Digitale Transformationsprozesse in der Forschung – Perspektiven für Methoden in der Technikfolgenabschätzung

Linda Nierling und Leonie Seng (Hrsg.)

KIT Scientific Publishing

Linda Nierling und Leonie Seng (Hrsg.)

Digitale Transformationsprozesse in der Forschung – Perspektiven für Methoden in der Technikfolgenabschätzung

Digitale Transformationsprozesse in der Forschung – Perspektiven für Methoden in der Technikfolgenabschätzung

Herausgegeben von
Linda Nierling und Leonie Seng

KIT Scientific Publishing

Institut für Technikfolgenabschätzung und Systemanalyse

Bild Umschlag: Ralf H. Schneider und Leonie Seng unter Verwendung der genAI Midjourney.

Impressum

Scientific Publishing

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark
of Karlsruhe Institute of Technology.
Reprint using the book cover is not allowed.

www.bibliothek.kit.edu/ksp.php | E-Mail: info@ksp.kit.edu | Shop: www.ksp.kit.edu

Print on Demand 2025 – Gedruckt auf FSC-zertifiziertem Papier

ISBN 978-3-7315-1391-9
DOI 10.5445/KSP/1000175516

Inhaltsverzeichnis

Einführung

Digitale Transformationsprozesse sind in vielen Lebensbereichen allgegenwärtig. Gesellschaftliche Veränderungsprozesse und digitale Technologien bedingen sich dabei gegenseitig (vgl. z. B. Grunwald 2022). Die Wechselwirkungen zwischen technischen Entwicklungen und gesellschaftlichen Prozessen ist traditionell Gegenstand der Technikfolgenabschätzung (TA). Schrape (2021, S. 84 ff.) beschreibt folgende „Phasen der digitalen Transformation“: die „Emergenz der ‚Informationsgesellschaft‘ als Begriff und Idee“ (1960er/ 70er-Jahre), die „beginnende Informatisierung der alltäglichen Lebenswelt (1980er/ 90er-Jahre), den „Aufstieg der Datenunternehmen und ‚Web 2.0‘-Diskurs“ (2000er-Jahre) sowie die „soziale Vergegenwärtigung der Digitalisierung“ (ab 2010).

In dieser Entwicklung hat Forschung nicht nur den Stellenwert einer Akteurin, sondern kann auch zum Gegenstand der digitalen Transformation selbst werden.[1] So werden zum einen digital-gestützte wissenschaftliche Methoden immer relevanter für Forschungsprojekte, zum anderen werden diese auch durch die rasante Entwicklung beispielsweise im Feld der generativen künstlichen Intelligenz (KI) beeinflusst, also der Automatisierung regelbasierten Wissens, wobei technische Systeme häufig die Regeln selbst als Teil einer Problemlösungsstrategie entwickeln (vgl. Heil 2021, S. 424). KI-Tools – obschon kontrovers diskutiert (vgl. Deutscher Ethikrat 2023) – werden mitunter als hilfreich für die wissenschaftliche Arbeit eingeordnet (Albrecht 2024; Jahnel und Heil 2024); angefangen von der Organisation, über die Textproduktion bis hin zur Sicherung wissenschaftlicher Ergebnisse.

[1] Die digitale Transformation von Forschung ist Gegenstand des Leibniz WissenschaftsCampus „Digital Transformation of Research – DiTraRe“, in dem die Auswirkungen und Potenziale der zunehmenden Digitalisierung des wissenschaftlichen Arbeitens untersucht werden. Die vorliegende Studie leistet hierin einen Beitrag aus dem Feld der Technikfolgenabschätzung. Der WissenschaftsCampus ist ein gemeinsames Projekt von FIZ Karlsruhe und dem Karlsruher Institut für Technologie, gefördert durch die Leibniz-Gemeinschaft von 2023 bis 2027.

Auf der Basis der reziproken Beeinflussung der Entwicklung von Technologien und Forschungsmethoden, entstand vor gut 20 Jahren an der Schnittstelle zwischen Sozialwissenschaften (*social science*), Informatik (*computer science*) und Maschinenbau (*engineering*) das Feld der *Computational Social Sciences* (CSS) (vgl. Edelmann et al. 2020). Hierbei geht es darum, soziale und soziotechnische Phänomene anhand von digitalen Verhaltensdaten (*digital behavioral data*) zu erforschen (vgl. Lazer et al. 2009; Lazer et al. 2020). Die Daten werden unter anderem von Online-Plattformen wie sozialen Netzwerken, von Geräten wie Smartphones oder mithilfe spezieller Software (zum Beispiel Browser-Erweiterungen) gesammelt. Je mehr Daten dabei anfallen, umso eine größere Rolle spielt „die maschinelle Erfassung und Auswertung großer Datenmengen" (Rieder 2021, S. 310), auch bekannt unter dem Schlagwort *Big Data*. Im Jahr 2018 schrieben Weyer et al. bezüglich Big Data, dass sich „in der deutschsprachigen soziologischen Diskussion nur wenige Arbeiten [finden lassen], die mit neueren datenanalytischen Methoden und großen Datensätzen arbeiten" (S. 116). Als Ursache dafür machen sie „unzureichende Kenntnisse neuer datenanalytischer Methoden" und „die Black Box algorithmischer Berechnungen" (ebd.) in den Sozialwissenschaften aus. Bei Lazer et al. 2009 findet sich zusätzlich die Einschätzung, dass der Zugriff auf proprietäre Daten eine der größten Herausforderungen der CSS sei: „Perhaps the thorniest challenges exist on the data side, with respect to access and privacy" (S. 722). Die vorliegende Studie erarbeitet daher den aktuellen Stand der Erreich- und Verwendbarkeit von Daten für die Technikfolgenabschätzung.

Die „Berücksichtigung der Einführungs- und Nutzungsbedingungen sowie insbesondere der Auswirkungsfelder technischer Innovationen" ist eine Kernaufgabe der TA (Dierkes 1994, S. 2004). Mit Blick auf gesellschaftliche Bezüge ist TA

> *„eine durch reflexiven Erkenntnisgewinn motivierte Forschungspraxis zur wissenschaftlichen Analyse von dynamischen und komplexen sozio-technischen Konstellationen in Politik beratender Absicht" (Böschen et al. 2021, S. 23).*

Bei der digitalen Transformation von Forschung überlagern sich beide Analyseperspektiven der Technikfolgenabschätzung, die Abschätzung der Reichweite technischer Innovationen sowie das genuine Erkenntnisinteresse der TA – die Analyse sozio-technischer Veränderungsprozesse. Gleichzeitig stellen diese neuen Anforderungen an TA-Methodologie. So erfordert die digitale Transformation von Forschung, wendet man diese auf die TA an, eine Überprüfung des eigenen methodischen Zuschnitts und eine Erweiterung des methodischen Zugriffs im Rahmen des TA-Methodenportfolios, um digitale Transformationsprozesse zu analysieren.

> *„TA methodology can be conceptualized along the assessment process with its requirements, boundary conditions. […] While TA methodology at the micro level builds the bridge to scientific knowledge in its usual understanding, the assessment at the macro level is a specificity of TA. Hence, TA methodology is a compilation of methods and procedures adapted to TA's assessment process and well reflected with respect to value-sensitive issues and contextual circumstances" (Grunwald 2019, S. 198 f.).*

Computergestützte Modellsimulationen zur Bewertung und Begründung politischer Maßnahmen sind für die Politikberatung innerhalb der TA schon heute relevant (Kaminski et al. 2023). Weitere TA-relevante Methoden mit Fokus auf Partizipation und Internetforschung, sind bspw. ‚living labs' (van Geenen und Kinder-Kurlanda 2022) oder auch Online-Tools zur Erfragung normativer Orientierungen von Teilnehmenden in Partizipationsprozessen (Mader et al. 2019). Riousset et al. 2024 fassen die Grenzen und Potenziale digitaler Methoden für die Technikfolgenabschätzung zusammen und beziehen dabei auch Sprachmodelle mit ein. Datengestützte Ansätze im Sinne der CSS haben bislang noch keinen Einzug in das Methodenportfolio der TA gehalten.

Die in diesem Band veröffentlichte Studie versuchen, diese Lücke zu schließen. Die Untersuchung von Wiedemann gibt aus der Perspektive der CSS einen Überblick, zum einen über die Entwicklung der deutschen Forschungslandschaft und stellt zum anderen auch Anwendungen im Forschungsbereich der Technikfolgenabschätzung zusammen.

Im Anschluss an ein einführendes Kapitel werden in Kapitel 2 methodische Zugänge zu Daten, d. h. sowohl Datenarten (Abschnitt 2.1) als auch Datenquellen (Abschnitt 2.2) und der Prozess der Datenerhebung selbst (Abschnitt 2.3) beschrieben. In Kapitel 3 wendet er sich möglichen Auswertungsmethoden zu, darunter der Inhaltsanalyse (Abschnitt 3.1), der Netzwerkanalyse (Abschnitt 3.2), der Simulation (Abschnitt 3.3) sowie der Statistik (Abschnitt 3.4). In Kapitel 4 stellt Wiedemann die CSS-Forschungslandschaft in Deutschland dar. Im fünften Kapitel wird eine „CSS-Toolbox für Einsteiger" präsentiert, in der verschiedene Programme und Anwendungsmöglichkeiten mit ihren Vor- und Nachteilen spezifisch für das Feld der TA gegenübergestellt werden. Kapitel 6 konkretisiert die vorherigen Ausführungen anhand des Fallbeispiels ‚Metaverse'-Diskurs.

Bei allem Potenzial des Einsatzes digitaler Methoden in der Forschung weist Wiedemann auch auf die Grenzen der Forschungsmethoden hin. Aufgrund notwendiger Kenntnisse in Statistik und Informatik, empfiehlt er, datenbasierte TA-Forschungsprojekte in interdisziplinären Teams durchzuführen oder Forschungsprojekte direkt aus der Informatik heraus mit sozialwissenschaftlichen Kooperationspartnern anzugehen. Diese interdisziplinäre Zusammenarbeit birgt Chancen und gleichzeitig Herausforderungen der Zeitlichkeit interdisziplinärer Zusammenarbeit aber auch der reflexiven Betrachtung und Einordnungen datenbasierter Forschungsmethoden:

> *„Die informatische Seite einer Kooperation betrachtet ihre Arbeit häufig als erledigt an einer Stelle, an der es für den sozialwissenschaftlichen Erkenntnisgewinn erst richtig interessant wird. Insofern ersetzt der Einsatz computergestützter Verfahren weder die Notwendigkeit sorgfältiger Validierung noch qualitativer Interpretation der auf Basis sehr großer Datenmengen abgeleiteten Erkenntnisse" (S. 46).*

In welcher Weise die digitale Transformation nicht nur Gegenstand der Analyse, sondern auch eine Transformation mit Blick auf Fragestellungen und Methoden der TA-Forschung selbst bedeutet, lässt sich heute noch nicht abschließend beschreiben. Wir möchten mit dieser vom ITAS beauftragten

Überblicksstudie eine Diskussionsgrundlage bieten, dies zum einen mit Blick auf TA-Forschungsmethoden zu reflektieren. Zum anderen möchten wir die TA-Community dazu einladen, auf Basis neuer methodischer Möglichkeiten auch neue Forschungsfragen mit Blick auf die digitale Transformation zu entwickeln.

Karlsruhe, im Juli 2024

Linda Nierling und Leonie Seng

Literatur

Albrecht, Steffen (2024): „ChatGPT als doppelte Herausforderung für die Wissenschaft. Eine Reflexion aus der Perspektive der Technikfolgenabschätzung.“ Gerhard Schreiber und Lukas Ohly (Hg.): KI:Text. Diskurse über KI-Textgeneratoren. Berlin: De Gruyter, S. 13–28. https://doi.org/10.1515/9783111351490-003

Böschen, Stefan; Grunwald, Armin; Krings, Bettina-Johanna; Rösch, Christine (2021): „Technikfolgenabschätzung – neue Zeiten, neue Aufgaben.“ In: *Technikfolgenabschätzung. Handbuch für Wissenschaft und Praxis*, hg. v. Stefan Böschen, Armin Grunwald, Bettina-Johanna Krings und Christine Rösch. Baden-Baden: Nomos, S. 15-40.

Deutscher Ethikrat (2023): „Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz. Stellungnahme.“ Online verfügbar unter https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf, zuletzt geprüft am 13.03.2024.

Dierkes, Meinolf (1994): „Technikfolgen-Abschätzung.“ In: *Handwörterbuch des Umweltrecht* 2. 2., hg. v. Otto Kimminic; überarbeitete Auflage. Berlin: Erich Schmidt Verlag, S. 2003-2019.

Edelmann, Achim; Wolff, Tom; Montagne, Danielle; Bail, Christopher (2020): “Computational Social Science and Sociology.“ In: *Annual Review of Sociology* 46, S. 61–81. https://doi.org/10.1146/annurev-soc-121919-054621

EK – Europäische Kommission (2020a): „Artificial intelligence – ethical and legal requirements.“ Online verfügbar unter https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/public-consultation_en, zuletzt geprüft am 21.03.2024.

EK (2020b): „Künstliche Intelligenz – ethische und rechtliche Anforderungen.“ Online verfügbar unter https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Kunstliche-Intelligenz-ethische-und-rechtliche-Anforderungen/feedback_de?p_id=8242911, zuletzt geprüft am 21.03.2024.

EK (2020c): „Künstliche Intelligenz – ethische und rechtliche Anforderungen.“ Online verfürbar unter https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Kunstliche-Intelligenz-ethische-und-rechtliche-Anforderungen/feedback_de?p_id=24212003, zuletzt geprüft am 21.03.2024.

Grunwald, Armin (2022): „Kapitel 3. Verantwortliche Gestaltung von KI-Systemen. Ethik, Recht und Technikfolgenabschätzung.“ In: *Künstliche Intelligenz. Ethik und Recht. Information und Recht 87*, hg. v. Thomas Hoeren und Stefan Pinell. München: C. H. Beck, S. 45-66.

Grunwald, Armin (2019): *Technology Assessment in Practice and Theory*. London: Routledge.

Heil, Reinhard (2021): „Künstliche Intelligenz/ Maschinelles Lernen.“ In: *Handbuch Technikethik* 2, hg. v. Armin Grunwald und Rafaela Hillerbrand. aktualisierte und erweiterte Auflage. Baden-Baden: Nomos, S. 424-428.

Jahnel, Jutta; Heil, Reinhard (2024): „KI-Textgeneratoren als soziotechnisches Phänomen – Ansätze zur Folgenabschätzung und Regulierung.“ In: *KI:Text. Diskurse über KI-Textgeneratoren*, hg. v. Gerhard Schreiber und Lukas Ohly. Berlin: De Gruyter, S. 341–354. https://doi.org/10.1515/9783111351490-021

Kaminski, Andreas; Gramelsberger, Gabriele; Scheer, Dirk (2023): “Modeling for policy and technology assessment: Challenges from computer-based simulations and artificial intelligence.“ In: *TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis* 28(1), S. 58-64. https://doi.org/10.14512/tatup.28.1.58

Lazer, David et al. (2020): “Computational social science: Obstacles and opportunities Data sharing, research ethics, and incentives must improve.“ In: *Science* 369 (6507), S. 1060-1062. https://10.1126/science.aaz8170

Lazer, David et al. (2009): „Computational Social Science.“ In: *Science* 323 (5915), S. 721-723. https://10.1126/science.1167742

Mader, Clemens; Hilty, Lorenz; Som, Claudia; Wäger, Patrick (2019): “Transparenz normativer Orientierungen in partizipativen TA-Projekten. Ein Software-basierter Ansatz.“ In: *TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis* 31(1), S. 62-63. https://doi.org/10.14512/tatup.31.1.62

Rieder, Gernot (2021): „Big Data.“ In: *Handbuch Technikethik* 2, hg. v. Armin Grunwald und Rafaela Hillerbran; aktualisierte und erweiterte Auflage. Baden-Baden: Nomos, S. 310-314.

Riousset, Pauline; Madsen, Anders Koed; Baya-Laffite, Nicolas; Villard, Lionel (2024): „Digital methods for technology assessment.“ In: *Handbook of Technology Assessment*, hg. v. Armin Grunwald. Cheltenham: Edward Elgar Publishing, S. 387-396.

Schrape, Jan-Felix (2021): „Digitalisierung und Technikfolgenabschätzung.“ In: *Technikfolgenabschätzung. Handbuch für Wissenschaft und Praxis*, hg. v. Stefan Böschen, Armin Grunwald, Bettina-Johanna Krings und Christine Rösch. Baden-Baden: Nomos, S. 83-96.

van Geenen, Daniela; Kinder-Kurlanda, Katharina (2022): „Meeting report: ‘Re-thinking/re-configuring participation’. Conference, 2021 (online).“ In: *TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis* 31(1), S. 62-63. https://doi.org/10.14512/tatup.31.1.62

Weyer, Johannes; Delisle, Marc; Kappler, Karolin; Kiehl, Marcel; Merz, Christina; Schrape, Jan-Felix (2018): „Big Data in soziologischer Perspektive.“ In: *Big Data und Gesellschaft. Eine multidisziplinäre Annäherung*, hg. v. Barbara Kolany-Raiser, Reinhard Heil, Carsten Orwat und Thomas Hoeren. Wiesbaden: Springer VS, S. 69-15.

Zusammenfassung

Diese Studie gibt einen Überblick über Gegenstände und Methoden der Computational Social Science. Schwerpunkte legt die Betrachtung auf die Entwicklung in der deutschen Forschungslandschaft sowie für mögliche Anwendungen im Forschungsbereich der Technikfolgenabschätzung. Vorgestellt werden Datenarten, -quellen und -zugänge sowie die gängigsten Auswertungsmethoden in der Forschung mit digitalen Daten. Für einen Einstieg werden Vorschläge für die Softwarenutzung auf unterschiedlichen Komplexitätsstufen gemacht.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ACL	Association of Computational Linguists
API	Application Programming Interface
CSS	Computational Social Sciences
DH	Digital Humanities
DIP	Dokumentations- und Informationssystem f. Parlamentsmaterialien
ITAS	Institut für Technikfolgenabschätzung und Systemanalyse
KI	Künstliche Intelligenz
KIT	Karlsruher Institut für Technikfolgenabschätzung
MCL	Meta Content Library
OCR	Optical-Character-Recognition-Software
TA	Technikfolgenabschätzung

Digitale Methoden für die Technikfolgenabschätzung

Überblicksstudie

Dr.-Ing. Gregor Wiedemann
Leibniz-Institut für Medienforschung
Hans-Bredow-Institut
g.wiedemann@leibniz-hbi.de

1 Technikfolgenabschätzung in Zeiten der Digitalisierung

Im Rahmen der zunehmenden Digitalisierung der Gesellschaft hat sich in den letzten Jahren die ‚Computational Social Science' (CSS) als neues wissenschaftliches Teilgebiet etabliert. In der CSS finden verschiedene Disziplinen mit jeweils unterschiedlichen Fachtraditionen und Wissenschaftsverständnissen zusammen, um mithilfe von computergestützten Methoden wie Netzwerkanalyse, Text- und Data-Mining oder Computersimulation soziale Prozesse und Phänomene zu untersuchen. Im Zuge des technologischen Fortschritts hat insbesondere die Bedeutung der systematischen Analyse sehr großer digitaler Datensätze (Big Data) an Bedeutung gewonnen.

1.1 Ziele

Analog zu anderen sozialwissenschaftlichen Bereichen können CSS-Methoden für die Technikfolgenabschätzung (TA) als Teilgebiet der Techniksoziologie einen großen Beitrag zur Erweiterung des Methodenspektrums leisten. Diesen methodischen Beitrag näher zu betrachten ist Gegenstand der vorliegenden Studie im Auftrag der Forschungsgruppe „Digitale Technologien und gesellschaftlicher Wandel"[1] des Karlsruher Instituts für Technologie (KIT), Institut für Technikfolgenabschätzung und Systemanalyse (ITAS). Im Zentrum der ITAS-Forschungstätigkeit stehen gesellschaftliche Prozesse der Digitalisierung sowie ihre Einordnung und Bewertung mit Blick auf Konzepte und Methoden der Technikfolgenabschätzung (TA), die selbst als spezifischer Forschungstyp im Feld der transformativen Forschung verortet wird. Die TA ist gekennzeichnet durch problem- bzw. folgenorientierte Forschung zur Produktion eines Wissens zum Handeln, welches häufig einen direkten Gesellschafts- und Beratungsbezug aufweist (vgl. u. a. Nentwich 2023, Abschnitt 1.1.1). Ihre

[1] https://www.itas.kit.edu/fg_digit.php, zuletzt geprüft am 04.06.2024.

Arbeitsweise ist daher notwendig inter- und transdisziplinär. Böschen et al. (2021, S. 23) definieren TA als

> *„eine durch reflexiven Erkenntnisgewinn motivierte Forschungspraxis zur wissenschaftlichen Analyse von dynamischen und komplexen sozio-technischen Konstellationen in Politik beratender Absicht“ (Herv. i. Orig.).*

Entsprechend breit gefächert ist das bereits in der TA etablierte Methodenspektrum, dass sich aus quantitativen Ansätzen wie Surveys und Inhaltsanalysen sowie qualitativen Auswertungen beispielsweise von Experteninterviews, Fokusgruppen und Mediendiskursen zusammensetzt. Die CSS haben das Potenzial, das bisherige Methodenspektrum in dreierlei Hinsicht zu erweitern:

- Datenerhebung: Die CSS erschließt neue, digitale Datensätze für die Erforschung von TA-Fragestellungen (digitale Textkorpora, Netzwerke, Bilder und Videos, Tracking und Log-Daten).
- Auswertungsverfahren: Zur Auswertung dieser Daten stehen (teil-) automatisierte computergestützte Verfahren zur Verfügung, mit denen die meist sehr großen Datenmengen überhaupt erst bewältigt werden können.
- Beteiligungsverfahren: Die in der TA etablierten partizipativen Forschungsmethoden können durch digitale Formate erweitert werden.

