der Plattform	Sind Twitter-User*innen repräsentativ für die stimmberechtigte Bevölkerung?
Sammlung Daten	Auswahl der Datenpunkte	Sind alle relevanten Parteien in den Daten berücksichtigt? Welcher Zeitraum wird für die Auswertung berücksichtigt?
Analyse Daten	Messung der Datenpunkte	Wie werden Tweets mit verschiedenen genannten Parteien gezählt? Wie wird aggregiert?

Tabelle 1 Übersicht über die mittels des TED-On identifizierten Fehlerpotentiale im Fallbeispiel der Bundestagswahlstudie von Tumasjan et al. (2010).

wird. Für die Arbeit mit digitalen Verhaltensdaten ist dies im Sinne der Nachvollziehbarkeit und Reproduzierbarkeit der Ergebnisse jedoch unbedingt wünschenswert.

> ***Menge und Verfügbarkeit digitaler Verhaltensdaten – eine Zeitenwende.***

Wie das Fallbeispiel aufzeigt, kann das TED-On dazu verwendet werden, den Prozess der kritischen Auseinandersetzung mit den eigenen Forschungsdesigns von Anfang an anzuleiten, und Forschende bei der Findung und Begründung ihrer Entscheidungen zu unterstützen.

To do: Qualitätsstandards und interdisziplinärer Austausch zu Datenqualität weiterentwickeln

Die Sozialwissenschaften erleben durch die immense Menge und Verfügbarkeit digitaler Verhaltensdaten gerade eine Zeitenwende. Dieser Beitrag versucht jedoch aufzuzeigen, dass selbst die innovativsten und umfangreichsten Daten unzureichend sind, wenn sie nicht von hoher Qualität sind. Diese Qualitätsdefizite können einerseits in den datengenerierenden Prozessen ihren Ursprung haben, aber auch im Umgang mit den Daten durch die Forschenden selbst. Digitale Verhaltensdaten werden immer häufiger genutzt. Gerade diese vermehrte Nutzung untermauert die Notwendigkeit zur Etablierung von Qualitätsstandards, um bestmögliche wissenschaftliche Rückschlüsse zu generieren, welche wiederum die Grundlage politischer Entscheidungen darstellen können. Die bisherigen Qualitätskonzepte (siehe für eine Übersicht zu Qualitätskonzepten Daikeler et al., 2022) und Dokumentationsstandards zu *TED-On* (z.B. TES-D von Fröhling et al., 2022), können nach den ersten Jahren des Ausprobierens insbesondere als erste Schritte gewertet werden, um Qualitätsstandards zu etablieren. Diese Standards sind, wie in unserem Beispiel erläutert, oftmals von dem Format und der Nutzung der Daten abhängig.

Die Ergänzung von Umfragedaten mit digitalen Verhaltensdaten hat neue Datenqualitätspotentiale, jedoch auch neue Herausforderungen, mit sich gebracht (Weiss & Stier, 2023, in diesem easy-Band). Datenqualitätsstandards hängen neu etablierten Datenformaten oftmals hinterher. Denken wir diese Entwicklung neuer Datenformate nun noch einen Schritt weiter, beispielsweise in die

Abb. 4 Datenerfassung von Bild-, Sprach-, Video und Sensordaten in einer virtuellen Realität, Photo by Eugene Capon, CC0 Public Domain

Richtung von *virtual* und *augmented reality*, oder auch „nur" in die automatisierte Analyse von Bild- und Videodaten, stehen wir mittel- und langfristig vor weiteren, nie dagewesenen Forschungsmöglichkeiten (z.B. zur Untersuchung von Vertrauen wie in Miller et al., 2019) – aber auch vor neuen Herausforderungen für die Datenqualität. Insbesondere durch die Verschmelzung von Text-, Video-, Sprach-, Bild- und Sensordaten (siehe Abbildung 4 für ein Beispiel) stehen die Sozialwissenschaften einer riesigen Datenquelle mit all ihrer potenziellen Datenverzerrungsproblematik gegenüber. Gerade vor diesem Hintergrund ist der Blick in die Qualitätsstandards von lang etablierten Disziplinen der Computer Science, der Umfrageforschung und der Sensortechnik essenziell und eine intensive Zusammenarbeit unerlässlich.