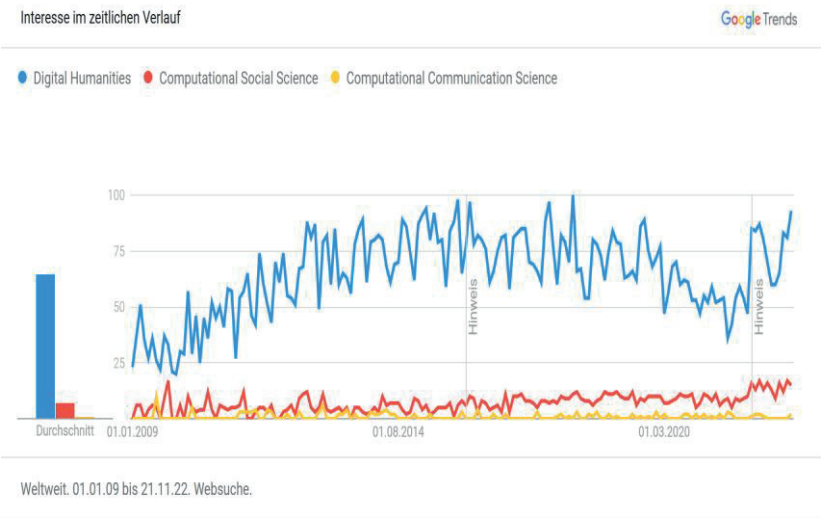

Abbildung 1.1: Entwicklung des Suchinteresses für disziplinäre Teilgebiete.

1.2 Computational turns

Die Forschung in den Sozial- und Geisteswissenschaften ist in den vergangenen zwei Jahrzehnten durch eine stetige Zunahme des Einsatzes digitaler Erhebungs- und Auswertungsmethoden gekennzeichnet, die in der Literatur häufig als ‚computational turn' charakterisiert wird. Dabei können die ‚Digital Humanities' (DH) als Vorreiter gelten, deren Anfänge bereits bis in die 1960er Jahre zurückreichen. Die wichtigste internationale Fachkonferenz der Community wird seit 2006 unter dem Titel „Digital Humanities" durchgeführt, hat ihren Ursprung aber unter anderem Namen bereits Anfang der 1990er Jahre. Der Fachverband „Digital Humanities im deutschsprachigen Raum" (DHd) wurde 2013 gegründet. Eine Beobachtung der die Entwicklung begleitenden Suchbegriffe mit Google Trends zeigt, dass die Sozialwissenschaften deutlich später und in geringer Quantität die Entwicklung hin zu einer Digitalisierung des Faches vollziehen (Abbildung 1.1). Der Artikel 2009 in Science erschienene Artikel „Computational Social Science" von Lazer et al. (2009) kann als namensprägend für das Feld angesehen werden. Seit 2015 findet die Interna-

tional Conference on Computational Social Science (IC²S²) als wichtigstes Forum statt. Darüber hinaus ist seit einigen Jahren eine weitere Ausdifferenzierung entlang etablierter Disziplinen in beispielsweise ‚Computational Communication Science' und ‚Computational Political Science' zu beobachten. Die Adaption des „computational turn" in den sozialwissenschaftlichen Teildisziplinen folgt dabei einer anderen Entwicklungslinie als die DH. Für letztere kamen die Impulse in erster Linie aus den Geisteswissenschaften selbst. Erst nach und nach wuchs das Interesse von Seiten der Informatik an dem neuen Forschungsfeld. So fand die DHd-Tagung 2014 beispielsweise unter dem Motto „methodischer Brückenschlag oder feindliche Übernahme?" statt (Meyer 2014). Im Gegensatz dazu wurde die CSS-Entwicklung von Anfang an maßgeblich von Akteuren vorangetrieben, die disziplinär außerhalb der Sozialwissenschaften angesiedelt sind. Forscher*innen aus der Physik, Mathematik und Informatik dominierten lange Zeit die Teilnehmerschaft an den einschlägigen Konferenzen und Workshops,[2] was der CSS bisweilen die Kritik einbrachte, sich zwar mit sozialen Phänomenen zu beschäftigen, aber leider keine Sozialwissenschaft zu betreiben. So mahnen Waldherr et al. (2021) an, die CSS-Forschung deutlich stärker mit sozialwissenschaftlichen Theorien mittlerer und großer Reichweite zu verzahnen, anstatt sich mit theorieloser Deskription der Beobachtung von Mustern in großen Datenmengen zu begnügen. Solche Forderungen, wie auch die jüngere Adaption des Attributs ‚computational' für die vielfältigen sozialwissenschaftlichen Teildisziplinen, verdeutlichen dabei ein Bestreben, die Erforschung und Nutzung digitaler Datenerhebungs- und Auswertungsmethoden in den Disziplinen selbst nachhaltig zu verankern und eine Verbindung mit den eigenen, etablierten Fachstandards anzustreben, welche die ‚frühe' CSS-Forschung nicht herzustellen vermochte. In diesem Zuge findet die Arbeit mit ‚Big Data' aus digitalen Beobachtungs- und Simulationsprozessen, Textsammlungen oder audio-visuellen Datenquellen mit Hilfe von Datenbanken, Text und Data Mining, maschinellem Lernen und statistischer Modellierung nach und nach Eingang in das Methodenrepertoire sowie die

[2] Weniger als 25 % der Teilnehmer*innen der zweiten IC²S² 2016 in Köln gaben an, eine sozialwissenschaftliche Ausbildung zu haben.

Lehrpläne von Soziologie, Politikwissenschaften, Kommunikations- und Medienforschung und vielen weiteren benachbarten Disziplinen.

1.3 Aufbau dieser Studie

Zur Unterstützung der TA bei der Erschließung dieser neuen Methoden beschreibt diese Studie im Folgenden wesentliche Aspekte des CSS-Forschungsfeldes und bewertet diese im Hinblick auf die Potenziale für das Anwendungsgebiet. Zunächst werden in Kapitel 2 Arten, Quellen und Zugänge für digitale Forschungsdaten vorgestellt. Kapitel 3 widmet sich den gängigen Auswertungsmethoden zur Bewältigung großer Datenmengen. Kapitel 4 beschreibt in Auszügen wichtige Akteure, Projekte und Events der CSS-Forschungslandschaft in Deutschland. Kapitel 5 schlägt eine Auswahl von Softwareprogrammen und -bibliotheken vor, mit denen eine Annäherung an CSS-Methoden für die TA-Forschung angegangen werden kann. Im vorletzten Kapitel 6 werden anhand eines Fallbeispiels zum Diskurs rund um das so genannte ‚Metaverse' einzelne Schritte und zu erwartende Ergebnisse eines CSS-Forschungsvorgehens für eine TA-Fragestellung exemplarisch aufgezeigt. Das siebte und letzte Kapitel listet noch einmal wichtige Lern- und Lehrressourcen auf, mit denen eine Erschließung des CSS-Forschungsfeldes über diese Studie hinaus fortgesetzt werden kann.

2 Datenarten, -quellen und -zugänge

Die CSS-Forschung ist im Wesentlichen durch die Entstehung, Beobachtbarkeit und Auswertbarkeit sehr großer Mengen ‚digitaler Spurendaten' mit Relevanz für soziale Prozesse und Phänomene geprägt. Vor diesen Hintergrund werden in diesem Kapitel verschiedene Arten und Bezugsquellen von digitalen Spurendaten beschrieben, die auch im Forschungsfeld TA von Interesse sein können. In den einzelnen Unterabschnitten werden einschlägige Forschungsarbeiten und Projekte beispielhaft und ohne Anspruch auf Vollständigkeit oder Repräsentativität für das jeweilige Forschungsgebiet aufgezählt.

2.1 Datenarten

2.1.1 Text

Digitale Texte sind eine der informationsreichsten Datenarten im CSS-Bereich. Unterschieden wird zwischen ‚born-digital' und retro-digitalisierten Dokumenten. Erste werden in der Regel durch die Digitalisierung von Archiven mit Hilfe von Scannern und Optical-Character-Recognition-Software (OCR) erzeugt. Die Qualität der Daten hängt hier maßgeblich von der Beschaffenheit der Ausgangsdokumente sowie der verwendeten Software ab. Zahlreiche Forschungsprojekte im Bereich der Digital Humanities sowie der Bibliotheks- und Informationswissenschaft beschäftigen sich mit der Verbesserung dieser Abläufe. ‚Born-digital' Textdaten entstehen dagegen durch die Nutzung digitaler Kommunikations- und Datenverarbeitungsinfrastruktur. Dabei kann es sich um digitale Archive organisationaler Einheiten handeln wie bspw. das Dokumentations- und Informationssystem für Parlamentsmaterialien (DIP) des Bundestages oder die Community-basierte Wissenssammlung Wikipedia. Die maßgebliche Quelle für digitale Textdaten in der CSS-Forschung sind

jedoch Internet-basierte Kommunikationsinhalte wie redaktionellen Nachrichtenartikel aus Online-Medienangeboten, Sammlungen von Webseiten, Blogs und Foren oder öffentliche Postings auf sozialen Medienplattformen. Während retrodigitale Archive meist ‚nur' wenige Zehn- bis hunderttausend Dokumente umfassen, können Social-Media-Analysen schnell deutlich größere Datenmengen umfassen. Für die Erhebung, Verarbeitung und ggf. Veröffentlichung bzw. langfristig Aufbewahrung digitaler Textdaten sind verschiedene Rechtsvorschriften zu beachten—in erster Linie Urheber bzw. Leistungsschutzrecht, allg. Persönlichkeitsrechte und das Datenschutzrecht. Ggf. müssen Lizenzen zur Nutzung z. B. digitaler Zeitungsarchive beschafft werden bzw. Zugänge zu Archiven von kommerziellen Anbietern eingekauft werden. Für wissenschaftliche Zwecke stellen § 60d UrhG zur Datenerhebung sowie § 27 BDSG bezüglich der Datenverarbeitung die maßgeblichen rechtlichen Grundlagen dar. Beispiele aus der CSS-Forschung mit Textdaten liefern die folgenden Veröffentlichungen:

- Die DFG-Initiative OCR-D fördert seit 2015 zahlreiche Projekte zur Verbesserung der Massendigitalisierung historischer, deutschsprachiger Dokumente und Dokumentarchive (Engl 2020).[1]
- Das Projekt Europeana zur Digitalisierung des kulturellen Erbes Europas macht hunderttausende historischer Zeitungen aus 20 Ländern von 1618 bis in die 1980er Jahre zugänglich (Pekárek und Willems 2012).[2]
- Das Projekt „Wissensmanagement von Altdokumenten aus Forschung, Verwaltung und Betrieb" (Eck, Hensel und Kappei 2020) befasste sich mit der Aufbereitung und Auswertung der umfangreichen Dokumentation des Forschungsbetriebes Schachtanlage ASSE II zu Zwecken der Endlagerung radioaktiver Abfälle (ca. 1 Mio. Seiten)[3].

[1] https://ocr-d.de/de/, zuletzt geprüft am 06.06.2024.

[2] https://www.europeana.eu/de/collections/topic/18-newspapers, zuletzt geprüft am 06.06.2024.

[3] Das Projekt wurde ab 2020 nicht weiter gefördert und das mit dem Projekt verbundenen Institut für Wissensmanagement und Wissenssynthese (IWW) abgewickelt. Zusammen mit dem Abschlussbericht wurde jedoch ein Großteil der bis dahin digitalisierten Archivdaten veröffentlicht und harrt der Auswertung.

- Das Projekt „ePol – Postdemokratie und Neoliberalismus“ untersuchte anhand eines Korpus von ca. 3,5 Millionen Zeitungsartikeln seit 1949 die Ökonomisierung politischer Argumentation (Schaal et al. 2016) sowie die Diskurse demokratischer Grenzziehung gegenüber politischen ‚Extremen‘ in Deutschland (Wiedemann 2016).
- Su et al. (2018) werten 243 Millionen Facebook-Kommentare im Zusammenhang mit den US-Wahlen 2016 aus und stellen fest, dass die Nutzung inziviler Sprache zwischen den politischen Lagern in den Vereinigten Staaten signifikant zugenommen hat.
- Rauh und Schwalbach (2020) stellen mit dem ParlSpeech-V2-Dataset einen Korpus mit mehr als 6 Mio. Parlamentsreden aus mehreren europäischen Ländern zu Verfügung. Weitere Parlamentskorpora sind über das Infrastrukturprojekt CLARIN mit der Initiative Parla-CLARIN[4] digital verfügbar gemacht worden.

Die Auswertung ‚kleinerer‘ Textkorpora in der Größenordnung von ca. 100.000 Dokumenten kann auf leistungsfähigen Einzelrechnern mit Hilfe von Textanalysesoftware bzw. eigenen geschriebenen Programmen ausgewertet werden. Für größere Datensammlungen empfiehlt sich der Einsatz von Datenbanksystemen und Volltextindex-Technologien auf leistungsfähigen Servern, um die Daten schnell durchsuch- und auswertbar zu machen. Dies gilt auch für den Einsatz von Machine Learning Technologien für die Automatische Sprachverarbeitung.

2.1.2 Audio-visuelle Inhalte

Zusätzlich zu Text hat sich insbesondere mit der weiteren Verbreitung von Social-Media-Angeboten die Verbreitung und der Konsum audio-visueller Daten (Bilder, Videos, Audiodokumente) massiv erhöht. Der Anbieter YouTube erlaubt das Teilen von Videos beliebiger Länge in personalisierten Kanälen. Plattformen wie Instagram setzen vor allem auf Fotos, die unter Nutzer*innen geteilt werden können. Der Anbieter TikTok verteilt mit Hilfe eines speziellen

[4] https://www.clarin.eu/resource-families/parliamentary-corpora, zuletzt geprüft am 06.06.2024.

Empfehlungsalgorithmus nutzergenerierte, kurze Videoclips. Messenger-Dienste wie WhatsApp und klassische soziale Netzwerke wie Facebook haben mit der sogenannten ‚Story'-Funktion eine Nutzungsart geschaffen, bei der Nutzer*innen Bild- und Video-Inhalte zeitlich begrenzt öffentlich sichtbar machen können. Über diese multimedialen Kanäle kommunizieren Nutzer*innen nicht nur Entertainmentinhalte, sondern verbreiten zunehmend auch politisch und gesellschaftlich relevante Informationen. Die visuell vermittelten Informationen können in der Regel über Feedback- und Kommentarfunktionen von anderen Nutzer*innen bewertet und diskutiert werden. Beispiele aus der CSS-Forschung mit Bild- und Videodaten liefern die folgenden Veröffentlichungen:

- Araujo, Lock und Velde (2020) machen einen Vorschlag für ein systematisches methodisches Vorgehen zur automatischen Inhaltsanalyse von Bilddaten mit maschinellem Lernen.
- Haim und Jungblut (2021) analysieren ca. 80.000 Fotos von Kandidierenden der Europawahl 2019 hinsichtlich unterschiedlicher Merkmale ihrer der Selbstdarstellung und ihrer Darstellung in den Medien.
- Nyhuis et al. (2021) untersuchen die Anwesenheit von Parlamentarier*innen zu unterschiedlichen Tagesordnungspunkten im Landtag von Baden-Württemberg anhand von Video-Aufzeichnungen der Parlamentssitzungen.
- Mit CASM stellen Zhang und Pan (2019) einen Ansatz zur automatischen Identifikation von Protestereignissen in Social-Media-Postings vor und wenden diesen auf einem Korpus aus ca. 9.5 Mio. Posts der chinesischen Plattform Weibo an.

Die jüngeren Veröffentlichungsdaten der Beispiele machen deutlich, dass audiovisuelle Inhalte eine zunehmend wichtige Datenquelle für die CSS darstellen. Gleichzeitig befindet sich die Methodenentwicklung für die CSS hier aufgrund zahlreicher Herausforderungen noch in ihren Anfängen. Die Erhebung der Daten ist aufgrund ihrer Größe, ihrer eingeschränkten Zugänglichkeit und ihres erhöhten Eingriffs in Persönlichkeitsrechte deutlich gegenüber Textdaten erschwert. Automatische Datensammelprogramme werden meist durch die Betreiber identifiziert und gesperrt. Algorithmisch oder über Follower-Netzwerke verbreitete Videos sowie zeitlich begrenzt verfügbare Medieninhalte aus den

‚Story'-Funktionen entziehen sich einem systematischen Datenerhebungsprozess durch die Forschenden nach reproduzierbaren Kriterien. Zusätzlich sind die Möglichkeiten zur automatischen Auswertung begrenzt. Für die Umwandlung von Audiospuren in transkribierten Text stehen auf maschinellem Lernen beruhende Speech-to-Text-Technologien zur Verfügung.[5] Für die Erkennung von Textinhalten auf Bildern (z. B. bei sog. Sharepics, die häufig in der politischen Kommunikation zum Einsatz kommen) steht OCR-Technologie zur Verfügung. Auf den erkannten Textinhalten kann wiederum die ganze Bandbreite von Analyseverfahren der automatischen Sprachverarbeitung zum Einsatz gebracht werden. Zur Erkennung von Bildinhalten gibt es mittlerweile leistungsfähige Klassifikationssoftware aus dem Forschungsbereich Computer Vision, die bspw. Objekte auf Fotos, Personen und Geschlechter bei Gesichtern oder Körperposen automatisch erkennen kann. Im Gegensatz zu Textinformation sind die automatisch erfassbaren Kategorien aus visuellen Inhalten jedoch bislang recht grob, so dass sie nur für wenige Forschungsfragen geeignet scheinen.

2.1.3 Netzwerke

Digitale Spurendaten sind das Ergebnis sozialer Prozesse, also der Interaktion von miteinander verbundenen Individuen. Diese Verbindungen als Netzwerke zu (re-)konstruieren und ihre Eigenschaften zu analysieren ist von Beginn an eine der bedeutsamsten Methodenansätze in der CSS-Forschung. Netzwerke bestehen aus Einheiten (Knoten) und deren Verbindungen (Kanten), die jeweils mit bestimmten Eigenschaften wie beispielsweise ein Gewicht oder eine diskrete Variablenbelegung verknüpft sein können. Kanten können zudem ungerichtet oder gerichtet sein und damit markieren, dass eine Beziehung nur in eine Richtung oder beidseitig gilt. Ziel der Netzwerkanalyse ist die Bestimmung wesentlicher Charakteristika eines Netzwerks zu einem Zeitpunkt wie beispielsweise seine Dichte, das Finden zentraler Knoten oder die Identifikation von Cliquen (vollständig verbundene Teilnetzwerke) und Clustern (eng

[5] YouTube bietet die Möglichkeit an, Transkripte von Videos zu erstellen und diese über die API herunterzuladen.

miteinander verbundene Knotengruppen) sowie die Veränderung dieser Charakteristika über die Zeit. Welche Informationen in Knoten und Kanten eines Netzwerks repräsentiert werden, kann ganz verschieden sein. Häufig aber weisen Netzwerke aus sozialen Prozessen die ‚small-world'-Eigenschaft auf, die besagt, dass zwei beliebige Knoten in der Regel über eine sehr kurze Kette von weiteren Knoten miteinander verbunden sind.

- Adamic (1999) zeigt, dass es sich bei der Link-Struktur zwischen Webseiten des World Wide Web um ein small-world Netzwerk handelt.
- Anhand von Interaktions-Netzwerken von Twitter-Nutzenden beobachten Vosoughi, Roy und Aral (2018), dass sich Falschmeldungen bis zu 100-mal schneller in dem sozialen Netzwerk ausbreiten als wahre Nachrichtenmeldungen.
- Münch et al. (2021) vermessen die deutsche „Twittersphäre", indem sie die Follower-Beziehungen der einflussreichsten deutschsprachigen Twitter-Accounts rekonstruieren.
- Co-Autor*innen-Beziehungen bei wissenschaftlichen Publikationen formen Netzwerke, die Gegenstand der Bibliometrie-Forschung sind. Wang, Song und Su (2022) beschreiben das CSS-Forschungsfeld der letzten 20 Jahre anhand eines Zitationsnetzwerks.

Netzwerke aus digitalen Spurendaten zu erheben kann, ähnlich wie bei Text, schnell zu sehr großen Datenmengen führen, die besonderer Technologien zur Speicherung, Auswertung und visuellen Darstellung benötigen. Für die TA interessant wären beispielsweise die Netzwerke zentraler Akteure (Unternehmen, Regierungsinstitutionen, Expert*innen), die aus der Kommunikation über bestimmte Technologien aus sozialen Medien oder der gemeinsamen Erwähnung in der Presseberichterstattung abgeleitet werden können.

2.1.4 Log- und Trackingdaten

Bei der Verwendung digitaler Technologien können einzelne Aktivitäten wie beispielsweise das Abrufen bestimmter Daten, das Anmelden eine*r Nutzer*inoder die Ausführung einer Suchanfrage in einem System als Datenpunkt in Protokoll-Dateien (Logs) festgehalten werden. Werden einzelne Ereignisse

anhand bestimmter Eigenschaften über die Zeit miteinander verknüpft, beispielsweise alle Interaktionen eine*r identifizierten Nutzer*in mit einem System, handelt es sich um Tracking-Daten bzw. „digital trace data“ (Howison, Wiggins und Crowston 2011). Die Daten können entweder Server-seitig erhoben werden und stehen dabei von vornherein gesammelt zur Verfügung. Alternativ werden sie Client-seitig erhoben, z. B. auf den Computern und mobilen Endgeräten von Nutzer*innen und anschließend an eine zentrale Stelle übertragen. Insofern es sich um personenbezogene Daten handelt, was in der Regel der Fall ist, setzen beide Ansätze eine bewusste Einwilligung der Beobachteten voraus. Mit solchen Daten lassen sich unter anderem typische Nutzungsweisen technischer Systeme sowie Muster individueller Verhaltensweisen analysieren. Beispiele wären besuchte Internetseiten während bestimmter Tageszeiten, häufigste Suchanfragen auf einer Webseite von Nutzer*innen aus bestimmten Regionen oder typische Wege von der Startseite eines Online-Shops über Produktansichten bis zur Bestellung oder zum Verlassen der Seite. In den CSS kommen Tracking-Daten beispielsweise zur Beobachtung räumlicher Bewegungsmuster von Personen, zur Messung von Reputationskennzahlen in sozialen Netzwerken oder zur Beobachtung der Nutzungsweisen mobiler Endgeräte und Apps zum Einsatz.

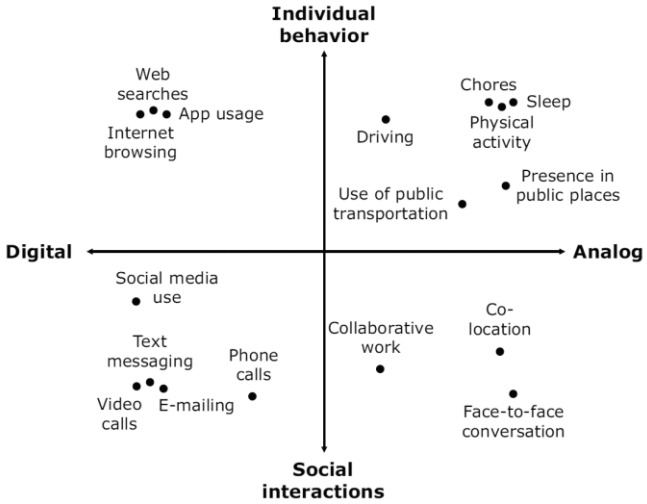

Abbildung 2.1: Beispiele für analoge und digitale Verhaltensweisen und Interaktionen, die anhand von digitalen Spurendaten untersucht werden können (Keusch und Kreuter 2021, 102).

Keusch und Kreuter (2021) unterscheiden vier Kategorien von Spurendaten in der CSS entlang zweier Dimensionen: 1. individuelle Spurendaten vs. Daten aus sozialer Interaktion und 2. die Beobachtung digitaler vs. analoger Phänomene (Abbildung 2.1). Ein Hauptvorteil bei der Forschung mit digitaler Spurendaten wird darin gesehen, dass es sich um nicht-reaktive, bzw. ‚unobstrusive' Methoden handele, bei denen Daten nicht durch Befragungseffekte verfälscht werden. Diese Eigenschaft gilt jedoch nicht für alle Studienarten. Beispielsweise erheben Forscher*innen Sensor- und App-Daten von Smartphones von Studienteilnehmer*innen, die dafür vorher freiwillig ihre Einwilligung erteilen. Häufig werden solche Studien durch Befragungen der Teilnehmer*innen über ihre Smartphones ergänzt (Elevelt, Lugtig und Toepoel 2019).

- Vosen und Schmidt (2011) gelingt es anhand von Daten des Dienstes Google Trends Verbraucherstimmungen und Konsumklimaindexe besser vorherzusagen, als dies mit Befragungsdaten gelingt.

- Ben-Zeev et al. (2015) zeigen, wie Smartphones zur Überwachung individueller Verhaltensindikatoren genutzt werden können, die Rückschlüsse auf die psychische Gesundheit der Nutzenden ermöglichen.
- Amaya et al. (2020) messen anhand von Interaktionsdaten im sozialen Netzwerks Reddit die Verbreitung von politischen Einstellungsdimensionen in der Bevölkerung.

Eine besondere Quelle von Log-Daten, die in der Regel ohne menschliche Beobachtungssubjekte auskommt, stellen Simulationsexperimente dar. Hierfür werden theoretische Annahmen über Abläufe in der sozialen Welt in Form eines Computerprogramms abstrahiert. Das Programm erzeugt mit variierten Parameter-Annahmen über die soziale Welt und abhängig von Zufallseinflüssen eine Vielzahl an Ausgaben über den Zustand bzw. die Veränderung der simulierten Welt.

2.2 Datenquellen

Die aktuelle CSS-Forschung basiert auf einer Vielzahl heterogener Datenquellen. Digitale soziale Spurendaten werden dabei häufig über Medienintermediäre als Vermittler zwischen Inhalte-Produzierenden und -Nutzenden erhoben.

2.2.1 Soziale Medien

In sozialen Medien treten Bürger*innen miteinander sowie mit Medienschaffenden, Expert*innen, Politiker*innen, staatlichen und nicht-staatlichen Institutionen in einen öffentlichen Austausch. Für die TA bergen verschiedene soziale Medienplattformen dabei unterschiedliche Potenziale, um die Rezeption von, den Umgang mit und die Erwartungen an neue Technologien zu beobachten. Wichtig ist zu beachten, dass Diskurse in den sozialen Medien nicht repräsentativ für die Gesamtbevölkerung sind und quantitative Ergebnisse daher nicht ohne weiteres verallgemeinert werden können. Gleichzeitig können über die Beobachtung von „Early Adopters“, „Pioniergemeinschaften“ (Hepp 2022), Expert*innengruppen und anderen Stakeholdergruppen besonders relevante Teildiskurse beobachtet werden. Wiedemann, Münch et al.

(2023) stellen Herausforderungen und Lösungen für einen typischen Forschungsprozess mit Social-Media-Daten vor, beginnend mit ethischen und rechtlichen Grundlagen der Forschung über computergestützte Erhebungs- und Auswertungsmethoden bis hin zu neuen Praktiken der Publikation von Forschungsdaten oder Analyseapplikationen. Für eine Orientierung beim Studiendesign besprechen Sen et al. (2021) ausführlich potenzielle Mess- und Repräsentationsfehler in der Erhebung und Auswertung von digitalen Spurendaten auf Online-Plattformen (vgl. Abbildung 2.2).

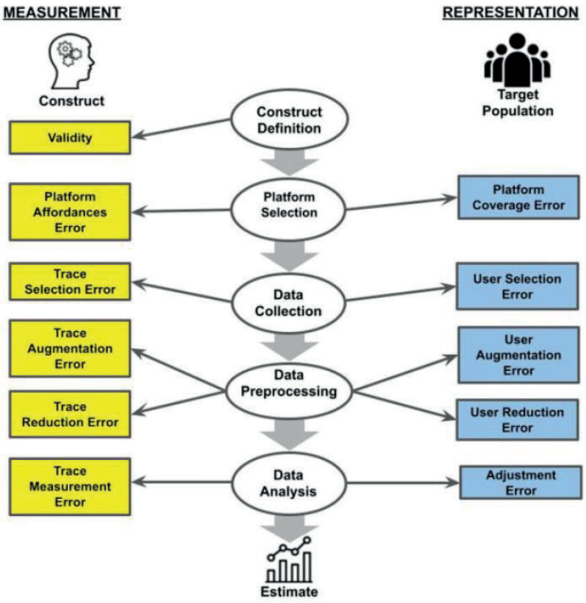

Abbildung 2.2: Potenzielle Mess- und Repräsentationsfehler im Lebenszyklus von Studien mit digitalen Spurendaten (Sen et al. 2021).

Tabelle 2.1: (Forschungs-)Datenzugänge einzelner Social-Media-Plattformen

Plattform	Datenzugang	Datenarten	Anmerkung
Facebook, Instagram	Meta Content Library and API[6]	Posts und Multimedia-Inhalte von öffentlichen Accounts, Anonymisierte Kommentare zu öffentlichen Posts	Antrag auf Zugang über das unabhängige Inter-university Consortium for Political and Social Research (ICPSR) der University of Michigan. Nutzungsbedingungen sehen Unterschrift der Forschenden und der rechtsverbindlichen Vertretung der Forschungseinrichtung vor.
Twitter/X	Antragsformular für Forschungsdatenzugang nach DSA Art. 40[7] Kommerzieller API-Zugang (z. B. „Pro“ mit 1M Posts für 5,000 EUR / Monat)[8]	X API v2 Endpunkte (Search, Filtered Stream), Nutzer*innen-Accounts und Posts	Kostenfreier Forschungsdatenzugang wird bislang kaum gewährt. Ablehnungen werden nur sehr allgemein begründet.
TikTok	TikTok Research Tools and API[9]	Daten zu Nutzer*innen-Accounts (inkl. Follower- und Followee-Beziehungen), Videos (Likes, Pins, Reposts)	Umfangreiches Antragsformular, dass sehr konkrete Beschreibungen des geplanten Projekts beinhaltet. Nutzungsbedingungen sehen z.T. fragwürdige Regeln vor (bspw. müssen Forschungsergebnisse vor Publikation an TikTok mitgeteilt werden)

[6] https://transparency.fb.com/researchtools/meta-content-library, zuletzt geprüft am 06.06.2024.
[7] https://developer.x.com/en/use-cases/do-research, zuletzt geprüft am 06.06.2024.
[8] https://developer.x.com/en/products/twitter-api, zuletzt geprüft am 06.06.2024.
[9] https://developers.tiktok.com/application/research-api, zuletzt geprüft am 06.06.2024.

YouTube	YouTube Researcher Program[10]	Erhöhte Rate-Limits zur Standard-API von YouTube[11]	Bewerbung setzt detaillierte technische Beschreibung des Forschungsvorhabens voraus.
Mastodon, BlueSky, Telegram	Nutzung der regulären APIs der Netzwerke mastodon[12], bluesky[13], telegram[14]	Account-Information und Post-Inhalte (inkl. Interaktionsmetriken)	Diese Netzwerke gewähren freien Zugang zu ihren APIs, s.d. alle Inhalte, die von den jeweils standardmäßig verbreiteten Clients angezeigt werden können, auch in eigenen Programmskripts abrufbar sind.
Reddit	Forschungsdatenzugang angekündigt im Mai 2024 auf reddit[15]	-	-

2.2.2 Facebook

Mit mehr als 30 Millionen wöchentlich aktiven Nutzenden ist das zum Konzern Meta gehörende Facebook das größte soziale Netzwerk in Deutschland (Koch 2022). Als hybride Plattform zwischen öffentlicher und privater Kommunikation bietet es seinen Nutzer*innen fein abgestufte Zugangskontrollen zu den geposteten Inhalten an. Neben den Kommunikationsinhalten, die an direkte Freunde verbreitet werden, bietet Facebook eine Gruppenfunktion, die es nicht miteinander verbundenen Nutzer*innen ermöglicht, miteinander zu interagieren, z. B. für einen Austausch über lokale, politische oder Freizeitinteressen. Die Plattform ermöglicht auch öffentliche Seiten für Prominente,

[10] https://research.youtube/, zuletzt geprüft am 06.06.2024.
[11] https://developers.google.com/youtube/v3/getting-started/?target=_blank, zuletzt geprüft am 06.06.2024.
[12] https://docs.joinmastodon.org/client/intro/, zuletzt geprüft am 06.06.2024.
[13] https://docs.bsky.app/, zuletzt geprüft am 06.06.2024.
[14] https://core.telegram.org/, zuletzt geprüft am 06.06.2024.
[15] https://www.reddit.com/r/reddit4researchers/, zuletzt geprüft am 06.06.2024.

Organisationen und Politiker*innen, die von jedermann verfolgt werden können. Unter Politiker*innen ist Facebook noch beliebter als Twitter/X (ca. 61 % aller Kandidat*innen der Wahl 2021 hatten eine öffentliche Seite) (Schmidt 2021). Posts auf öffentlichen Seiten können mit öffentlichen Kommentaren durch andere FB-Nutzer*innen zu einem lebendigen Meinungs- und Gedankenaustausch führen. Bei politischen Themen nehmen (rechts-)populistische politische Inhalte einen unverhältnismäßig hohen Anteil an der Nutzer*innen-interaktion ein (Thiele und Turnšek 2022). Zugang zu öffentlichen Facebook-Post, nicht jedoch zu Kommentaren, ist über die ebenfalls zum Meta-Konzern gehörende Plattform CrowdTangle verfügbar. Die Erhebung von Kommentardebatten auf Facebook ist daher nur sehr eingeschränkt über Web Scraping möglich. Zudem hat Crowd-Tangle eine Abschaltung seiner Dienste zum 14. August 2024 angekündigt.[16] Als Ersatz baut Meta derzeit die Meta Content Library (MCL) aus. Die MCL bietet über eine Weboberfläche sowie über eine API (siehe Abschnitt 2.3) Zugriff auf die gesamten öffentlichen Archiv-Inhalte von Facebook und Instagram (zum Beispiel Instagram-Accounts und Posts und Facebook-Seiten, Posts, Gruppen, Events, Profile und Kommentare). Allerdings kann nur ein kleiner Teil der Daten von besonders öffentlich bekannten Personen und Organisationen aus der MCL heraus exportiert werden. Analysen auf größeren Datenbeständen und mit weniger öffentlich bekannten Accounts müssen innerhalb der von Meta zur Verfügung gestellten Infrastruktur (‚clean room environment') durchgeführt werden.

2.2.3 Twitter/X

Mit ca. 8 Millionen Nutzenden (Koch 2022) ist X, vormals Twitter, in Deutschland eine eher kleine Plattform. Für den politischen und gesellschaftlichen Diskurs war sie jedoch lange Zeit von besonderer Bedeutung, da viele Politi-

[16] Unter https://foundation.mozilla.org/de/campaigns/crowdtangle-petition/ (zuletzt geprüft am 06.06.2024) hat die Mozilla Foundation eine Petition gestartet, mit der Meta aufgefordert wird, das branchenführende Transparenz-Tool CrowdTangle in einem global wichtigen Wahljahr nicht abzuschalten. Tatsächlich werden aber schon seit Jahresbeginn 2022 keine neuen Nutzeranmeldungen für den Dienst zugelassen.

ker*innen, Journalist*innen, Wissenschaftler*innen und sonstige Expert*innen sie als direkten Kommunikationskanal in jeweils ihre thematische Teilöffentlichkeit hinein nutzen. So unterhielten beispielsweise rund 38 % aller Kandidat*innen zur Bundestagswahl 2021 einen Twitter/X-Account (Schmidt 2021). Auf Twitter/X wurden in den letzten Jahren immer wieder Debatten angestoßen, die auch den gesellschaftlichen Diskurs außerhalb der sozialen Netze geprägt haben (z. B. #metoo oder die #blacklivesmatter-Bewegung). Für die TA von besonderem Interesse sind Einschätzungen zu neuen Technologien, wie Fortschritten bei Impfstoffen oder im Bereich Künstliche Intelligenz (KI), die auf der Plattform von ganz unterschiedlichen gesellschaftlichen Teilgruppen diskutiert werden. Beispielsweise sind allein im Dezember 2022 mehr als 1,2 Mio. Tweets, die das Keyword ‚ChatGPT' enthalten, veröffentlicht worden (für erste Analysen der Twitter-Debatten zu ChatGPT siehe Heaton et al. 2024 und Leiter et al. 2024). Die Erhebung von Twitter-Daten kann anhand von Schlagworten oder Nutzer*innen-Accounts erfolgen. Für letztere Zugriffsstrategie müssen Listen mit für ein Forschungsthema einschlägigen Accounts recherchiert und gepflegt werden. Die Datenbank öffentlicher Sprecher*innen (DBöS) des Projekts Social-Media-Observatory (Schmidt et al. 2022) stellt beispielsweise Accountlisten von deutschen Parlamentarier*innen, Parteigliederungen, Journalist*innen und Medienangeboten bereit. Der Forschungsdatenzugang für wissenschaftliche Zwecke ist von Twitter in den Jahren 2019 bis 2023 vorbildhaft gestaltet worden. Über die Academic API war die Abfrage von bis zu 10 Millionen Tweets kostenfrei möglich. Dieser Zugang wurde im Juni 2023 abgeschaltet und durch kommerzielle API-Zugänge ersetzt, mit denen nun deutlich weniger Daten zu sehr hohen Preisen eingekauft werden können (siehe Tabelle 2.1). Forschenden aus dem EU-Raum wird zudem ein Bewerbungsformular[17] auf einen nicht näher beschriebenen Forschungsdatenzugang aufgrund des Digital Services Act (DSA, Artikel 40) angeboten. Erste Erfahrungen von Forscher*innen den Zugang zu beantragen wurden jedoch von X mit schwammigen Begründungen abgewiesen (Stokel-Walker 2024), so

[17] https://developer.x.com/en/use-cases/do-research, zuletzt geprüft am 09.09.2024.

dass zu befürchten ist, dass ein Forschungsdatenzugang über diesen Weg auch künftig nicht freiwillig eingeräumt werden wird.

2.2.4 Instagram

Anders als Facebook konzentriert sich das ebenfalls zu Meta gehörende Instagram-Netzwerk auf die öffentliche Kommunikation mittels Foto- und Videoinhalten und richtet sich im Durchschnitt an ein viel jüngeres Publikum als die beiden vorherigen Plattformen. Mit ca. 30 Millionen Konten ist es eine der beliebtesten Plattformen mit dem größten Anteil an Nutzer*innen zwischen 25 und 34 Jahren (Koch 2022). Die Plattform ermöglicht für öffentliche Profile das Hochladen von Bildern, die kommentiert werden können. Die Nutzer*innen können anderen Nutzenden folgen, um deren Profil zu verfolgen. In den letzten Jahren hat die ‚Story'-Funktion, mit der dauerhafte und flüchtige Inhalte veröffentlicht werden können, an Popularität gewonnen. Außerdem wurde Instagram als Plattform für politische Kommunikation immer wichtiger und wie bei Facebook erhalten (rechts-)populistische politische Inhalte einen unverhältnismäßig hohen Anteil an Nutzer*inneninteraktion im Vergleich zu nicht-populistischen politischen Inhalten (Larsson 2021). Aufgrund des Charakters der primär visuellen Kommunikation und der Schwierigkeiten bei der Erhebung großer Datensätze ist die Erforschung von Instagram-Inhalten sehr viel komplexer. Mit der MCL bietet Meta inzwischen jedoch einen niedrigschwelligen Forschungsdatenzugang an, mit dem Posts und Kommentare gesammelt und ausgewertet werden können (siehe Abschnitt Facebook).

2.2.5 YouTube

Auf YouTube können Nutzer*innen Videos öffentlich hochladen und kommentieren. Über Kanäle verteilen YouTuber*innen Inhalte mit konzeptuell ausgearbeiteten Formaten in verschiedenen Sparten wie Comedy, politischer Kommentar oder Gaming an ihre Abonnent*innen und können so mehrere

hunderttausend bis Millionen Views erreichen. Über 36 % der Deutschen geben an, YouTube beinah täglich zu benutzen.[18] Über eine Forschungs-API bietet YouTube Zugriff auf Metadaten zu den Videos sowie auf automatisch bzw. in Einzelfällen auch manuell erstellte Transkriptionen der Audiospuren an. Damit können Videos in Bezug auf die nicht-visuellen Informationsanteile hin untersucht werden. Zusätzlich können die Debatten untersucht werden, bei denen Nutzer*innen Videos zunächst kommentieren und andere Nutzer*innen Antworten auf die Kommentare verfassen können. In einer Studie zu Wissenschaftskanälen auf YouTube beobachten Dubovi und Tabak (2020), dass sich in Nutzerdebatten durchaus Muster gemeinsamer deliberativer Wissensproduktion finden, die über ein bloßes Mitteilen von Information hinausgehen. In diesem Sinne können Nutzerdebatten zusätzlich zu thematisch einschlägigen Videos einen interessanten Forschungsgegenstand für die TA darstellen.

2.2.6 TikTok

TikTok ist eine Social-Media-Plattform, auf der Nutzer*innen kurze Videos von sich selbst oder anderen teilen können. Die meisten Inhalte auf der Plattform sind Musikvideos, Comedy-Sketche und Lifestyle-Inhalte, zunehmend aber auch politische Inhalte. TikTok hat weltweit über 1 Milliarde monatliche aktive Nutzer*innen. In Deutschland sind es rund 12 Millionen (Koch 2022). Die Nutzerschaft von TikTok ist hauptsächlich jung, mit einem großen Anteil an Nutzer*innen im Alter von 16 bis 24 Jahren. Seit Herbst 2023 bietet TikTok in Europa einen Forschungsdatenzugang über eine API an, mit der Account-Daten, Video-Metadaten und Kommentare angefragt werden können. Wie nützlich TikTok-Daten für die Technikfolgenschätzung sind, lässt sich derzeit schwer abschätzen. Die sehr junge Zielgruppe ermöglicht einen Einblick in Einstellungen und Nutzungsweisen der heranwachsenden Generation. Der Fokus auf Entertainment lässt jedoch vermuten, dass für die TA relevante

[18] https://de.statista.com/statistik/daten/studie/777782/umfrage/tagesreichweite-von-videoportalen-in-deutschland, zuletzt geprüft am 06.06.2024.

Inhalte auf der Plattform deutlich unterrepräsentiert sind. Zunehmend erschließen jedoch auf bestimmte Themenbereiche wie Medizin oder technische Innovationen spezialisierte Influencer größere Publikumskreise auf der Plattform.[19]

2.2.7 Telegram

Der Messenger-Dienst Telegram hat seine Nutzerbasis in Deutschland in den letzten Jahren stetig ausgebaut (Statista berichtet von ca. 8 Millionen Nutzer*innen im Jahr 2019).[20] Aufgrund der besonderen Funktionen, die Telegram seinen Nutzenden bietet, spielt es eine zunehmend wichtige Rolle bei der Gestaltung politischer und gesellschaftlicher Diskurse (Wiedemann et al. 2023). Telegram erlaubt öffentliche Chatgruppen mit bis zu 200.000 Mitgliedern und ‚Channels' ohne Teilnehmerbegrenzung. Das Unternehmen, das hinter Telegram steht, behauptet, sichere Kommunikationstools anzubieten, um die Privatsphäre seiner Nutzer*innen zu schützen und behördlichen Forderungen nach Zensur der veröffentlichten Inhalte zu widerstehen. Diese Eigenschaften machten Telegram zu einem attraktiven Kommunikationsort für (rechtsextreme) Populisten und Anhänger von Verschwörungstheorien, insbesondere seit dem Ausbruch der Corona-Pandemie (Boberg et al. 2020). Für die TA wäre die Kommunikation in diesen Bereichen ggf. für Studien von Interesse, die mehr über die Rezeption, insbesondere über Vorurteile und Ängste gegenüber neuen Technologien wie Impfstoffen oder Mobilfunkstandards in gesellschaftlichen Randbereichen erfahren wollen. Auch wenn die in Telegramgruppen gepflegten Verschwörungsmythen bisweilen skurril anmuten, steht zu befürchten, dass gerade in bestimmten Regionen mittlerweile Ausstrahlungskraft in nennenswerte Teile der Bevölkerung hinein entfalten (Kiess et al. 2022; Wiedemann et al. 2023).

[19] https://www.deutschlandfunkkultur.de/medfluencer-krankheiten-soziale-medien-100.html, zuletzt geprüft am 06.06.2024.

[20] https://de.statista.com/infografik/26390/umfrage-zur-nutzung-des-telegram-messengers-in-deutschland, zuletzt geprüft am 06.06.2024.