Zudem bleibt zu bedenken, dass Forschende oft noch keinen Zugang zu den neu entstehenden Datenmengen haben. Das liegt insbesondere an Datensicherheitsbedenken und kommerziellen Interessen vieler Plattformen. Aktuell ermöglichen insbesondere YouTube, Reddit und Mastodon für Forschende kostenlosen und weitreichenden Zugang zu veröffentlichten Daten, während bei vielen anderen Plattformen (z.B. Facebook, Instagram, TikTok und neuerdings Twitter) der Zugang zu den Daten stark eingeschränkt wird (Bruns, 2019; Freelon, 2018). Der richtige Umgang mit dieser Problematik ist eine weitere kurz- und mittelfristige Herausforderung für die sozialwissenschaftliche Forschung.

Zuletzt wird auch das hier diskutierte Problem der Repräsentation der Stichprobe auf die Gesamtbevölkerung erhalten bleiben. So wird es trotz hoher Nutzendenzahlen in der nächsten Dekade vermutlich nicht möglich sein, Rückschlüsse für die gesamte deutsche Bevölkerung über soziale Medien zu erheben.

Literatur

Amaya, A., Biemer, P. P. & Kinyon, D. (2020). Total error in a big data world: adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology, 8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society, 22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Freelon, D. (2018). Computational research in the post-API age. *Political Communication, 35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Groves, R. M. & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hsieh, Y. P. & Murphy, J. (2017). Total twitter error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, C. Tucker & B. T. West (Hg.), *Total survey error in practice* (S. 23–46). https://doi.org/10.1002/9781119041702.ch2

Jungherr, A., Jürgens, P. & Schoen, H. (2012). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment". *Social science computer review, 30*(2), 229–234. https://doi.org/10.1177/0894439311404119

Sen, I., Flöck, F., Weller, K., Weiß, B. & Wagner, C. (2021). A total error framework for digital traces

of human behavior on online platforms. *Public Opinion Quarterly, 85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Daikeler, J., Sen, I., Birkenmaier, L., Froehling, L., Gummer, T., Silber, H., Lechner, C. & Weiß, B. (2022). *Assessing Data Quality in the Age of Digital Social Research: A Systematic Review. Daikeler.* ESA RN 21 Quantitative Methods Mid-Term Conference, Salamanca, Spain.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Kim, M. & Lu, Y. (2020). Testing partisan selective exposure in a multidimensional choice context: Evidence from a conjoint experiment. *Mass Communication and Society, 23*(1), 107–127. https://doi.org/10.1080/15205436.2019.1636283

King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science, 331*(6018), 719–721. https://doi.org/10.1126/science.1197872

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science, 343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Moeller, J. & Helberger, N. (2018). *Beyond the filter bubble: Concepts, myths, evidence and issues for future debates.* https://hdl.handle.net/11245.1/478edb9e-8296-4a84-9631-c7360d593610

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you.* Penguin Press.

Ross Arguedas, A., Robertson, C., Fletcher, R. & Nielsen, R. (2022). *Echo chambers, filter bubbles, and polarisation: A literature review.* Reuters Institute for the Study of Journalism.

Scharkow, M., Mangold, F., Stier, S. & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences, 117*(6), 2761–2763. https://doi.org/10.1073/pnas.1918279117

Tumasjan, A., Sprenger, T., Sandner, P. & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Web and Social Media, 4*(1), 178–185. https://doi.org/10.1609/icwsm.v4i1.14009

Weiß, J. & Stier, S. (2023). Die Verknüpfung von digitalen Verhaltensdaten und Umfragedaten. *easy_social_sciences 68,* 31-38. https://doi.org/10.15464/easy.2023.04

Leon Fröhling

GESIS – Leibniz-Institut für Sozialwissenschaften

E-Mail leon.froehling@gesis.org

Leon Fröhling ist wissenschaftlicher Mitarbeiter und Doktorand in der GESIS-Abteilung Computational Social Sciences. Er forscht zu digitalen Verhaltensdaten, Datenqualität und dem Zusammenspiel von Technologie und Gesellschaft.

Lukas Birkenmaier

GESIS – Leibniz-Institut für Sozialwissenschaften

E-Mail lukas.birkenmaier@gesis.org

Lukas Birkenmaier ist wissenschaftlicher Mitarbeiter und Doktorand in der GESIS-Abteilung Survey Design and Methodology. Seine Forschungsschwerpunkte liegen im Bereich digitale Verhaltensdaten, Polarisierung und Validität.

Jessica Daikeler

GESIS Leibniz-Institut für Sozialwissenschaften

E-Mail jessica.daikeler@gesis.org

Dr. Jessica Daikeler arbeitet als Post-Doc in der GESIS-Abteilung Survey Design and Methodology. Sie forscht zu Repräsentativität von digitalen Verhaltensdaten und Umfragedaten, Bereitschaft zur Teilung von Daten, Modus- und Deviceeffekte in Befagungen.