2.2.8 Reddit

Reddit ist eine Social-News-Plattform, auf der Nutzer*innen Inhalte in Form von Links, Texten und Bildern teilen und diskutieren können. Die Plattform ist in verschiedene sogenannte Subreddits unterteilt, die sich mit bestimmten Themen beschäftigen wie Technologie[21], Politik, Unterhaltung und Sport. Die meisten Inhalte auf Reddit sind Textinhalte, aber es gibt auch viele Bilder und Links zu Videos. Die in Subreddits geposteten Inhalte werden in den jeweiligen Communities unterschiedlich stark diskutiert. Up-/Down-Votes von Usern für Posts und für Kommentare sorgen für eine Sortierung der Inhalte, bei der fundierte Sachbeiträge und hilfreiche Meinungsäußerungen besonders sichtbar werden. Laut einer Schätzung von 2021 hat Reddit weltweit über 430 Millionen monatliche aktive Nutzer*innen.[22] In Deutschland sind es rund 3 Millionen (Koch 2022). Die Nutzenden von Reddit ist eher männlich und tendenziell jünger, viele Nutzer*innen sind im Alter von 18 bis 29 Jahren. Die Aufteilung in thematische Subreddits und die darin aktiv bewertenden und diskutierenden Communities machen Reddit-Daten interessant für wissenschaftliche Forschungszwecke. Für die TA von Interesse wären beispielsweise Einblicke in Trends und Meinungen von Nutzer*innen zu bestimmten Technologien und Einschätzungen zu deren Auswirkungen auf die Gesellschaft. Da im Jahr 2023 der kostenfreie Zugang zur Reddit API deutlich eingeschränkt wurde, sind Forschende zwischenzeitlich zum von Scraping (siehe Abschnitt 2.3) von Inhalten gezwungen, wenn sie aktuelle Debatten beforschen möchten.[23]

2.2.9 LinkedIn

LinkedIn ist eine Business-orientierte Social-Media-Plattform, die es Nutzenden ermöglicht, ihre beruflichen Kontakte zu verwalten, zu erweitern und zu

[21] Z. B. https://www.reddit.com/r/ArtificialInteligence/; https://www.reddit.com/r/OpenAI/, zuletzt geprüft am 09.09.2024

[22] https://www.statista.com/topics/5672/reddit, zuletzt geprüft am 06.06.2024.

[23] Im Mai 2024 kündigt ein Vertreter des Reddit CTO-Teams unter r/reddit4researchers (zuletzt geprüft am 06.06.2024) einen Forschungsdatenzugang in absehbarer Zeit an.

nutzen. Die Plattform bietet Funktionen wie die Möglichkeit, ein professionelles Profil zu erstellen, sich mit anderen Fachleuten zu vernetzen, Jobangebote zu suchen und zu teilen und an Online-Kursen und Veranstaltungen teilzunehmen. Zunehmend nutzen Expert*innen die Plattform für ihre professionelle Selbstdarstellung, in dem sie Blog-Artikel und Diskussionsbeiträge zu Fachthemen verfassen. Laut Schätzungen von 2021 hat LinkedIn weltweit über 740 Millionen registrierte Nutzer*innen. In Deutschland sind es rund 20 Millionen. Die Nutzerschaft von LinkedIn ist hauptsächlich erwachsen, mit einem großen Anteil an Nutzenden im Alter von 25 bis 44 Jahren und einen höheren Anteil von männlichen Nutzer*innen. Die zunehmende Bedeutung der Plattform für professionelle Kommunikationskontexte zeigt sich beispielsweise in der Tatsache, dass 46 % aller Bundestagsabgeordneten 2022 ein LinkedIn-Profil besitzen.[24] Seit August 2023 stellt LinkedIn im Zuge des DAS Forscher*innen aus der EU auf Antrag einen Forschungsdatenzugang bereit.[25]

2.2.10 Mastodon, BlueSky und Threads

Die Veränderungen auf Twitter/X im Nachgang der Übernahme durch Elon Musk haben zu einem Aufschwung bei den Nutzer*innenzahlen und Aktivitäten anderer Kurztext-Plattformen geführt. Mastodon hat als soziales Netzwerk in der zweiten Jahreshälfte 2022 größere Aufmerksamkeit als offene und freie Alternative auf sich gezogen. Mastodon ist ein dezentrales Netzwerk von Servern, die über das ActivityPub-Protokoll miteinander kommunizieren. Die dezentrale Architektur ermöglicht es, dass auf unterschiedlichen Instanzen unterschiedliche Community-Regeln und Moderationsstandards gelten sowie User nicht ohne weiteres ausgesperrt werden können. Dadurch fehlt allerdings auch ein netzwerkweit operierender Empfehlungsalgorithmus, der potenziell ‚virale' Posts prominent in den Newsfeeds der Nutzer*innen platziert. Stattdessen erhält man nur die Posts der Accounts, denen man aktiv folgt, in chronologischer Reihenfolge dargestellt. Damit passiert es häufig, dass User*innen

[24] https://de.linkedin.com/pulse/wie-linkedin-ist-der-deutsche-bundestag-martin-fuchs, zuletzt geprüft am 06.06.2024.

[25] https://www.linkedin.com/help/linkedin/answer/a1645616, zuletzt geprüft am 06.06.2024.

für die interessante Informationen verpassen. Nach anfänglich großem Interesse in wissenschaftlichen und politisch aktiven Communities auf Twitter/X zu Mastodon umzuziehen, sind die Aktivitäten dort jüngst wieder enorm zurückgegangen. Nach 2,5 Millionen Nutzer*innen im Dezember 2022 loggten sich im Januar nur noch 1,8 Millionen ein (Nicholas 2023). Jedoch haben sich vor allem wissenschaftliche Communities in dieser Zeit erstaunlich schnell entlang ihrer (teil-)disziplinären Interessen vernetzt, so dass womöglich auch künftig bestimmte Expert*innendiskurse auf der Plattform geführt werden, deren Beobachtung für die TA interessant sein könnte. Eine vielversprechende die 2023 viel Aufmerksamkeit auf sich zog, ist BlueSky. Die schnell wachsende Community (5,6 Mio. Nutzer*innen weltweit, davon ca. 225.000 aus Deutschland[26]) tauscht sich über ähnliche Themen aus wie (zuvor) beim Wettbewerber Twitter/X, ist aber noch deutlich weniger von Aktivitäten von Bots und toxischen User*innen, aber auch von deutlich weniger prominenten Accounts im Vergleich zu Twitter/X geprägt. Über eine offene API-Schnittstelle können Post-Inhalte, Konversationen und Nutzerprofile für Forschungszwecke abgefragt werden. Die Kurztext-Plattform Threads vom Meta-Konzern ist seit Dezember 2023 in Europa verfügbar und zieht seitdem mehr und mehr Nutzer*innen an. Durch seine Empfehlungsalgorithmen versucht Meta politische, bzw. politisch kontroverse Inhalte von der Plattform fernzuhalten und mehr das Unterhaltungssegment zu bedienen. Der Konzern kündigte zudem eine Schnittstelle zum ActivityPub Protokoll von Mastodon an, so dass eine Integration mit dem dezentralen Fediverse, ein Verbund verschiedener sozialer Netzwerke über ein gemeinsames Protokoll, möglich erscheint. Dies würde auch einen vereinfachten Forschungsdatenzugang ermöglichen. Wie sich aber die Inhalte und die Nutzerzahlen entwickelt und welche Potenziale für die TA in den neuen, (noch) kleinen Plattformen steckt, wird sich erst in den nächsten Jahren zeigen.

[26] https://www.namepepper.com/bluesky-statistics, zuletzt geprüft am 06.06.2024.

2.2.11 (Nachrichten-)Webseiten

Zur Beobachtung medialer Diskurse werden in der CSS-Daten häufig direkt von Webseiten der Informationsanbieter automatisiert und in großer Zahl erhoben (siehe Abschnitt 2.3.2). Von Online-Medien lassen sich in der Regel Titel, Teaser und Datum der erscheinenden Artikel frei erfassen. Für Artikel, die nicht hinter einer Bezahlschranke verborgen sind, lassen sich auch Volltexte herunterladen. Für systematische und vollständige Erhebungen von Artikeln einzelner Publikationsorgane zu bestimmten Themen wird häufig auf Zeitschriftendatenbanken zurückgegriffen. Anbieter wie LexisNexis[27], Genios[28] und Wiso[29] bieten im deutschsprachigen Raum gegen Bezahlung Zugriff auf die Volltexte fast aller Zeitschriften sowie der regionalen und überregionalen Tagespresse. Eine kleine Auswahl von Zeitungskorpora über mehrere Jahrgänge ist kostenlos über das Projekt „Digitaler Wortschatz der deutschen Sprache“ (DWDS) verfügbar.[30]

Je nach Forschungsinteresse werden auch große Mengen öffentlicher Texte auch bei anderen Webseiten wie Pressemitteilungen von Unternehmen, Regierungsstellen oder NGOs erhoben. Beispielsweise analysiert Grimmer (2010) die Themenagenda in mehr als 24.000 Pressemitteilungen von US-Senator*innen. Pfetsch et al. (2016) erheben und analysieren Hyperlink-Netzwerke von Webseiten zum Thema Lebensmittelsicherheit und vergleichen die Akteurs- und Link-Strukturen in vier Ländern miteinander. Für die TA interessante Daten könnten beispielsweise auch in den Artikeln und Nutzerkommentaren thematisch einschlägiger Blogs, Foren und Petitionsseiten[31] erhoben werden. Die Recherche und das Erheben von Webseitendaten aus unterschiedlichen Quellen sowie deren Vereinheitlichung für eine folgende Analyse stellen oft

[27] https://www.lexisnexis.de, zuletzt geprüft am 06.06.2024.
[28] https://www.genios.de, zuletzt geprüft am 06.06.2024.
[29] https://www.wiso-net.de, zuletzt geprüft am 06.06.2024.
[30] https://www.dwds.de/d/k-zeitung, zuletzt geprüft am 06.06.2024.
[31] Auf E-Petitionen des Deutschen Bundestages oder Seiten wie change.org diskutieren tausende Bürger*innen aktuelle Themen und Forderungen an die Politik.

einen erheblichen Arbeitsaufwand dar, da Grundgesamtheiten für eine Forschungsfrage relevanter Webseiten in der Regel nicht bekannt sind und deren heterogene technische Realisierung meist individuell angepasste Werkzeuge benötigt.

2.2.12 (Online-)Umfragen und Tracking

Daten aus Online-Umfragen liefern in der CSS-Forschung zunehmend ergänzend Daten zu nicht-reaktiven Messungen digitaler Spurendaten, die beispielsweise aus Smartphone Apps oder Internetbrowsern stammen. Die Daten gelangen in der Regel über freiwillige ‚Datenspenden' an die Forschenden. Stier et al. (2022) werten Daten aus der Browser-History von über 7.000 Befragungsteilnehmer*innen in sieben Ländern aus, um zu untersuchen, wie sich algorithmische Recommender-Systeme auf den Nachrichtenkonsum von sozialen Mediennutzer*innen auswirken. Dabei geht es weniger darum, wie in der klassischen Survey-Forschung Aussagen mit Repräsentativität in Bezug auf eine bestimmte Grundgesamtheit zu erzeugen, sondern Einstellungen und Kenntnisstände in Bezug zu anderen Variablen zu setzen, die beispielsweise aus der Beobachtung von Verhalten in digitalen Umwelten stammen. Die Kombination verschiedener Datenquellen sowie rechtlichen und ethischer Anforderungen an die Datenerhebung und -auswertung machen solche kombinierten Forschungsdesigns besonders komplex. Zu besseren Orientierung und Planung solcher Studien schlagen Boeschoten et al. (2022) ein methodisches „Total Error Framework" für ein datenschutzgerechtes Arbeiten mit Datenspenden vor. Silber et al. (2022) messen beispielsweise, welchen Einfluss bestimmte Faktoren wie demografische Daten oder Anreizstrukturen auf die Entscheidung von Umfrageteilnehmer*innen haben, ihre digitalen Spurendaten aus der Nutzung von Smartphone Apps oder sozialen Medien für die Forschung zur Verfügung zu stellen.

Jenseits von digitalen Spurendaten kommen Online-Umfragen auch in Experimentaldesigns zum Einsatz. Hierfür werden Teilnehmer*innen in Gruppen eingeteilt und mit systematisch variierenden Inhalten bzw. Aufgaben konfrontiert und anschließend befragt. Mit Hilfe statistischer Modelle lassen sich anschlie-

ßend systematische Zusammenhänge zwischen unterschiedlichen ‚Treatments' und resultierenden Antworten aufdecken. Roberts et al. (2014) stellen mit ihrem Structural Topic Model einen interessanten Ansatz für die Auswertung offener Fragen.

2.2.13 Open (Government) Data

Zunehmend wichtiger in der CSS-Forschung werden offene Datenbestände, die in der Regel über Repositorien oder Schnittstellen von regierungsamtlichen Stellen oder Unternehmen zur Verfügung gestellt werden. Als offen gelten Daten zu denen jede*r freien, unentgeltlichen Zugriff hat und die zu jedem Zweck genutzt, bearbeitet und weitergegeben werden dürfen.[32]. Als einschränkende Bedingungen sind lediglich Urhebernennung und die Gewährleistung von Offenheit bei Weitergabe (‚Share-alike') zugelassen, die über spezifische Lizenzen wie die „Datalizenz Deutschland – Zero 2.0"[33] festgeschrieben werden können. Um offen bereitgestellt werden zu können, dürfen Regierungs- und Verwaltungsdaten nicht personenbezogen oder sicherheitsrelevant sein sowie keine Betriebs- und Geschäftsgeheimnisse enthalten. Zentrale Suchportale wie GovData[34] erfassen die dezentral von Ämtern, Ministerien und Firmen bereitgestellten Informationen über spezialisierte Metadatenformate und erlauben so die Recherche zehntausenden von Datensätzen. Für Forschungsdaten, insbesondere im Bereich Sozialwissenschaften, ist das Kriterium der Offenheit aufgrund rechtlicher und forschungspraktischer Einschränkungen vielfach nicht ohne Weiteres umsetzbar. Hier hat sich stattdessen das FAIR-Prinzip etabliert (Wilkinson et al. 2016): Daten sollten demnach auffindbar (Findable), zugänglich (Accessible), interoperabel und maschinenlesbar (Interoperable) und nachnutzbar (Reusable) verfügbar werden. Die Zugänglichkeit darf dabei begründet auf bestimmte Kreise, zum Beispiel Forschende, die eine bestimmte Nutzungsvereinbarung unterzeichnen, eingeschränkt sein. Analog zu Open

[32] https://opendefinition.org/od/2.1/de, zuletzt geprüft am 06.06.2024.
[33] https://www.govdata.de/dl-de/zero-2-0, zuletzt geprüft am 06.06.2024.
[34] https://www.govdata.de, zuletzt geprüft am 06.06.2024.

Government Data gibt es für die Recherche von Forschungsdaten zentrale Repositorien[35] und Such-Indizes[36].

Für die TA-Forschung könnten Open Data Quellen in vielfältiger Hinsicht von Interesse sein. Eine Denkrichtung wäre, den Ausbau und die Nutzung öffentlicher Infrastrukturen[37] zu beobachten und ggf. mit anderen Daten wie Einstellungen aus Survey- oder Social-Media-Daten zu kontrastieren. Zur Beobachtung TA-relevanter, politischer Diskurse bieten sich Parlamentsdokumentationen unterschiedlicher administrativer Gliederungen an. Beispielsweise stellen Rauh und Schwalbach (2020) mit dem ParlSpeech V2 dataset ein Korpus aus über 6 Mio. Parlamentsreden aus 9 demokratischen Ländern zur Verfügung. Ein vollständig aufbereitetes Korpus der Plenardebatten des Deutschen Bundestages von 1949 bis 2021 wird bereitgestellt von Blaette und Leonhardt (2022). Zusätzlich zu Plenardebatten lassen sich andere Arten von Drucksachen des Bundestages über dessen Dokumentations- und Informationssystem über offene Schnittstellen beziehen.[38]

2.3 Datenerhebung

Für die CSS-Forschung werden in der Regel sehr große Mengen digitaler Daten erhoben, deren Potenziale und Herausforderungen unter dem Begriff ‚Big Data' diskutiert werden (Mahrt und Scharkow 2014). Der folgende Abschnitt beschreibt übliche Wege, an diese Daten zu gelangen.

[35] Zum Beispiel das CERN betriebene Portal https://zenodo.org, zuletzt geprüft am 06.06.2024.

[36] Zum Beispiel das EU-finanzierte Portal OpenAIRE Explore https://explore.openaire.eu, zuletzt geprüft am 06.06.2024.

[37] Beispiele wären Orte und Verfügbarkeit von E-Auto Ladestationen in Deutschland (https://www.govdata.de/web/guest/suchen/-/details/elektro-ladestationen-in-deutschland) oder durchschnittliche Zugverspätungen nach Stadt/Region aus dem API Marketplace der Deutschen Bahn (https://developers.deutschebahn.com/db-api-marketplace/apis/product/timetables, zuletzt geprüft am 06.06.2024.

[38] https://dip.bundestag.de/\"uber-dip/hilfe/api, zuletzt geprüft am 06.06.2024.

2.3.1 APIs

Der direkteste Weg, digitale Forschungsdaten zu erheben, ist die Abfrage über dedizierte Schnittstellen bei einem Datenanbieter. An zu diesem Zweck bereit gestellte Application Programming Interfaces (API), kann innerhalb eines Softwareprogramms eine Anfrage für einen Datensatz, der bestimmten Suchkriterien erfüllt, gesendet werden.[39] Die API antwortet dem Programm mit einer Liste der gefundenen Datensätze in einem bestimmten Format, das dann vom anfragenden Programm weiterverarbeitet werden kann.[40] Jünger (2021) stellt in seiner ‚kurzen Geschichte der APIs' drei Phasen der Entwicklung dar. Nach Aufbau und Eroberung von Daten-APIs insbesondere durch die großen digitalen Player wie Facebook und Google befinden wir uns seit dem Jahr 2015 in einer Phase der Besorgnis und Abschottung. Skandale wie der massenhafte Datenmissbrauch zur Wahlbeeinflussung auf Facebook durch die Firma Cambridge Analytica (Venturini und Rogers 2019) haben dazu geführt, dass seitdem Datenzugänge stark eingeschränkt und damit auch die Erhebung von Forschungsdaten insbesondere bei Sozialen Medienplattformen erschwert wurde (Bruns 2019). Spezielle Forschungsdatenzugänge existieren derzeit für Twitter/X[41], TikTok[42] und YouTube. Einen erfreulich offenen Zugang zu seinen redaktionellen Inhalten erlaubt die britische Tageszeitung ‚The Guardian'. Über eine API lassen sich alle in Print und auf der Webseite erschienen Artikel

[39] Anfragen werden typischerweise als Aufruf einer URL per HTTP GET oder POST-Request an eine REST-Schnittstelle gestellt. Eine API-Dokumentation beschreibt, welche Anfragen an welche URLs mit welchen Parametern möglich sind und wie die zugehörigen Antworten aussehen. Ein Beispiel für eine solche Dokumentation liefert die API-Dokumentation des Bundestagsinformationssystems unter https://search.dip.bundestag.de/api/v1, zuletzt geprüft am 06.06.2024.

[40] Typischerweise werden Antworten im semi-strukturierten JSON- oder XML-Format übermittelt, die von entsprechenden Standardbibliotheken in gängigen Programmiersprachen leicht weiterverarbeitet werden können. Seltener werden als tabellarische CSV-Formate ausgegeben.

[41] Seit der Übernahme des Dienstes durch Elon Musk im zweiten Halbjahr 2022 besteht große Unsicherheit über die Fortführung des Forschungsdatenzugangs. Im Februar 2023 wurde eine Bezahlschranke für alle APIs angekündigt.

[42] In der Startphase nur für Forscher*innen in den USA und mit fragwürdigen Nutzungsbedingungen.

im Volltext beziehen.[43] Um die Möglichkeit komplette Kopien von Datenbanken zu erzeugen einzuschränken, werden APIs in der Regel mit einem Authentifikationssystem und einem Rate Limit geschützt, welches die Datenmenge, die ein einzelner API-Zugang beziehen darf, einschränkt.

2.3.2 Data Crawling und Scraping

Können oder sollen öffentlich verfügbare Daten von Anbietern nicht direkt über eine API abgefragt werden, können Techniken des Crawling und des Scraping zum Einsatz kommen (Munzert et al. 2015). Crawling bezeichnet die automatische Identifikation von Verlinkungsstrukturen innerhalb einer oder zwischen mehreren Webseiten, die (potenziell) relevante Inhalte enthalten. Ein typisches Vorgehen ist das sogenannte Snowball-Sampling bei dem ausgehend von einer Seed-Liste relevanter Seiten, alle von dort aus verlinkten Seiten besucht und ihrerseits nach Relevanz bewertet werden (Pfetsch et al. 2016). Als Relevanzkriterium kann beispielsweise das Vorhandensein bestimmter Schlüsselbegriffe auf der Webseite automatisch getestet werden. Dieser Vorgang wird für die als relevant identifizierten Seiten so lange wiederholt, bis eine gewisse Sättigung erreicht ist, also kaum noch neue relevante Seiten gefunden werden, oder die Stichprobe eine gewisse Größe überschritten hat, die im Rahmen der Forschungsarbeit als bewältigbar erscheint. Scraping bezeichnet die automatisierte Auswahl und das Herunterladen gezielter Teile einer Webseite wie beispielsweise die Werte aus einer Tabelle oder der Volltextartikel einer Online-Nachrichtenseite. Weggelassen werden sollten dabei in der Regel Menü-, Header und Footer-Elemente einer Webseite sowie eventuelle Werbeanzeigen (boilerplate removal). Die Auswahl relevanter Textteile wird dabei in der Regel anhand von Struktur-Informationen des HTML-Quelltextes[44] einer Webseite vorgenommen. Darin lassen sich über Anfragesprachen wie XPATH[45] gezielt bestimmte Abschnitte einer Webseite selektieren und

[43] Zum Beispiel liefert die folgende Anfrage alle Artikel, die den Bundeskanzler Olaf Scholz am 01. Januar 2023 erwähnen: https://content.guardianapis.com/search?q=Scholz&format=json&from-date=2023-01-22&api-key=test, zuletzt geprüft am 06.06.2024.

[44] https://www.w3schools.com/html, zuletzt geprüft am 06.06.2024.

[45] https://www.w3schools.com/xml/xml_xpath.asp, zuletzt geprüft am 06.06.2024.

darin enthaltene Texte und Attribut-Informationen extrahieren. Da Webseitenquelltexte in ihrem Aufbau nur wenig inhaltlichen Standards folgen,[46] ist das Scraping ausgewählter Inhalte häufig ein arbeitsintensiver Prozess, der individuelle Anpassungen an die jeweils zu analysierende Webseite bedarf. Tools wie Trafilatura (Barbaresi 2021) ermöglichen durch den Einsatz zahlreicher Heuristiken eine automatische Extraktion relevanter Inhalte mit für viele Fälle ausreichender Qualität.

Da viele Inhalteanbieter das Scraping ihrer Daten einerseits für Suchmaschinen ermöglichen und andererseits ein Kopieren ihrer Daten durch automatische Programme verhindern wollen, wird das Scraping zunehmend durch technische Einrichtungen eingeschränkt. Die jeweilige Richtlinie für das automatische Scrapen einer Website-Domain kann in der Datei robots.txt vermerkt werden und sollte auch durch Forschende bei der Datenerhebung beachtet werden. Mittlerweile erfordern viele Webseiten ein JavaScript-Rendering ihrer Inhalte sowie die aktive Zustimmung zu Verhaltensregeln auf der Webseite zum Beispiel durch Anklicken eines Buttons. In diesem Fall ist das das Scraping nur über eine (langsame) Browser-Fernsteuerung möglich, bei der ein Programmscript die Zielseite in einem speziellen Browser aufruft und Informationen aus der dort gerenderten Seite extrahiert.[47] Das Browser-Test-Tool Selenium WebDriver[48] ist hierbei der de-facto Standard. Plattformen wie Instagram können allerdings über solche Browserfernsteuerungen erzeugtes, unauthentisches Nutzerverhalten detektieren und sperren betreffende Accounts mit Hinweis auf Verstoß gegen ihre Nutzungsbedingungen. Je nach Forschungsfrage und dafür notwendige Datenquelle kann das Scraping somit erheblichen Aufwand zum Finden von Umwegen um die technischen Verhinderungsmaßnahmen herum erfordern oder gänzlich unmöglich sein.

[46] Wenn korrekt implementiert, können semantische HTML5-Tags wie *article* oder *section* die Selektion relevanter Dokumentteile erheblich erleichtern.

[47] Siehe https://tm4ss.github.io/docs/Tutorial_1_Web_scraping.html (zuletzt geprüft am 06.06.2024) für eine kurze Einführung in das Scraping mit dem Paket „rvest".

[48] https://www.selenium.dev/, zuletzt geprüft am 06.06.2024.

2.3.3 OCR und PDF-Datenextraktion

Liegen auszuwertende Daten teilweise noch nicht digitalisiert oder nur als Sammlungen von (gescannten) PDF-Dokumenten vor, kann es für kleinere Dokumentmengen sinnvoll sein, die Digitalisierung selbst vorzunehmen. Kommerzielle OCR-Programme wie ABBYY FineReader oder Adobe Acrobat wandeln Bilddokumente zuverlässig in maschinenlesbaren Text um. Gegebenenfalls muss erkannter Text nachbearbeitet werden, um etwa überflüssige Zeilenumbrüche zu entfernen oder Silbengetrennte Worte wieder zusammenzufügen. Hierfür bedarf es in der Regel individuell an die Daten angepasster Programmskripte, um eine optimale Aufbereitung für weitere Analysen zu gewährleisten. Als Open Source Alternative lässt sich die Software Tesseract OCR[49] in Programmscripte zur Texterkennung aus Bilddaten einbauen. Diese Option ist besonders interessant für die Extraktion von Text aus automatisch erhobenem Bildmaterial, das beispielsweise auf einer Social-Media-Plattform gepostet wurde. Auf diesen Plattformen ist die Einbindung von Text in Bilder, insbesondere im Zusammenhang mit zeitlich begrenzt verfügbaren Inhalten (bei Instagram und Facebook sog. ‚Stories') in den letzten Jahren besonders populär geworden. Einer automatischen Auswertung ist dieser Text allerdings nur per OCR zugänglich.

Die Extraktion von Text aus bereits digital vorliegenden Dateien im PDF und anderen Dokument-Formaten (DOCX, HTML etc.) wird unterstützt durch darauf spezialisierte Programmbibliotheken, die sich leicht in Auswertungsskripten verwenden lassen. Das wohl umfangreichste Werkzeug zum Auslesen von Daten aus beliebigen Dateiformaten ist Apache Tika.[50] Die für kollaborative investigative Recherche im journalistischen Bereich entwickelte Software LiquidInvestigations[51] nutzt Tika zur Prozessierung sehr großer, unbekannter Datenmengen aus zugespielten Leaks. Für kleinere und auf übliche Formate

[49] https://tesseract-ocr.github.io, zuletzt geprüft am 06.06.2024.
[50] https://tika.apache.org, zuletzt geprüft am 06.06.2024.
[51] https://github.com/liquidinvestigations/docs/wiki, zuletzt geprüft am 06.06.2024.

beschränkte Dokumentsammlungen gibt es beispielsweise Pakete wie readtext für die Programmiersprache R.[52]

2.3.4 Crowdsourcing / Online-Surveys

Crowdsourcing ist eine kostengünstige Möglichkeit, große Datenbestände mit zusätzlichen Informationen anzureichern und diese so für die CSS-Forschung zu erschließen. Beim Crowdsourcing werden große Aufgaben in kleinste Teilaufgaben zerteilt, die dann über eine Plattform wie Amazon Mechanical Turk[53] an einzelne Clickworker vermittelt und von diesen gegen Bezahlung ausgeführt werden. Typische Aufgaben sind das Zuordnen von Kategorien zu Textdaten entweder zur Erzeugung von Trainingsdaten für einen maschinellen Lernprozess oder zur Validierung von dessen Ergebnissen. Die Herausforderung bei dieser Art der Datenerhebung ist, die Aufgaben so zu formulieren, dass sie von verschiedenen Arbeiter*innen zuverlässig (reliabel) und korrekt (valide) ausgeführt werden können. Porter, Verdery und Gaddis (2020) machen Vorschläge für eine Best Practice Anleitung zum Einsatz von Crowdsourcing in der CSS-Forschung. Die Plattform eignet sich beispielsweise für die Durchführung von Umfragen. Weinberg, Freese und McElhattan (2014) untersuchen, inwieweit sich Plattformen, auf denen sich Teilnehmer gegen Bezahlung an Umfragen beteiligen sich als Alternative zu repräsentativen Bevölkerungsumfragen eignen. Sie kommen zu dem Ergebnis, dass die Rekrutierung für Umfragen via Crowdsourcing vergleichbar gute Ergebnisse liefert wie Repräsentativbefragungen, wenn die erhobenen Statistiken entlang demografischer Faktoren wie Alter und Geschlecht korrigiert werden. Zudem eignet sich Crowdsourcing auch für die Durchführung von Experimentalstudien. Horton, Rand und Zeckhauser (2011) ein Beispiel dafür, wie sich Mechanical Turk für verschiedene Experimentaldesigns im Bereich der Wirtschaftswissenschaft nutzen lässt. Beispielsweise testen sie grundlegende Hypothesen des Faches über menschliche Verhaltensmuster, in dem sie die Auswirkungen verschiedener Priming-Strategien auf das Antwortverhalten

[52] https://cran.r-project.org/web/packages/readtext/, zuletzt geprüft am 06.06.2024.
[53] https://www.mturk.com/, zuletzt geprüft am 06.06.2024.

von Befragten beobachten. Zallot et al. (2021) diskutieren zahlreiche Herausforderungen und Probleme beim Einsatz von Crowdsourcing in der Forschung. Sie kommen zu dem Schluss, dass Stichproben auf Basis von Crowdsourcing wahrscheinlich
weniger problematisch sind als vielfach übliche Stichproben aus Grundgesamtheiten von Studierenden oder unbekannte Grundgesamtheiten anonymer Online-Umfragen. Für die TA könnte per Crowdsourcing realisierte Experimentalstudien in Verbindung mit Online-Umfragen interessante Daten über die Nutzungsweisen oder die Bedingungen zur Akzeptanz neuer Technologien erhoben werden.

2.3.5 Datenspenden

Für Datenspenden werden typischerweise projektspezifische Softwaretools zur Datensammlung in einem technischen Endgerät auf der Nutzerseite (Aufzeichnungen eines Fitnesstrackers, Browserverläufe oder Smartphone App Nutzungsdaten) entwickelt, die mit Einwilligung der Nutzenden die Daten an die Forschenden übermitteln. Ein populäres Beispiel stellte die Smartphone App „Corona Datenspende“ des Robert Koch-Instituts dar,[54] mit der Daten von Fitnessarmbändern und Smartwatches an das RKI zur Beobachtung des Verlaufs der Corona-Pandemie gesendet werden konnten. Alternativ zur Sammlung mittels einer extra zu installierenden Software können auch Datenexporte von Accounts großer Social-Media-Plattformen gespendet werden. Die auf EU-Ebene eingeführte Datenschutzgrundverordnung verpflichtet Datenverarbeiter über die von ihnen gespeicherten Daten Auskunft zu geben. Anstelle Anfragen individuell beantworten zu müssen, räumen große Internetplattformen daher ihren Nutzer*innen häufig die Möglichkeit ein, ihre Daten direkt in ihrem Onlineprofile abzurufen. Mit OSD2F stellen Araujo et al. (2022) eine Software zur Verfügung, mit der Nutzer*innen diese Datenexporte mit Forschenden in datenschutzgerechter Weise teilen können.

[54] https://corona-datenspende.de, zuletzt geprüft am 06.06.2024.

3 Auswertungsmethoden

Das folgende Kapitel beschreibt Analysezugänge zu den verschiedenen Datenarten in der CSS-Forschung. Dabei wird ein Schwerpunkt auf die algorithmische Auswertung sehr großer Datenmengen gelegt. Etablierte Methoden der quantitativen Statistik, die in der CSS ebenfalls eine große Rolle spielen, werden dabei nur am Rande betrachtet. Allgemein wird in der CSS-Forschung häufig der für andere Wissenschaftsbereiche als selbstverständlich angesehene Hinweis gegeben, den Analysen eine konkrete und relevante Forschungsfrage oder Hypothese zum Forschungsgegenstand voranzustellen. Dass dies in der CSS besonders betont wird, liegt daran, dass die zugrundeliegenden großen und vielfach zunächst unbekannten Datenbestände häufig in den ersten Forschungsschritten weitgehend theorielos mit komplexen Methoden exploriert werden. Dabei zutage tretende, interessante Muster sind dann schon häufig als Ergebnis der Forschung präsentiert (Waldherr et al. 2021). Allerdings erlaubt diese Arbeitsweise nur schwer einen Anschluss an existierende Theorien und relevante Vorarbeiten. In diesem Sinne betonen Lazer et al. (2021) in ihrer Darstellung des CSS-Forschungsprozesses (vgl. Abbildung 3.1) die Rolle von wissenschaftlicher Motivation als Ausgangspunkt, welche die Datenerhebung beeinflussen sollte und nicht umgekehrt.[1] Die in der CSS-Forschung vorrangig ausgewerteten Phänomene können in drei Bereiche unterteilt werden, entlang derer sich jeweils ein weitläufiges Methodenspektrum zu deren Auswertung entfaltet: 1. Kommunikationsinhalte (Inhaltsanalyse), 2. soziale Relationen (Netzwerkanalyse) und 3. Eigenschaften von Akteuren (deskriptive Statistik, Inferenzstatistik).

[1] Tatsächlich verlief die CSS-Forschungspraxis bislang häufig umgekehrt. Je nach Verfügbarkeit von Daten wird überlegt, welche Fragen mit diesen beantwortet werden könnten.

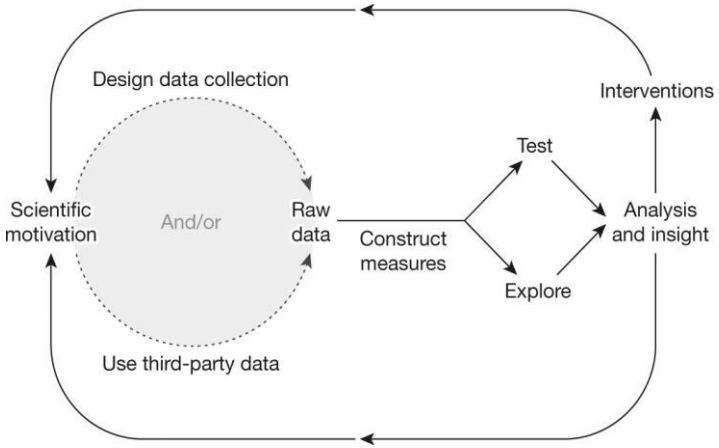

Abbildung 3.1: Empirische Messung dient als Brücke zwischen wissenschaftlicher Motivation und Erkenntnis im CSS-Forschungsprozess (Lazer et al. 2021).

3.1 Inhaltsanalyse

Für die von Kommunikationsinhalten sehr großer Datenmengen sind in den vergangenen Jahrzehnten immer komplexere automatische Verfahren entwickelt worden, die immer besser in der Lage sind Bedeutung zu erfassen (Wiedemann 2013). Von der einfachen Suche nach Zeichenketten und darauf aufbauenden Wortzählungen verlief die Entwicklung über latente semantische Modelle zur Messung von Dokumentinhalten hin zu heute gängigen Sprachmodellen auf Basis neuronaler Netze, die Wort-, Satz- und Dokumentbedeutung sowie auch Bildinhalte in hochdimensionalen Vektoren, sogenannten ‚embeddings', enkodieren. Wiedemann und Fedtke (2021) erläutern anhand neuer Verwendungskontexte des Begriffs ‚Goldstücke' in den sozialen Medien im Nachgang zur Flüchtlingskrise 2015, wie diese zunehmend komplexen Verfahren der automatischen Sprachverarbeitung semantische Bedeutung modellieren, Bedeutungswandel erfassen und damit schließlich eine Integration von quantitativen und qualitativen Analyseperspektiven ermöglichen. Einen aktuellen, deutschsprachigen Überblick über Verfahren des ‚text mining' geben Biemann, Heyer und Quasthoff (2022). Nicht mehr ganz aktuell, aber sehr gut

auf Artikellänge komprimiert, führen Grimmer und Stewart (2013) in die verschiedenen Verfahren der automatischen Inhaltsanalyse ein.

Um Textinhalte automatisch verarbeiten zu können, sind je nach Zustand der gesammelten Daten und eingesetzter Auswertungsmethode umfangreiche Vorverarbeitungsschritte notwendig. Übliche Vorverarbeitungsschritte, die leicht mit Hilfe spezialisierter Softwarebibliotheken wie quanteda (Benoit et al. 2018) oder udpipe (Wijffels 2022) in R oder nltk (Bird, Klein und Loper 2009) für Python umgesetzt werden können, sind:

- die Zerlegung von Texten in Sätze und Einzelworte (Tokenisierung), um Worte als Einzelereignisse in den Daten sichtbar zu machen sowie Messungen auf Satzebene zu ermöglichen.
- die Reduktion von Einzelworten in ihre Wortstämme (Stemming) oder Grundform (Lemmatisierung) sowie die Umwandlung in Kleinbuchstaben[2], um semantisch ähnliche Ereignisse in Textdaten zusammenzuführen.
- die Entfernung von sehr häufigen Funktionswörtern einer Sprache (Stoppworte) sowie von sehr seltenen Begriffen, die in einer statistischen Analyse keine wichtige Information beitragen.
- die Entfernung von Satzzeichen, Zahlen, Hashtag-Symbolen, URLs, Emojis oder redundanten Textteilen, die je nach Analysezweck nicht als bedeutungstragend angesehen werden.

Methoden der automatischen Inhaltsanalyse lassen sich in datengetriebene, Struktur-entdeckende und ‚überwachte', also Struktur-klassifizierende Verfahren unterscheiden. Erstere eignen sich für die Exploration eines unbekannten Datenbestandes. Mit letztere können Theorien operationalisiert und Hypothesen getestet werden. Als wichtigster Meilenstein der letzten Jahre gilt die Einführung von Sprachmodellen auf Basis der Transformer-Architektur. Das als initiale Variante von Google veröffentlichte BERT-Modell (Devlin et al. 2019)

[2] Dieser Schritt wird vor allem für englische Texte vorgenommen. In deutschen Texten kann bei einer entsprechenden Lemmatisierung darauf verzichtet werden.

kann mit Milliarden von öffentlich verfügbaren Sprachdaten vortrainiert werden, so dass ein umfangreiches Wissen über Sprache und semantische Bedeutungszusammenhänge in den Modellparametern abgebildet werden kann. Durch dieses Vortrainieren erreichen große Sprachmodelle dieser Architektur inzwischen menschliche Leistungen in vielen Aufgabentypen des Sprachverstehens oder übertreffen diese sogar. Mit dieser technischen Revolution gibt es für nahezu jede Aufgabe im Natural Language Processing eine Transformer-basierte Variante.

Eines der meistgenutzten datengetriebenen Verfahren ist die Extraktion von Schlüsselworten bzw. -phrasen, welche die wichtigsten inhaltlichen Aspekte eines Textes wiedergeben können.[3] Eine besondere Art von Schlüsselbegriffen sind Eigennamen, also Bezeichnungen für Personen, Organisationen, Orte, Währungen, Produkte etc. Mit Hilfe von Verfahren der Named Entity Recognition ist es möglich, Eigennamen aus Texten automatisch zu extrahieren. Mit Hilfe von Frequenzanalysen lassen sich anschließend Verwendungshäufigkeiten zum Beispiel von Begriffen oder Eigennamen über die Zeit oder zwischen Querschnitten des Datenkorpus vergleichen (Wiedemann und Fedtke 2021). Auf diese Weise lassen sich beispielsweise Begriffe finden, deren Verwendung über die Zeit signifikant zu oder abgenommen hat. Über einzelne Worte hinaus gehend erlaubt die Kookkurrenzanalyse die Beobachtung von Bedeutungsmustern anhand der statistischen Bewertung des Auftretens zweier Wortformen in einem wohldefinierten Textfenster. Zum Beispiel tauchen die beiden Begriffe ‚Bundestag' und ‚wählen' überzufällig häufig bezogen auf ihr Einzelvorkommen gemeinsam in einem Satz auf. Durch statistische Tests lassen sich solche Kookkurrenzen automatisch aus Texten extrahieren und zu ganzen Kookkurrenznetzwerken visualisiert werden, die der inhaltlichen Charakterisierung eines Dokumentkorpus dienen können. Maas (2016) veranschaulicht beispielsweise anhand von Kookkurrenznetzwerken zu Artikeln rund um die PISA-Studien in deutschen Nachrichtentexten die Veränderung des öffentlichen Diskurses zu diesem Thema. Auf Ebene der Analyse ganzer Dokumente ermöglicht das seit einigen Jahren in den Sozialwissenschaften etablierte Topic

[3] Eine auf BERT-basierende Implementierung findet sich hier: https://maartengr.github.io/KeyBERT/, zuletzt geprüft am 06.06.2024.

Modeling eine Aufteilung großer Dokumentmengen in überlappende, thematische Cluster (Maier et al. 2018). Dokumente sind dabei eine Wahrscheinlichkeitsverteilung über Themen und Themen eine Wahrscheinlichkeitsverteilung über Worte. Auf diese Weise werden bei der Modellinferenz vollkommen datengetrieben semantisch kohärente Wortcluster identifiziert, die sich häufig als Themen interpretieren lassen. Die so gefundenen Themen können zur Filterung des Dokumentkorpus, zum Finden ähnlicher Dokumente, zur Beobachtung ihrer Verbreitung über die Zeit oder zur Verknüpfung mit weiteren Dokumentvariablen wie Autorenschaft verwendet werden. Kieslich, Došenović und Marcinkowski (2022) führen beispielsweise mit Topic Modeling eine Themenanalyse der deutschen Medienberichterstattung über Künstliche Intelligenz durch und vergleichen dabei unter anderem überregionale und regionale Medien. Mit BERTopic (Grootendorst 2022) steht eine BERT-basierte Variante des Topic Modeling zur Verfügung, die nicht nur hervorragend interpretierbare Topics liefert, sondern durch den Einsatz multilingualer Transformermodelle eine sprachübergreifende Themenmodellierung ermöglicht.

Unter den überwachten Verfahren der automatischen Inhaltsanalyse ist die Diktionärsanalyse einer der weit verbreitetsten und am einfachsten zu handhabenden Ansätze. Diktionäre sind systematisch zusammengestellte Wortlisten, die eine oder mehrere semantische Kategorien repräsentieren sollen. Das Vorkommen von Diktionärsbegriffen in Zieldokumenten wird dann als Treffer für die entsprechende Kategorie gezählt. Auf diese Weise lassen sich abstrakte semantische Konzepte in großen Textmengen erfassen und für quantitative Auswertungen nutzbar machen. Problematisch am Diktionärsansatz ist allerdings, dass die Begriffslisten eine Kategorie nicht vollständig abbilden (Synonymie-Problem) und andererseits durch das Beobachten einzelner Worte ein je nach Kontext variierender Bedeutungsgehalt nicht erfasst werden kann (Polysemie-Problem). Das erste Problem kann teilweise durch neue Verfahren des Word Embeddings gelöst werden, indem Worte nicht mehr als Zeichenfolgen, sondern deren Bedeutungsrepräsentation in einem vortrainerten Vektorraum beobachtet werden. Auf diese Weise können semantisch ähnliche Begriffe in der

Quantifizierung von Kategorien berücksichtigt werden.[4] Abstrakte Konzepte oder komplexere inhaltliche Kategorien wie Argumentstrukturen oder politische Forderungen können jedoch mit Diktionären nur schwer erfasst werden. Hierfür bieten sich Verfahren der maschinellen Textklassifikation an, die auf Basis manuell kodierter Trainingsdaten lernen, eine Zielkategorie in neuen Daten wiederzufinden. Ein häufig genutztes Verfahren der Textklassifikation ist die Sentiment-Analyse, bei der es darum geht, eine positive, neutrale oder negative Haltung in Texten automatisch zu erkennen. Für deutsche Texte haben Guhr et al. (2020) ein BERT-basiertes Modell veröffentlicht, dass auf über 5 Millionen Beispielen trainiert wurde. Als Eingabe fungiert hier jeweils ein ganzer Satz bzw. kurzer Text, dessen sequenziell gelesene Bedeutung als Ganzes interpretiert wird. So können beispielsweise auch Verneinungen bzw. sprachlich markierte Umkehrungen von Stimmungen durch das Modell erkannt und korrekt klassifiziert werden. Für spezielle Forschungsfragen müssen in der Regel neue Kategoriesysteme festgelegt, dafür passende Trainingsdaten erzeugt und anschließend neue Modelle damit trainiert und evaluiert werden. Fedtke und Wiedemann (2020) nutzen beispielsweise ein selbst trainiertes Klassifikationsmodell zur Einteilung von Facebook-Kommentaren in Hassrede, Gegenrede und neutrale Posts mit dem anschließend Dynamiken von Nutzerdebatten analysiert werden können.

Am Anfang einer automatischen Inhaltsanalyse steht also die Auswahl der geeignetsten Verfahren für die Bearbeitung der Forschungsfrage. Hierbei können vielfältige Ansätze zu komplexen Forschungsdesigns verbunden werden. Wiedemann (2016) kombiniert beispielsweise Diktionärs-basiertes Filtern, Schlüsselwortextraktion, Kookkurrenzanalyse, Topic Modeling und Textklassifikation zu einem komplexen Forschungsdesign, mit dem ein thematisch äußerst breiter Diskurs rund um Fragen der Demokratie in deutscher Nachrichtentexten ausgewertet wird. Die Herausforderung beim Einsatz einzelner Verfahren wie auch in deren Kombination besteht in der Definition der jeweils am besten geeigneten Vorverarbeitung der Texte, der Auswahl geeigneter

[4] Das R-Paket dictvectoR (Thiele 2022) bietet eine einfach zu nutzende Implementierung dieses Verfahrens an.

Parameter bei der eigentlichen Analysemethode, der systematische Interpretation der Ergebnisse sowie der gründlichen Validierung der darauf aufbauenden Erkenntnisse. Diese erfolgt in der Regel durch einen qualitativen Blick in einen Ausschnitt der Daten.

3.2 Exkurs ‚large language models'

Während der Erstellung dieser Überblicksstudie ist im Jahr 2023 eine KI-Revolution bislang ungekannten Ausmaßes im Bereich der Computerlinguistik und den mit dieser Disziplin verbundenen Anwendungsbereichen wie den CSS in Gang gekommen. Zum Jahresende 2022 veröffentlichte die Firma OpenAI die Anwendung ChatGPT, welche den Nutzer*innen über ein Chat-Interface Zugriff auf ein mit riesigen Datenmengen trainiertes großes Sprachmodell (large language model, LLM) erlaubt (OpenAI und Achiam et al. 2024). Optimiert für eine möglichst natürliche und nutzerorientierte, dialogische Interaktion erlaubt das KI-Modell die Beantwortung von Experten-Fragen und Aufgaben in allen möglichen Bereichen (z. B. Medizin, Biologie, Chemie, Recht), maschinelle Übersetzung und die Unterstützung bei der Entwicklung von Computerprogrammen. Ähnliche Funktionalitäten bieten OpenAIs Wettbewerber wie Googles Gemini-Modell oder Meta LLaMa Modell-Serie sowie eine Reihe von Open-Source Modellen kleinerer Unternehmen wie Mistral AI. Ziems et al. (2024) bescheinigen der Technologie transformatives Potenzial für die CSS, insbesondere für die automatische Kodierung von Text- und Bild-Inhalten. So zeigen erste Studien, dass LLMs ähnlich gut in der Lage sind Texte zu kategorisieren wie menschliche Kodierer, wenn ihnen nur die Kategorie-Beschreibungen als Code-Buch zur Verfügung stehen (Gilardi et al. 2023; Törnberg 2023). Derzeit entwickelte, multimodale Modelle können zudem Bild- und Video-Inhalte textlich beschreiben, dargestellten Text extrahieren und Audiospuren transkribieren, so dass Multimedia-Inhalte künftig deutlich leichter für die Analyse sehr großer Datenmengen in Betracht kommen. Der große Vorteil für anwendende Forscher*innen ist, dass sie sich über oben angesprochene Fragen des ‚Textpreprocessing' oder aufwändige Verfahren der Trainingsdatenerstellung für komplexe Codebücher keine Gedanken mehr machen müssen. LLMs erfassen und kodieren die Textinhalte allein auf Basis

natürlichsprachlicher Aufgabenbeschreibungen und sind somit deutlich leichter und intuitiver nutzbar. Im Vergleich zu methodischen Vorgehensweisen, die menschliche Kodierer*innen an irgendeinem Schritt im Prozess einsetzen, werden sich je nach Größe der zu bearbeitenden Datensätze und nach Komplexität der zu erfassenden Kategorien vollautomatische Inhaltsanalysen künftig deutlich kostengünstiger durchführen lassen. Forschende müssen bei der Nutzung von APIs wie beispielsweise zu OpenAIs derzeit leistungsfähigsten Modell GPT-4 (OpenAI und Achiam 2024) lediglich optimale Aufgabenbeschreibungen formulieren (sog. ‚prompt engineering') und die von der KI zurückgelieferten Resultate in ausreichender Sample-Größe validieren, um Fehlerraten und etwaige Biases des Modells einschätzen zu können. Gleichzeitig wirft die Nutzung dieser Services in der Forschung neue Fragen hinsichtlich des Datenschutzes und von Open Science Standards, wie bspw. der Gewährleistung von Reproduzierbarkeit von Ergebnissen, auf. Die Methodendiskussion in den CSS wird hierzu zeitnah Antworten liefern und Best Practices vorschlagen.

3.3 Netzwerkanalyse

Die Analyse sozialer Netzwerke hat eine lange Tradition in der empirischen Sozialforschung. Mit dem Aufkommen digitaler Technologie, den damit verfügbaren digitalen Verhaltensdaten und den gestiegenen Rechenkapazitäten ist die Netzwerkanalyse zu einem der größten und wichtigsten Teilbereiche der CSS-Forschung geworden. Auf der Mikroebene interessiert sich die Netzwerkanalyse für die Eigenschaften einzelner Knoten in Bezug auf das Netzwerk, zum Beispiel ihre Zentralität im Netzwerk. Auf der Makroebene werden dagegen Strukturmuster von Teilnetzwerken oder des gesamten Netzwerks ausgewertet. Hier sind beispielsweise Größe und der Grad der Verbundenheit des gesamten Netzwerkes und die Identifikation von Clustern von Interesse.

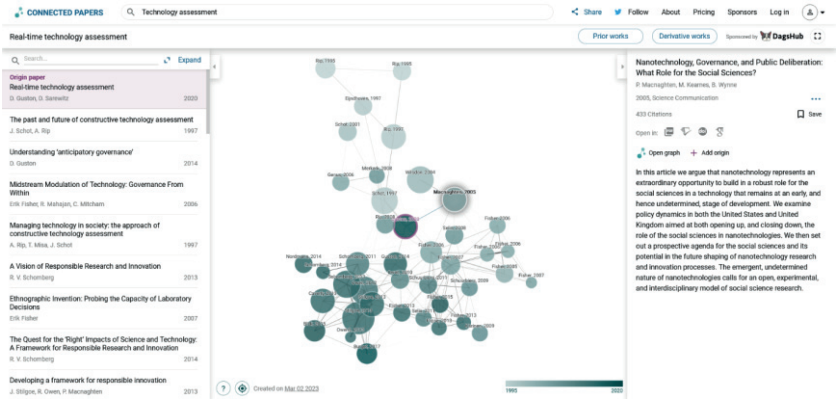

Abbildung 3.2: Beispiel für ein Netzwerk inhaltlich ähnlicher Forschungsartikel der Technikfolgenabschätzung auf Basis überlappender Literaturverweise (https://www.connectedpapers.com, zuletzt geprüft am 06.06.2024).

Für eine Netzwerkanalyse muss in der Regel zunächst das Netzwerk aus den vorliegenden Daten rekonstruiert werden. Für ein Zitationsnetzwerk wissenschaftlicher Publikationen aus einem Fachgebiet etwa würde als erstes ein Korpus aus wissenschaftlichen Publikationen des Feldes auf Basis einer Datenbank wie Scopus[5] zusammenstellen (siehe Abbildung 3.2 für ein Beispiel). Anschließend müssten Autor*innenlisten zu den Forschungsartikeln aus einer öffentlichen Datenbank[6] oder aus dem Volltext für jeden Artikel des Korpus extrahiert werden. Um die Datenqualität zu verbessern, können danach unterschiedliche Schreibweisen von Namen derselben Person vereinheitlicht werden. Sind diese Daten von zufriedenstellender Qualität, kann mit der Rekonstruktion des eigentlichen Netzwerks begonnen werden. Eine Möglichkeit wäre, Autor*innen als Knoten zu modellieren. Kanten zwischen Knoten könnten immer dann gezogen werden, wenn diese gemeinsame Verfasser*innen eines Artikels sind oder wenn diese gemeinsam in einem Artikel zitiert werden. Alternativ könnten auch wissenschaftliche Artikel als Knoten oder

[5] https://www.scopus.com, zuletzt geprüft am 06.06.2024.

[6] Zum Beispiel lassen sich Autor*innen zur DOI eines Papers mit der Datacite API abfragen: https://support.datacite.org/docs/api-get-doi, zuletzt geprüft am 06.06.2024.

Artikel und Autor*innen gemeinsam in einem Netzwerk modelliert werden, mit deren Auswertung dann jeweils andere Aussagen getroffen werden könnten. Ein einmal definiertes Netzwerk besteht in seiner einfachsten Form aus einer zweispaltigen Tabelle, in denen zwei eindeutige Knotennamen in jeder Zeile eine Kantenverbindung beschreiben. Weitere Spalten können Attribute von Kanten wie die Stärker ihrer Assoziation oder von Knoten festhalten. Diese Tabelle dient als Eingabe für Softwareprogramme zur Netzwerkanalyse, die einerseits Kennzahlen, Muster und Strukturen berechnen und andererseits Netzwerke visuell darstellen können. Dies kann beispielsweise die Identifizierung von Clustern oder Gruppen von Elementen im Netzwerk, die Bestimmung der zentralen Knotenpunkte im Netzwerk oder die Messung der Eigenschaften des Netzwerks wie der Dichte oder des Durchmessers umfassen. Die Darstellung als Netzwerk-Graph auf einer Ebene mit Hilfe eines Layout-Algorithmus der stark verbundene Knoten näher zueinander platziert, ermöglicht die visuelle Interpretation und Exploration des Netzwerks. Für ein Netzwerk von Autor*innen wäre dies beispielsweise besonders zentrale Personen als prägende Wissenschaftspersönlichkeiten des Feldes oder thematisch nahe beieinander arbeitende Communities. Wie auch bei der automatischen Inhaltsanalyse müssen die Ergebnisse der Netzwerkanalyse mit Bezug zur Forschungsfrage interpretiert werden. Schlussendlich müssen die Ergebnisse der Netzwerkanalyse sorgfältig validiert werden, um sicherzustellen, dass sie korrekt und robust sind. Dies kann beispielsweise durch einen qualitativen Blick in Datenausschnitte oder Hinzuziehen anderer Datenquellen und bereits veröffentlichter Forschungsliteratur erfolgen.

3.4 Simulation

In den Computational Social Sciences nimmt Simulation einen bedeutsamen Platz ein. Durch die Erstellung von Computersimulationen können komplexe soziale Phänomene untersucht und Vorhersagen über das Verhalten von Individuen und Gruppen in ihrer Einbettung in soziale Systeme getroffen werden. Simulationen können auch dazu beitragen, experimentelle Bedingungen für die Untersuchung von Sozialverhalten in einer kontrollierten Umgebung zu schaf-

fen, was in der Realität oft schwierig oder unmöglich ist. In der CSS-Forschung gebräuchliche Ansätze der Simulation sind beispielsweise System Dynamics (Papachristos 2019) und Agent-based modeling (AM, Gilbert 2004). System Dynamics ist eine Methode der Modellierung und Simulation, die verwendet wird, um das Verhalten von komplexen dynamischen Systemen zu untersuchen. Es basiert auf der Idee, dass soziale Systeme, wie Wirtschaften, Gesellschaften und Organisationen, aus einer Vielzahl von miteinander verbundenen Elementen bestehen, die sich im Laufe der Zeit gegenseitig beeinflussen und verändern. Interaktionen und Rückkopplungsschleifen zwischen den Elementen eines Systems werden in diesem Ansatz in einem eher qualitativen Prozess auf der Makroebene modelliert und ggf. anhand mathematischer Modelle in Verbindung mit empirischen Daten quantifiziert. Agent-based modeling versucht dagegen die Mikro- und Makro-Ebene sozialer Interaktion zu verbinden, indem ‚Agenten' als kleine Programme in der Simulation miteinander nach klar definierten Regeln in Interaktion treten und dadurch auf der Ebene des Gesamtsystems bestimmte Muster erzeugen. Gelingt es dabei Muster zu reproduzieren, die sich auch in empirischen Beobachtungsdaten zeigen, liefert AM Hinweise darauf, dass diese Muster durch die in den Agenten einprogrammierten, minimalen Regeln und Verhaltensweisen erklärbar sind. Beispielsweise untersucht Waldherr (2014) mit Hilfe von AM, durch welche Faktoren sich thematische Aufmerksamkeitswellen (issue-attention cycles) in der Nachrichten-Berichterstattung erklären lassen. In experimentellen Vorgehensweisen können AM-Studien dann Variablen kontrolliert variieren, um deren Einfluss auf das Modell-Ergebnis besser zu verstehen. Simulationsansätze können für die Technikfolgenabschätzung enormes Potenzial entfalten, wenn sie für die Untersuchung der Interaktion von neuen Technologien mit menschlichen Nutzenden fruchtbar gemacht werden können. Beispielstudien finden sich vornehmlich in der Umwelt-, Klima- und Energieforschung (Castro et al. 2020) sowie im Gesundheitsbereich (Djanatliev et al. 2012).

3.5 Statistik

Eine große Rolle in der CSS-Forschung spielt freilich auch die klassische Statistik in ihren vielfältigen Varianten. Dabei werden häufig mit den oben

beschriebenen Methoden der Inhaltsanalyse, Netzwerkanalyse oder Simulation Daten produziert, die mit statistischen Verfahren weiter untersucht werden. Deskriptive Statistik wird eingesetzt, um die erhobenen bzw. aus anderen Verfahren produzierten Daten zu beschreiben sowie Muster und Trends in den Daten darzustellen. Dazu kommen Maße wie Mittelwerte, Standardabweichungen und Histogramme zum Einsatz. Mit Verfahren der Inferenzstatistik wird darüber hinaus versucht, Schlussfolgerungen über eine gesamte Population auf Basis einer Stichprobe zu ziehen. Dazu werden statistische Tests verwendet, um die Signifikanz von Unterschieden zwischen Gruppen zu bestimmen, Konfidenzintervalle zu berechnen oder Hypothesen zu testen. Dabei komme immer häufiger Modellbildungen der Bayesschen Statistik zum Einsatz, die genutzt werden kann, um die Wahrscheinlichkeit von Hypothesen in Abhängigkeit von Beobachtungsdaten zu berechnen und zu aktualisieren. Predictive Modeling wird in der CSS-Forschung für die Vorhersage von zukünftigen Ereignissen auf Basis historischer Daten genutzt. Es verwendet statistische Modelle, um Muster in Daten zu erkennen und diese Muster zu nutzen, um Vorhersagen zu treffen. Predictive Modeling wird vielfach für Probleme und Fragen der Wirtschaft, dem Marketing, der Medizin und der Kriminalitätsbekämpfung eingesetzt. Causal Inference befasst sich dagegen mit der Untersuchung von Ursache-Wirkungs-Beziehungen zwischen Variablen. Es versucht, durch statistische Analyse zu bestimmen, ob eine bestimmte Variable tatsächlich die Ursache für eine Veränderung in einer anderen Variable ist. Causal Inference ist in vielen Bereichen der Forschung und Politik relevant, um Entscheidungen auf Basis von Ursache-Wirkungs-Beziehungen zu treffen. (Engel 2021) bezeichnet kausale und vorhersagende Inferenz gar als zwei Hauptziele der CSS-Forschung. Anhand eines Strukturgleichungsmodells auf Basis von Befragungsdaten wird in dem einführenden Artikel ein ursächlicher Zusammenhang von Vorurteilsstrukturen auf extrem rechte politische Einstellung gezeigt. Andres und Slivko (2021) stellen in ihrer Studie zu den Auswirkungen des Netzwerkdurchsetzungsgesetzes (NetzDG) in Deutschland fest, dass das Gesetz als ursächlich für den Rückgang von Hassrede in deutschen Twitter/X-Posts angesehen werden kann. Die Autor*innen beobachten in ihrem Untersuchungszeitraum von drei Jahren einen Rückgang von rechtswidrigen Tweetinhalten um 10 %, den sie mittels eines ‚difference-in-differences'-Ansatz mit der Vergleichsgruppe österreichischer Tweets ursächlich dem

Gesetz Inkrafttreten des Gesetzes zuschreiben. Für die Messung des ‚Hassanteils' in Tweets kommt ein Textklassifikationssystem zum Einsatz, das jedem Tweet einen Toxizitätsscore vergibt. Kausale Inferenz lässt sich aber auch gut direkt mit Textdaten realisieren. Eine anschauliche Einführung hierfür geben Grimmer, Roberts und Stewart (2022).

4 Die CSS-Forschungslandschaft in Deutschland

Seit mehreren Jahren nimmt das Wachstum und die Ausdifferenzierung der Computational Social Sciences auch in Deutschland erheblich an Fahrt auf. Dies Kapitel beschreibt die Entwicklung entlang von Institutionen, wissenschaftlichen Vernetzungs- und Publikationsorten sowie einschlägigen Projekten. Ziel ist es auch hier, anstatt einer vollständigen Auflistung eher einen Überblick zu geben.

4.1 Institute, Lehrstühle und Forschungsabteilungen

Die CSS-Forschung ist in Deutschland zunehmend institutionell verankert. Als eine der ersten außeruniversitären Forschungseinrichtungen nahm bereits 2011 das Alexander von Humboldt Institut für Internet und Gesellschaft (HIIG) seine Arbeit auf, zunächst jedoch noch nicht mit einem klar konturierten CSS-Fokus. Als europaweit erste sozialwissenschaftliche Einrichtung richtete zwei Jahre später GESIS eine Abteilung Computational Social Science ein. Seitdem wurden drei weitere Institute für Digitalisierungs- und Internetforschung gegründet und an bereits bestehenden Instituten spezielle CSS-Abteilungen eingerichtet (siehe Tabelle 4.1). Im internationalen Vergleich begann der Aufbau des Forschungsfeldes dabei etwas später als im angloamerikanischen Raum, ist mittlerweile aber auf eine beachtliche Größe angewachsen. International vernetzen sich die deutschen Internetforschungsinstitute im Network of Internet & Society Centers.[1]

[1] https://networkofcenters.net, zuletzt geprüft am 02.07.2024.

Gleichzeitig wurde der Aufbau der CSS-Forschung an den deutschen Universitäten durch Einrichtung neuer Lehrstühle und Studiengänge stark vorangetrieben. Lehrstühle mit der Denomination CSS bestehen unter anderen an der LMU[2] und TU München[3], der Universität Stuttgart[4] sowie der RWTH Aachen[5]. Weitere CSS-Professuren in Bremen und an der FAU Erlangen-Nürnberg sind derzeit im Berufungsverfahren. An zahlreichen Lehrstühlen mit Bezeichnungen wie „Steuerung innovativer und komplexer technischer Systeme“[6] (Universität Bamberg), „Digital Social Science“[7] (Universität Hamburg) oder „Data Science in den Wirtschafts- und Sozialwissenschaften“[8] (Universität Mannheim) wird zudem ebenfalls teilweise seit vielen Jahren CSS-Forschung betrieben. Als eigenständiger Studiengang lässt sich CSS derzeit belegen in Aachen und Bamberg. Darüber hinaus werden CSS-Schwerpunkte mittlerweile in der Methodenausbildung vieler grundständiger Sozialwissenschaftsstudiengänge angeboten.

[2] https://www.css.soziologie.uni-muenchen.de, zuletzt geprüft am 06.06.2024.
[3] https://www.hfp.tum.de/css/startseite, zuletzt geprüft am 06.06.2024.
[4] https://www.sowi.uni-stuttgart.de/abteilungen/css, zuletzt geprüft am 06.06.2024.
[5] https://www.sth.rwth-aachen.de/cms/fachgruppe/Die-Fachgruppe/~jwcez, zuletzt geprüft am 06.06.2024
[6] https://www.uni-bamberg.de/politikdigital/team/prof-dr-andreas-jungherr/, zuletzt geprüft am 06.06.2024.
[7] https://www.wiso.uni-hamburg.de/fachbereich-sowi/professuren/oberg, zuletzt geprüft am 06.06.2024.
[8] https://www.bwl.uni-mannheim.de/strohmaier, zuletzt geprüft am 06.06.2024.

Tabelle 4.1: Außeruniversitäre Forschungseinrichtungen in Deutschland und international mit dem Schwerpunkt Digitalisierung und Computational Social Science (Größe bezieht sich auf die Anzahl an Forschenden in den Bereichen CCS / Digitale Kommunikation / Internet Research).

Institut/ Webseite	Sitz	CSS seit	Größe
Alexander von Humboldt Institut für Internet und Gesellschaft (HIIG)[9]	Berlin	2011	ca. 50
Leibniz-Institut für Sozialwissenschaften (GESIS)[10]	Mannheim, Köln	2013	ca. 30
Center for Advanced Internet Studies (CAIS)[11]	Bochum	2016	ca. 40
Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI)[12]	Hamburg	2016	Ca. 20
Weizenbaum-Institut für die vernetzte Gesellschaft[13]	Berlin	2017	ca. 105
Bayerisches Forschungsinstitut für Digitale Transformation (bidt)[14]	München	2018	ca. 15
Oxford Internet Institute (OII)[15]	Oxford, UK	2001	ca. 50
UMass Computational Social Science Institute (CSSI)[16]	Amherst, USA	2010	ca. 80

[9] https://www.hiig.de, zuletzt geprüft am 06.06.2024.
[10] https://www.gesis.org/institut/ueber-uns/abteilungen/computational-social-science, zuletzt geprüft am 06.06.2024.
[11] https://www.cais-research.de, zuletzt geprüft am 06.06.2024.
[12] https://leibniz-hbi.de/de/forschung/forschungsprogramme/media-research-methods-lab, zuletzt geprüft am 06.06.2024.
[13] https://www.weizenbaum-institut.de, zuletzt geprüft am 06.06.2024.
[14] https://www.bidt.digital, zuletzt geprüft am 06.06.2024.
[15] http://www.oii.ox.ac.uk/, zuletzt geprüft am 06.06.2024.
[16] https://www.cssi.umass.edu, zuletzt geprüft am 06.06.2024.

4.2 Konferenzen, Journale, Netzwerke

Der Aufbau von Instituten, Abteilungen und Lehrstühlen wird begleitet durch die Entstehung neuer wissenschaftlicher Austausch- und Publikationsorte.

4.2.1 Konferenzen

Als wichtigster Vernetzungsort der CSS-Community kann die International Conference on Computational Social Science (IC²S²) gelten, die seit 2015 jährlich ausgerichtet wird.[17] Die Konferenz ist von Anbeginn durch ihren hohen Grad an Interdisziplinarität geprägt, wobei besonders in den ersten Jahren nicht-sozialwissenschaftliche Disziplinen wie Physik und Informatik deutlich dominierten. Die Organisation der Konferenz wird mittlerweile über einen eigenen Fachverband International Society for Computational Social Science (ISCSS) getragen.[18] Bereits seit dem Jahr 2000 führt die Association of Internet Researchers (AoIR) ihre Jahreskonferenzen[19] durch, bei denen ebenfalls viele CSS-Themen, die soziale, kulturelle, politische, rechtliche, ästhetische, wirtschaftliche und/ oder philosophische Aspekte des Internets behandeln, im Mittelpunkt stehen. Die Konferenz ist ebenfalls von Beginn an sehr interdisziplinär aufgestellt und wird unter anderem von Forschenden der Kommunikations- und Medienwissenschaften gerne besucht.

Seit einigen Jahren hat sich auch von Seiten der Informatik das Interesse an CSS deutlich verstärkt. Sichtbar wird das an der Einrichtung von Workshops und dedizierten Tracks der Top-Konferenzen großer Fachverbände. So widmet die Association of Computational Linguists (ACL) seit 2016 der Verbindung von Natural Language Processing und CSS einen eigenen Workshop.[20] Die ACL Top-Konferenzen (ACL, EACL, EMNLP, NAACL) werben seit ein paar

[17] Die IC²S² 2023 findet vom 17.-20. Juli in Kopenhagen, Dänemark statt: https://www.ic2s2.org, zuletzt geprüft am 06.06.2024.
[18] https://iscss.org, zuletzt geprüft am 06.06.2024.
[19] https://aoir.org, zuletzt geprüft am 06.06.2024.
[20] https://sites.google.com/site/nlpandcss, zuletzt geprüft am 06.06.2024.

Jahren in ihren Calls direkt um CSS-Einreichungen. Die Informatik-Fachverbände Association for the Advancement of Artificial Intelligence (AAAI) und Association for Computing Machinery (ACM) widmen bereits seit mehreren Jahren eigene Konferenzen den Fragestellungen rund um die Analyse von Internet- und Social-Media-Daten, die zunehmend von der CSS-Community frequentiert werden.[21]

4.2.2 Journale

Die Entwicklung des Faches spiegelt sich ebenso auf Ebene der wissenschaftlichen Zeitschriften. Seit seiner erste Ausgabe 1983 versammelt Social Science Computer Review[22] Pioniere der CSS-Forschung noch lange bevor das Schlagwort in der akademischen Welt die Entwicklung als solche zusammenfasste. Gesellschaftliche Fragen im Zeitalter der Digitalisierung werden seit mehreren Jahren in Zeitschriften wie New Media & Society[23], Big Data & Society[24] und EPJ Data Science[25] veröffentlicht. finden sich in Darüber hinaus haben etablierte Top-Journale einzelner Disziplinen sowie disziplinübergreifende Journals CSS-Sonderausgaben veröffentlicht. So zum Beispiel Nature,[26] und Political Analysis[27], die Meilensteine der Methodenforschung im CSS-Bereich in

[21] Die wichtigsten Konferenzen: The International AAAI Conference on Web and Social Media (ICWSM), ACM Web Conference (TheWebConf), International ACM Conference on Web Science (WebSci).

[22] Impact Factor: 4,42 / 5-Year Impact Factor: 5,21, https://journals.sagepub.com/home/ssc, zuletzt geprüft am 06.06.2024.

[23] Impact Factor: 5,31 / 5-Year Impact Factor: 7,24, https://journals.sagepub.com/home/nms, zuletzt geprüft am 06.06.2024.

[24] Impact Factor: 8,73 / 5-Year Impact Factor: 10,51, https://journals.sagepub.com/home/bds, zuletzt geprüft am 06.06.2024.

[25] Impact Factor: 3,63 / 5-year Impact Factor: 5,15, https://epjdatascience.springeropen.com, zuletzt geprüft am 06.06.2024.

[26] https://www.nature.com/collections/cadaddgige, zuletzt geprüft am 06.06.2024.

[27] Impact Factor: 9,02, https://www.cambridge.org/core/journals/political-analysis, zuletzt geprüft am 06.06.2024.

mehreren virtuellen Sonderausgaben gebündelt haben.[28] Die weitere Ausdifferenzierung des Feldes wird in jüngster Zeit anhand von neu gegründeten Zeitschriften deutlich, wie dem Journal of Computational Social Science[29] und Computational Communication Research[30], die teilweise oder gänzlich Open Access publizieren.

4.2.3 Netzwerke und Verbände

Neben den bereits oben im Zusammenhang mit den einschlägigen Konferenzen genannten CSS-spezifischen Fachverbänden gibt es auch in den klassischen Fachverbänden Bestrebungen, die CSS institutionell zu verankern. Insbesondere die Kommunikationswissenschaft nimmt hier eine Vorreiterrolle ein. 2017 wurde die Computational Methods Division der International Communication Association (ICA)[31] gegründet. Die Fachgruppe ‚Methoden' der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft beschäftigt sich vorrangig mit CSS-Methoden.[32] Im Jahr 2020 wurde aus der Fachgruppe heraus eine Arbeitsgruppe ‚Computational Social Science in der Lehre' gegründet, welche die Vereinheitlichung und Verbesserung des Lehrangebots für CSS-Methoden voranbringen möchte. 2021 nahm das DFG-Netzwerk ‚Potenziale und Herausforderungen der Computational Communication Science am Beispiel von Online-Protest' seine Arbeit auf.[33] Von Seiten der Informatik wurde bei der German Society for Computational Linguistics and Language Technology (GSCL) 2020 die Arbeitsgruppe ‚Computational Linguistics for Political and Social Sciences' ins Leben gerufen, die informatori-

[28] https://www.cambridge.org/core/journals/political-analysis/special-collections, zuletzt geprüft am 06.06.2024.

[29] https://www.springer.com/journal/42001/, zuletzt geprüft am 06.06.2024.

[30] https://computationalcommunication.org/ccr, zuletzt geprüft am 06.06.2024.

[31] https://www.icahdq.org/group/compmethds, zuletzt geprüft am 06.06.2024.

[32] https://www.dgpuk.de/de/methoden-der-publizistik-und-kommunikationswissenschaft.html, zuletzt geprüft am 06.06.2024.

[33] https://gepris.dfg.de/gepris/projekt/464913012, zuletzt geprüft am 06.06.2024.

sche Methoden entlang der Anforderungen aus den Anwenderdisziplinen weiter entwickeln möchte.[34] Die klassischen sozialwissenschaftlichen Fachverbände DGS, DVPW und DGfP unterhalten derzeit keine dezidierten CSS-Sektionen.

4.2.4 Einschlägige Projekte

Seit den frühen 2010er Jahren kommt es unter anderem durch die Einrichtung entsprechender Förderlinien des BMBF[35] oder Schwerpunktprogrammen der DFG zu einer verstärkten interdisziplinären Zusammenarbeit zwischen Informatik und anwendenden Disziplinen der Geistes- und Sozialwissenschaften. Folgende Projekte stellen die große Bandbreite der Kooperationen beispielhaft dar:

- ODYCCEUS:[36] Das 2021 angeschlossene Horizon-2020 Projekt ‚Opinion Dynamics and Cultural Conflict in European Space' hat sich methodische Fortschritte bei der Erforschung globaler Systemdynamiken in sozialen Krisenlagen zum Ziel gesetzt. Verbunden wurden CSS-Methoden der automatischen Inhaltsanalyse mit Verfahren der Spieltheorie und der Netzwerkanalyse, z. B. zur Erforschung von Polarisierungsdynamiken und zur Ausbreitung von Desinformation. Entstanden ist unter anderem die Plattform PENELOPE[37], die eine Reihe von den im Projekt entwickelten bzw. für das Projekt angepassten Werkzeuge, wie beispielsweise TwitterExplorer auflistet. Diese Werkzeuge wurden in eine einheitliche Analyse- und Visualisierungspipeline integriert, mit der unter anderem argumentative Auseinandersetzungen in Nachrichtentexten und Nutzerdebatten untersucht werden können (Willaert et al. 2022).

[34] https://old.gscl.org/en/arbeitskreise/cpss, zuletzt geprüft am 06.06.2024.
[35] Beispielsweise die 2013 gestartete Förderung von „Forschungs- und Entwicklungsvorhaben aus dem Bereich der eHumanities", https://www.bmbf.de/bmbf/shareddocs/bekanntmachungen/de/2013/01/804_bekanntmachung.html, zuletzt geprüft am 17.06.2024.
[36] https://www.odycceus.eu, zuletzt geprüft am 17.06.2024.
[37] https://penelope.digitalmethods.net, zuletzt geprüft am 17.06.2024.

- MARDY:[38] Das seit 2018 im SPP-1999 ‚Robust Argumentation Machines (RATIO)' geförderte, interdisziplinäre Projekt aus Computerlinguistik, maschinellem Lernen und Politikwissenschaft entwickelt neue computergestützte Modelle und Methoden zur Analyse von Argumenten im politischen Diskurs. Im Fokus steht insbesondere die Erfassung der Dynamik des diskursiven Austauschs über kontroverse Themen im Zeitverlauf. Dazu wurden in MARDY Verfahren zur (teil-)automatischen Erstellung sogenannter Diskursnetzwerke entwickelt, in denen politische Forderungen und Argumente mit den diese vertretenden politischen Akteuren visuell dargestellt werden (Haunss et al. 2020).
- SMO:[39] Seit 2020 wird das Social Media Observatory am HBI als virtuelle Forschungsinfrastruktur innerhalb des Forschungsinstitut Gesellschaftlichen Zusammenhalt (FGZ) aufgebaut. Das SMO zielt auf eine Langzeiterhebung der öffentlichen Kommunikation auf ausgewählten sozialmedialen Plattformen und Online-Nachrichtenmedien. Auf der Grundlage systematisch zusammengestellter Listen von Sprecherkategorien, wie z. B. Parlamentariern oder Medienorganisationen, werden sowohl statistische als auch inhaltliche Daten erhoben, um den deutschen Social-Media-Diskurs im Vergleich zu Massenmedien zu untersuchen. Aggregierte Ergebnisse werden über interaktive Dashboards veröffentlicht. Einem Do-it-yourself-Ansatz folgend, stellt das SMO außerdem verschiedene Tools, kuratierte Datensätze und dokumentierte Workflows zur Verfügung, mit denen Forscher*innen eigene thematische Ad-hoc-Datensammlungen durchführen können.

[38] https://www.mardy-project.eu, zuletzt geprüft am 17.06.2024.
[39] https://smo.leibniz-hbi.de, zuletzt geprüft am 17.06.2024.

- MeMo:KI:[40] In dem am CAIS durchgeführten Projekt ‚Meinungsmonitor Künstliche Intelligenz' werden Umfrageforschung mit Verfahren der automatischen und manuellen Inhaltsanalyse von Nachrichten und Social-Media-Daten kombiniert, um die Bevölkerungsmeinung und Berichterstattung über Künstliche Intelligenz in Deutschland zu erforschen. Die Datenerhebungen werden monatlich wiederholt, so dass ein umfangreicher Längsschnitt des Meinungsbildes sichtbar wird. Aktuelle Erkenntnisse, wie dass der die Technologie ChatGPT der Firma open.ai bereits ca. 30 % der Deutschen bekannt ist,[41] werden in regelmäßigen Pressemitteilungen und Fact Sheets veröffentlicht.

[40] https://www.cais-research.de/forschung/memoki, zuletzt geprüft am 17.06.2024.

[41] https://www.cais-research.de/news/chatgpt-wie-viele-menschen-kennen-dich-bereits, zuletzt geprüft am 17.06.2024.

5 CSS-Toolbox für Einsteiger

In den Digital Humanities wurde lange und ausführlich diskutiert, ob es eine Anforderung an den geisteswissenschaftlichen Nachwuchs sein kann, sich für die digitale Datenerhebung und -Auswertung Programmierkenntnisse anzueignen. Initiativen wie ‚forText'[1] machen mit ihren Sammlungen von ‚tools' für die Analyse (literarischer) Texte deutlich, dass es eine große Bandbreite zwischen der Nutzung bestehender Tools/off-the-shelf (OTS) Software und der Programmierung maßgeschneiderter Lösungen gibt. Dies trifft ebenfalls auf die CSS zu, wobei größere Datenumfänge, besondere Dateneigenschaften oder Analyseerfordernisse die Nutzung von OTS-Lösungen häufiger als in den DH erschweren dürften. Für erste Einblicke in große digitale Sammlungen von Netzwerken oder Inhaltsdaten erscheinen OTS-Programme vielfach ausreichend. Die Beantwortung einer komplexen Forschungsfrage wird aber in der Regel die Entwicklung eigener Auswertungsroutinen auf Basis modulari-sierter Bausteine (Programmbibliotheken bzw. -pakete) erfordern, für deren Einsatz ein gewisses Kenntnis- und Erfahrungslevel notwendig ist. Hierzu hilfreich sind die zahlreich online und frei verfügbaren Lern- und Lehrinhalte, die auf den in der Regel als Open Source-Software veröffentlichten Programmierumgebungen beruhen.

Nützliche Tool-Übersichten liefern forText[2] und das Wiki des Social Media Observatory[3]. Das folgende Kapitel gibt einen Überblick über verschiedene Ansätze der Softwarenutzung in den CSS von einem einfachen Einstieg mit Software as a Service (SaaS) durch Drittanbieter, über die Nutzung von OTS-Programmen auf lokalen Systemen bis hin zur Entwicklung eigener Analyse-Skripte anhand ausgewählter Beispiele für die einzelnen Kategorien.

[1] https://fortext.net/ueber-fortext, zuletzt geprüft am 17.06.2024.
[2] https://fortext.net/tools, zuletzt geprüft am 09.09.2024.
[3] https://smo-wiki.leibniz-hbi.de, zuletzt geprüft am 09.09.2024.

5.1 Software as a Service

Zur Überbrückung der Technologielücke zwischen Entwicklungen in der Informatik und den anwendenden Disziplinen setzte sich im Bereich der Digital Humanities frühzeitig ein Gedanke der Service-orientierten Softwareentwicklung bzw. des Angebots von Software als Service für die Geisteswissenschaften durch (Gold 2009). Dies führte von Entwicklung von einzelner Webservices[4], über komplexe Webanwendungen zur Auswertung kleinerer Korpora bis hin zu groß angelegten virtuellen, digitalen Infrastrukturen[5], die Services entwickeln, standardisieren und zentral bereitstellen. Dieser Ansatz wurde jedoch schon früh kritisiert, da die erheblichen Aufwände zur Entwicklung, Verallgemeinerung, Standardisierung und zentralen Bereitstellung von Diensten zur Auswertung digitaler Daten mit den Anforderungen von Innovation und der Erforschung neuer Modelle unvereinbar sind (van Zundert 2012). Tatsächlich konnten während der 10-jährigen Förderphase von CLARIN-D zahlreiche digitale Sprachressourcen erschlossen sowie digitale Werkzeuge und Standards entwickelt werden. Mit den Erfordernissen der sich rasant entwickelnden informatischen Methoden konnten die in dieser Zeit entstandenen Services jedoch nicht mithalten, so dass sie mittlerweile als nicht mehr zeitgemäß gelten und nur noch selten genutzt werden dürften.

Für die CSS wirken die Schwierigkeiten und Probleme des SaaS Ansatzes umso stärker. Die oftmals sehr großen Datenmengen erschweren den Import und die Verarbeitung in Webanwendungen. Anforderungen an den Datenschutz, die beispielsweise eine Verarbeitung von personenbezogenen Daten innerhalb der EU vorschreiben, schließen bestimmte, außerhalb der EU gehostete Webanwendungen von vornherein aus. Trotzdem können über den SaaS-

[4] Zum Beispiel WebLicht zur syntaktischen und morphologischen Annotation von Sprachdaten (https://weblicht.sfs.uni-tuebingen.de, zuletzt geprüft am 17.06.2024).

[5] Die *Common Language Resources and Technology Infrastructure* (CLARIN) wurde von der EU seit 2012 mit über 165 Mio. Euro gefördert. Der deutsche Ableger CLARIN-D ist mittlerweile in der Nationale Forschungsdateninfrastruktur (NFDI) als Konsortium Text+ aufgegangen (https://www.nfdi.de/textplus, zuletzt geprüft am 17.06.2024.).

Ansatz bereitgestellte Tools einen nützlichen, ersten Einstieg in CSS-Methoden darstellen, indem sie einen niedrigschwelligen Zugang zu einfachen Methodenanwendungen bieten.

- Voyant Tools[6] ermöglichen eine überschaubare Menge von Dokumenten oder URLs für Webseiten mit einer einfachen, explorativen Inhaltsanalyse quantitativ auszuwerten (Sinclair und Rockwell 2012). Die Webanwendung bietet Wortfrequenz, Konkordanz-Analysen und Kookkurrenzanalysen, mit denen man sich leicht einen Überblick über das wichtigste Vokabular und auffällige Sprachgebrauchsmuster in einzelnen langen Dokumenten oder einem kleinen Korpus verschaffen kann.
- CATMA[7] ist eine in Deutschland offen entwickelte Forschungssoftware, die als Webanwendung qualitative und quantitative Auswertungen von mittleren Dokumentmengen unterstützt (Meister et al. 2019). Qualitative Ansätze lassen sich über die Kodierung von Dokumentabschnitten mit Hilfe selbst erstellter Codebücher analog zu Softwareprodukten wie MAXQDA umsetzen. Für quantitative Auswertungen können Vokabularstatistiken und qualitative Kodierungen gemeinsam ausgewertet werden.
- OpenFraming AI[8] ist eine neue SaaS-Anwendung aus dem Bereich der Kommunikationswissenschaften, die Topic Modeling mit Ansätzen des überwachten maschinellen Lernens zur automatischen Kategorisierung von Dokumenten in Frames verbindet (Guo et al. 2023). Der Ansatz stellt eine interessante Kombination von datengetriebenem Clustering und deduktiver Kodierung dar und richtet sich explizit an Forschende ohne Programmierkenntnisse.

Alle diese SaaS-Anwendungen setzen voraus, dass man seine Daten in vorherigen Schritten bereits erhoben hat und diese zumindest teil-strukturiert vorliegen. Für einen Einstieg in die CSS-Forschung mit großen Datenbeständen

[6] https://voyant-tools.org, zuletzt geprüft am 17.06.2024.
[7] https://catma.de, zuletzt geprüft am 17.06.2024.
[8] https://github.com/dnaaun/openFraming, zuletzt geprüft am 02.07.2024.

könnten daher auch Angebote von Interesse sein, die nicht nur Auswertungsmöglichkeiten, sondern auch Daten anbieten. Aufgrund von Urheberrechtsproblemen sowie finanziellen und personellem Aufwand werden solche Datensammlungen in aller Regel nur von kommerziellen Unternehmen angeboten. Für die Beobachtung von redaktionellen Massenmedien und sozialen Medien gibt es eine Reihe von Anbietern, die Zugang zu Nachrichtenmedien und beispielsweise Twitter/X-Datensammlungen ermöglichen.[9] Für die Bearbeitung akademischer Forschungsfragen sind diese Anbieter jedoch in der Regel weniger interessant. Gegen die Nutzung solcher Services sprechen zum einen die hohen Kosten und eine Optimierung der Auswertungstools auf Markt- und Markenbeobachtung anstelle auf die Beantwortung komplexer sozialwissenschaftlicher Forschungsfragen in einer Kombination aus qualitativen und quantitativen Analyseperspektiven. Vor allem aber sprechen gegen den Einsatz solcher kommerziellen Systeme die gravierenden Widersprüche zu einzelnen Open Science Prinzipien. Da Datengrundlagen, Details zu Erhebungsstrategien und Auswertungsalgorithmen als Geschäftsgeheimnisse nicht öffentlich bekannt gemacht werden, ist vielfach weder eindeutig, auf welche Grundgesamtheit sich die Auswertungen beziehen, noch ist die Anforderung von Nachvollziehbarkeit und Reproduzierbarkeit der Forschungsergebnisse einlösbar. Von der Nutzung solcher kommerziellen SaaS-Anbieter muss daher dringend abgeraten werden. Studien, die unter Nutzung solcher Dienste erstellt wurden, dürften mit hoher Wahrscheinlichkeit in blind review-Verfahren für die Veröffentlichung in renommierten Fachjournalen ausgeschlossen werden.

[9] Zeitreihenanalysen für Schlagwortsuchen in Nachrichtenmedien oder Twitter-Daten ermöglichen beispielsweise https://radiosphere.de, https://brandwatch.com und https://www.meltwater.com. Eine Ausnahme aus dem akademischen Bereich ist https://mediacloud.org, Schlagwortsuchen, Wortfrequenzstatistiken und Themensuchen für eine große Zahl internationaler Nachrichtenmedien als Open Source SaaS-Plattform und über API-Schnittstellen anbieten, alle Links zuletzt geprüft am 17.06.2024.

5.2 Off-the-shelf Software

Alternativ zu Webanwendungen, die über den Browser genutzt werden können, bietet sich die lokale Installation einer Forschungssoftware auf dem eigenen Rechner bzw. Server an. Daten müssen für die Analyse dann nicht mehr die eigene Organisation verlassen und können auch in großer Menge, je nach verfügbarer Hardware, verarbeitet werden. Ähnlich wie bei SaaS-Anwendungen müssen Daten in der Regel vor Nutzung der Software erhoben und zur weiteren Verarbeitung gegebenenfalls in ein für das Programm lesbares Format konvertiert werden. OTS-Forschungssoftware ist meist für sehr spezifische Analysezwecke entwickelt worden und erlaubt dementsprechend nur sehr bestimmte Analyseabläufe, die bisweilen für bestimmte Daten oder Forschungsfragen nicht passgenau zugeschnitten sind. Folgende aufgelistet sind OTS-Programme, die eine größere Bandbreite an Analysekapazitäten zur Verfügung stellen, die zu komplexeren Analyseabläufen kombiniert werden könnten:

- MAXQDA[10] und andere primär für die qualitative Datenanalyse (QDA) entwickelte Softwareprogramme haben in den jüngeren Versionen immer mehr Funktionalitäten zur automatischen Inhaltsanalyse integriert (Rädiker und Kuckartz 2019). Mit MAXQDA lassen sich beispielsweise größere Mengen an Tweets erheben bzw. laden und mit Erweiterungsmodulen wie MAXDictio sprachstatistisch auswerten (häufigste Begriffe, Keywords). Eine automatische Kodierung kann entlang von Schlüsselbegriffen oder Hashtags vorgenommen und anschließend analog zu manueller Kodierung über Code-Überschneidung, -Nähe oder -Kookkurrenz ausgewertet werden. Über das Austauschformat REFI-QDA (Evers et al. 2020) lassen sich die kodierten Dokumente und Codebücher für andere QDA-Programme oder eigene Skripte exportieren.

[10] https://www.maxqda.com, zuletzt geprüft am 17.06.2024.

- QDAMiner[11] folgt ebenfalls dem Ansatz, qualitative und quantitative Auswertungsverfahren für Mixed-Method-Analysen miteinander zu verknüpfen. Anders als MAXQDA bleibt die kommerzielle Software nicht bei einfachen Frequenzanalysen und Sprachstatistik stehen, sondern enthält mit der Erweiterung WordStat viele Funktionen aus dem NLP-Bereich wie automatische Textklassifikation, Eigennamenerkennung, Geolokalisierung von Eigennamen und Topic Modeling.[12]
- CorpusExplorer[13] ist eine kostenlose Open Source Software, die in erster Linie für Forschungsfragen im Bereich der Korpuslinguistik entwickelt wurde, aber darüber hinaus auch für andere Text Mining-Interessierte nützlich sein dürfte. Laut Entwickler stellt der CorpusExplorer über 50 interaktiven Auswertungsmöglichkeiten mit einer einfachen Bedienung bereit, darunter Textakquise, Taggen oder die grafische Aufbereitung von Ergebnissen (Rüdiger 2021).

[11] https://provalisresearch.com/products/qualitative-data-analysis-software/, zuletzt geprüft am 17.06.2024.

[12] https://provalisresearch.com/products/content-analysis-software/, zuletzt geprüft am 17.06.2024.

[13] https://notes.jan-oliver-ruediger.de/software/corpusexplorer-overview, zuletzt geprüft am 17.06.2024.

- Interactive Leipzig Corpus Miner (iLCM)[14] ist eine kostenlose Open Source-Forschungssoftware, die Anforderungen an State-of-the-art-Verfahren zur automatischen Inhaltsanalyse auf sehr großen Textmengen mit einer leichten Bedienbarkeit über grafische Benutzeroberflächen sowie Transparenz und individuelle Anpassbarkeit der Analyseabläufe miteinander vereint. Die Software bietet unter anderem Verfahren für Sprachstatistische Analysen, Topic Modeling, Textklassifikation und Eigennamenerkennung. Die Datenhaltung in einer Datenbank samt Volltextindex ermöglicht einen schnellen Zugriff auch auf sehr große Dokumentmengen (Niekler et al. 2018). Die Implementierung aller Analyseabläufe in der Skript-Sprache R lässt sowohl die Nachvollziehbarkeit der Auswertungsalgorithmen direkt im Quellcode als auch deren Projektspezifische Anpassung bzw. Erweiterung zu. Auf diese Weise werden Bedienbarkeit und Flexibilität der Analysefunktionen miteinander kombiniert, wie dies bei anderen OTS-Programmen bislang nicht der Fall ist. In diesem Sinne nutzt beispielsweise das Deutsche Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) den iLCM zur Entwicklung neuer Evaluierungsprozesse.[15]
- TwitterExplorer[16] bietet einen einfachen Einstieg in die Analyse von Twitter-Netzwerken. Die Open-Source-Software bietet Funktionalitäten zum Erheben, Verarbeiten und visuellen Exploration von Twitter-Daten in einer grafischen Benutzeroberfläche (Pournaki et al. 2021).
- Gephi[17] ist eines der am meistgenutzten Open Source-Programme zur Netzwerkanalyse. Die Software bietet umfangreiche Funktionen zur Visualisierung und Auswertung sehr großer Netzwerke wie beispielsweise die Berechnung von Zentralitätsmaßen oder Community-Clustern (Bastian, Heymann und Jacomy 2009).

[14] https://ilcm.informatik.uni-leipzig.de/, zuletzt geprüft am 18.06.2024.

[15] https://www.deval.org/de/methoden-standards/unsere-methodenprojekte/text-mining, zuletzt geprüft am 18.06.2024.

[16] https://twitterexplorer.org/, zuletzt geprüft am 18.06.2024.

[17] https://gephi.org/, zuletzt geprüft am 18.06.2024.

Über grafische Benutzeroberflächen, gute Dokumentation und ausgewählte Funktionalitäten bieten viele OTS-Programme einen guten Einstieg in CSS-Methoden mit sehr großen Datenmengen. Die hier aufgelisteten Programme sind in der wissenschaftlichen Forschung weit verbreitet. Die Open Source-Programme genügen zudem den wissenschaftlichen Anforderungen von Nachvollziehbarkeit und Reproduzierbarkeit. Für die kommerziellen Programme gilt dies nur eingeschränkt, so dass neben den zum Teilerheblichen Lizenzkosten auch wissenschaftsethische Argumente gegen deren Einsatz sprechen könnten.

5.3 Wissenschaftliches Programmieren

Fortgeschrittene Anwender*innen von OTS-Programmen kommen schnell an den Punkt, an denen die Software nicht mehr alle ihre Bedarfe nach Vorverarbeitung der Daten, spezifischen Auswertungsabläufen oder visuellen Ergebnisdarstellungen erfüllt. In diesem Fall empfiehlt es sich, einzelne Forschungsschritte der Datenerhebung, Auswertung und Visualisierung oder den gesamten Ablauf selbst zu programmieren. Hierfür stehen Open-Source-Programmierumgebungen und eine große Menge gut dokumentierter Softwarebibliotheken bereit, welche die maßgeschneiderte Entwicklung komplexer Analyseabläufe mit State-of-the-Art-Methoden möglich machen. Der Ansatz, Daten und Analysen über Programmskripte und öffentliche Repositorien verfügbar zu machen, stellt die Anforderungen an Nachvollziehbarkeit und Reproduzierbarkeit eines Forschungsablaufs bestmöglich sicher.

Python[18] und R[19] sind die am häufigsten verwendeten Programmiersprachen in den CSS. Beide Sprachen können die meisten Aufgaben bewältigen. Für die Auswahl spielt die Verfügbarkeit bestimmter Programmbibliotheken, die zum Einsatz kommen sollen, aber auch persönliche Vorliebe eine wichtige Rolle. Python ist eine allgemeine Programmiersprache, die flexibler sein kann und

[18] https://www.python.org/, zuletzt geprüft am 24.06.2024.
[19] https://www.r-project.org, zuletzt geprüft am 24.06.2024.

vor allem im Bereich maschinelles Lernen mit neuronalen Netzen einen Vorsprung vor R hat. R ist eine statistische Programmiersprache, die auf die Manipulation von Vektoren und Faktoren spezialisiert ist und damit eine einfachere Syntax für bestimmte Aufgaben zur Verfügung stellt. Beide Sprachen werden jedoch immer ähnlicher, da Module und Pakete ähnliche Funktionen bieten. Quellcode für ein Forschungsprojekt kann im Team sowie öffentlich mit der gesamten Forschungscommunity über eine Software wie GitLab[20] oder ein darauf aufbauendes Portal wie GitHub[21] geteilt werden. Git-Tools ermöglichen die Versionskontrolle und das kollaborative Arbeiten an Quellcode und stellen Werkzeuge für Projektmanagement bereit, mit denen sich Softwareprojekte jeder Größe managen lassen. Eine gute Einführung in wissenschaftliches Programmieren sowohl mit Python als auch mit R bieten van Atteveldt, Trilling und Calderón (2022) als Open Access Buch an.[22]

5.3.1 Interaktive Notebooks

Für die Unterstützung wissenschaftlicher Datenauswertung mit R oder Python empfiehlt sich die Nutzung interaktiver Dokument-Formate, sogenannte Notebooks. Diese erlauben die Kombination von Textbeschreibungen, Layoutanweisungen und ausführbaren Programmteilen. Die Programmteile können Ergebnisse wie einzelne Zahlenwerte, Tabellen oder komplexe Grafiken erzeugen, die wiederum selbst direkt ins Dokument eingebunden werden können. Über Zitationsanweisungen können wissenschaftliche Vorarbeiten, auf denen eine Auswertung beruht, wie beispielsweise eine bestimmte verwendete Analysemethode, einfach im Text referenziert und automatisch in ein Quellenverzeichnis eingebunden werden. Auf diese Weise ist es möglich eine Vielzahl von Dokumentarten (Webseiten, HTML- und PDF-Dokumente, Präsentationen) für die Vermittlung komplexer CSS-Analyseprozesse und daraus abgeleiteter Ergebnisse einfach zu erstellen und für andere Forschende vollständig reproduzierbar zu kommunizieren. So wird beispielsweise das bereits

[20] https://about.gitlab.com, zuletzt geprüft am 24.06.2024.
[21] https://github.com, zuletzt geprüft am 24.06.2024.
[22] http://cssbook.net, zuletzt geprüft am 24.06.2024.

oben erwähnte Buch „Computational Analysis of Communication“ (van Atteveldt, Trilling und Calderón 2022) öffentlich auf GitHub mit dem interaktiven Dokument-System quarto[23] kollaborativ erstellt[24]. Wiedemann und Niekler (2017) stellen die Inhalte eines Einführungskurses zum Text Mining für Sozialwissenschaftler*innen in der Programmiersprache R als Sammlung von R-Notebooks bereits.[25] Für Python-basierte Skripte können funktional ähnliche Notebooks mit der interaktiven Webapplikation Jupyter erstellt werden.[26] Für einen einfachen Einstieg in die Arbeit mit diesen Technologien, beispielsweise in Lehrveranstaltungen, stellen Anbieter Cloud-basierte Services zur Verfügung, mit denen R-Skripte / Notebooks und Jupyter Notebooks auf virtuellen Rechnern gestartet und ausgeführt werden können.[27]

5.3.2 Interaktive Dashboards

Die visuelle Darstellung und Exploration von Kennzahlen ist in der Auswertung sehr großer Datenbestände regelmäßig einer der wichtigsten Schritte im Forschungsprozess. Zur Unterstützung dieses Schrittes werden in CSS-Forschungsabläufen häufig interaktive Dashboards erstellt, auf denen aggregierte Rohdaten als Zusammenstellung informativer Grafiken dargestellt und bei fortlaufenden Datensammlungen in Echtzeit aktualisiert werden können. Mit Hilfe von Filtern lassen sich durch die Nutzer*in Datenausschnitte entlang bestimmter Dimensionen der Daten selektieren, mit denen sich dann beispielsweise verschiedene Subgruppen miteinander vergleichen lassen. Gene-rische Dashboard-Applikationen wie Apache Superset[28], Graphana[29] oder Kibana[30] lassen sich direkt an Datenbanken anbinden und bieten eine breite Auswahl

[23] https://quarto.org, zuletzt geprüft am 24.06.2024.
[24] https://github.com/vanatteveldt/cssbook, zuletzt geprüft am 24.06.2024.
[25] https://tm4ss.github.io, zuletzt geprüft am 24.06.2024.
[26] https://jupyter.org, zuletzt geprüft am 24.06.2024.
[27] Beispielsweise https://mybinder.org als freier Service von den Jupyter-Entwickler*innen oder https://posit.cloud als kommerzielles Angebot der Entwickler der R-Programmierumgebung RStudio, alle Links zuletzt geprüft am 24.06.2024.
[28] https://superset.apache.org, zuletzt geprüft am 24.06.2024.
[29] https://grafana.com, zuletzt geprüft am 24.06.2024.
[30] https://www.elastic.co/de/kibana, zuletzt geprüft am 24.06.2024.

vorgefertigter Module zur Auswertung und visuellen Darstellung. Mit Frameworks wie R-Shiny[31], Streamlit[32] oder Plotly Dash[33] lassen sich darüber hinaus einfache Dashboards bis hin zu komplexen Data Analytics Apps nach projektspezifischen Anforderungen selbst programmieren.

5.3.3 Hilfestellung durch KI

Der Einstieg in die wissenschaftliche Programmierung zur Datenauswertung kann für Sozialwissenschaftler*innen, die in ihrer Ausbildung erst seit wenigen Jahren standardmäßig mit Programmcode in Berührung kommen, eine größere Herausforderung darstellen. Die den Programmiersprachen jeweils spezifische Syntax, das Verständnis für Datenobjekte, die Konzeption von Funktionen und Auswertungsalgorithmen sowie die Nutzung vorhandener Softwarebibliotheken erfordern einen hohen Einsatz auch für fortgeschrittene Anwender*innen. In naher Zukunft ist hierbei eine erhebliche Erleichterung durch KI-Agenten zu erwarten, die entweder über externe Services aufgerufen oder direkt in die Entwicklungsumgebungen integriert genutzt werden können. Services wie ChatGPT[34] oder die darauf basierende, neue Version des GitHub Copilot[35] sind unter anderem in der Lage, Programmcode auf Basis verbaler Beschreibungen zu generieren, Fehler in Programmcodes zu finden und verbessern sowie Programmteile aus einer Programmiersprache in eine andere zu übersetzen. Es ist zu erwarten, dass sich Zugänglichkeit zu CSS-Methoden durch diese Technologien erheblich erleichtert wird.

5.4 Fallbeispiel: ‚Metaverse'-Diskurs

In diesem Kapitel werden Ansätze der automatisierten Inhaltsanalyse für große Datenmengen exemplarisch anhand einer Auswertung des Twitter-Diskurses

[31] https://shiny.rstudio.com, zuletzt geprüft am 24.06.2024.
[32] https://streamlit.io, zuletzt geprüft am 24.06.2024.
[33] https://plotly.com/dash, zuletzt geprüft am 24.06.2024.
[34] https://chat.openai.com, zuletzt geprüft am 24.06.2024.
[35] https://github.com/features/copilot, zuletzt geprüft am 24.06.2024.

zum ‚Metaverse' vorgestellt. Der Begriff Metaverse beschreibt eine Menge virtueller Welten, die in einem virtuellen Universum miteinander verbunden sind. Der grundlegenden Idee zufolge ermöglichen die vernetzten Welten immersive Erfahrungen, die von Menschen in Echtzeit erlebt und interagiert werden können. Es ist ein Konzept, das von Science-Fiction-Literatur und Filmen inspiriert wurde und von Technologieunternehmen wie Facebook als nächste Stufe der digitalen Evolution betrachtet wird. Facebook CEO Mark Zuckerberg hat im Oktober 2021 öffentlich seine Vision des Metaverse dargelegt samt seiner Einschätzung, dass es sich hierbei um den größte technischen Entwicklungsfortschritt seit Einführung des Internets handele. Seit dieser Ankündigung firmiert der Facebook-Konzern unter dem Namen „Meta" und lenkt Investitionen in die Technologieentwicklung, um Metaverse zu einem wichtigen Teil seiner Plattformen zu machen.

Wie wird diese Vision in verschiedenen Diskurs-Gemeinschaften rezipiert? Was verstehen Menschen in Deutschland unter dem Begriff Metaverse und wie bewerten sie diese Entwicklung? Diese Fragen können unter anderem durch die Beobachtung und Auswertung öffentlicher Debatten beantwortet werden. Die Social-Media-Plattform Twitter versammelt Medienakteure, Expert*innen, Unternehmen und potenzielle Nutzer*innen des Metaverse (vgl. Kapitel 2.2.1). Zwischen dem 01.10.2021 und dem 31.10.2022 sind auf Twitter ca. 102.000 deutschsprachige Tweets (ohne Retweets) verfasst worden, welche entweder den Begriff ‚Metaverse' oder ‚Metaversum' enthalten. Zusätzlich sind ca. 14.000 Antworten (replies) auf diese Tweets verfasst worden. Tweets und Antworten wurden über die Academic API mit dem Programm twarc[36] heruntergeladen und für eine explorative Analyse in einem interaktiven Dashboard mit semantischen Kategorien angereichert. Eine erste Sichtung der Daten ergab, dass der Begriff Metaverse extrem häufig für sogenanntes Hashtag-Spamming genutzt wird. Dabei werden populäre Hashtags an Tweets verwendet, um Recommender-Systeme dazu zu bringen, beliebige andere Inhalte (z. B. Werbelinks) auf der Plattform sichtbarer zu machen. Um solche Tweets aus der Analyse möglichst auszuschließen, wurden Tweets mit drei oder mehr

[36] https://twarc-project.readthedocs.io, zuletzt geprüft am 24.06.2024.

aufeinanderfolgenden Hashtags als Spam definiert und aus dem Datensatz entfernt. Für eine Topic Model Analyse wurden aus den verbleibenden ca. 45.000 Tweets Nutzernamen (@mentions), URLs, NFT-Tokens, Stoppworte und Satzzeichen entfernt und der verbleibende Text in Kleinbuchstaben umgewandelt. Mit der Software BERTopic (Grootendorst 2022) wurden die Tweets in 32 thematische Cluster unterteilt, die anhand ihrer Top 10 spezifischsten Begriffe inhaltlich beschrieben sind. Zudem wurden die Tweets mit einem deutschsprachigen Sentiment-Klassifikationsmodell in positiv, negativ und neutral klassifiziert (Guhr et al. 2020). Die Ergebnisse aus dem Topic Modeling und der Sentiment-Klassifikation wurden mit der Software Apache Superset zusammen mit den Metadaten der Tweets in einem interaktiven Dashboard zusammengeführt, dass die Auswertung von Themenzusammensetzungen, deren emotionale Bewertungen, in der Debatte aktive Nutzer*innen und externe Verlinkungen über den Verlauf des Beobachtungszeitraums gestattet (vgl. Abbildung 5.1).[37]

Eine erste Auswertung mit Hilfe des Dashboards zeigt auf, dass die Metaverse-Debatte tatsächlich erst mit Zuckerbergs Ankündigungen zur Zukunft des Facebook-Konzerns an Fahrt aufnimmt, die Aufmerksamkeit in der zweiten Jahreshälfte 2022 aber deutlich nachlässt. Thematisch dominieren technische Aspekte insbesondere mit Bezug zu virtueller Realität, Crypto-Währungen und sogenannten Non-Fungible Tokens, mit denen unter anderem im Metaverse bezahlt und Eigentum an digitalen Kunstwerken deklariert werden kann. Weniger stark präsente Themen sind unter anderem erwartete Auswirkungen des Metaverse auf Arbeits- und Wirtschaftsbeziehungen, Gaming oder Sexualität. Die Sentimentanalyse zeigt, dass die Debatte von Anfang deutlich stärker von negativen als von positiven Kommentaren geprägt ist. Positive Tweets haben oft einen Werbecharakter und verlinken auf externe, kommerzielle Angebote. Negative Tweets bezweifeln häufig den Sinn der Technologie oder machen sich gar in sarkastischer Weise über die Vision Zuckerbergs lustig. Bei aktiven Teilnehmer*innen der Debatte lassen sich zwei Gruppen unterscheiden. Die aktivsten Nutzer*innen des Hashtags scheinen den Begriff vor allem

[37] https://superset.notorious-project.com/superset/dashboard/metaverse/, zuletzt geprüft am 24.06.2024.

zu nutzen, um die Themen Kryptowährungen und Blockchain in ihrem Interesse zu pushen. Eine Stichprobe der aktivsten Accounts lässt Muster von unauthentischem Nutzer*innenverhalten erkennen, das auf Social Bots schließen lässt. Eine interessantere Auswahl aktivster Nutzer liefert die Filterung nach verifizierten Accounts. Hier wird sichtbar, dass redaktionelle Medien wie t3n, Horizont und Heise online die deutsche Metaverse-Debatte maßgeblich prägen. Dies sind dann auch die Webseiten, auf die neben anderen sozialen Medienplattformen und NFT-Art Brokern am meisten extern verlinkt wird. Diese vorläufigen Erkenntnisse bedürfen einer sorgfältigen Validierung anhand qualitativer Ausschnitte aus Datenmaterial sowie einer Vertiefung der Betrachtung einzelner Themen, die effizient mit Hilfe des Dashboards weiter vorgenommen werden kann.

Abbildung 5.1: Interaktives Dashboard (Apache Superset) zur Auswertung von ca. 102.000 deutschsprachigen Tweets mit Bezug zum ‚Metaverse' bzw. ‚Metaversum' im Zeitraum Oktober 2021 bis Oktober 2022.

6 Lernen und Lehren

Mittlerweile existiert eine Vielzahl an Literatur, Lerninhalten, Anleitungsvideos und Online-Kursen zum Einstieg in CSS-Methoden. Die folgenden Listen geben einen (unvollständigen) Überblick über weitere Quellen zum Einstieg in die CSS.

6.1 Einführende Bücher

Folgende Monografien und Sammelbände geben einen Überblick über die aktuelle CSS-Forschungslandschaft bzw. einen Einstieg in konkrete Methoden wie der automatischen Inhaltsanalyse.

- Alvarez, Michael (2016). Computational Social Science: Discovery and Prediction. Cambridge UK: Cambridge University Press. https://doi.org/10.3917/rfsp.674.0740a
- Biemann, Chris; Heyer, Gerhard; Quasthoff, Uwe (2022): Wissensrohstoff Text. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-35969-0
- Blätte, Andreas; Behnke, Joachim; Schnap, Kai-Uwe; Wagemann, Claudius (Hg.) (2018): Computational Social Science: Die Analyse von Big Data. Baden-Baden: Nomos. https://doi.org/10.5771/9783845286556
- Cioffi-Revilla, Claudio (2017): Introduction to Computational Social Science: Principles and Applications. Cham: Springer. https://doi.org/10.1007/978-3-319-50131-4
- Engel, Uwe; Quan-Haase, Anabel; Liu, Sunny Xun; Lyberg, Lars (Hg.) (2021): Handbook of Computational Social Science. London: Routledge. https://doi.org/10.4324/9781003024583

- Grimmer, Justin; Roberts, Margaret; Stewart, Brandon (2022): Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton; Oxford: Princeton University Press. https://doi.org/10.3917/res.238.0331
- Ignatow, Gabe; Mihalcea, Rada (2017): An Introduction to Text Mining: Research Design, Data Collection, and Analysis. Thousand Oaks: Sage. https://doi.org/10.4135/9781506336985
- Jünger, Jakob; Gärtner, Chantal (2022): Computational Methods für die Sozial- Und Geisteswissenschaften. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-37747-2
- Robinson, David; Silge, Julia (2017) Text Mining with R: A Tidy Approach. https://www.tidytextmining.com/, zuletzt geprüft am 01.07.2024.
- Stützer, Cathleen; Welker, Martin; Egger, Marc (Hg.) (2018): Computational Social Science in the Age of Big Data: Concepts, Methodologies, Tools, and Applications. Köln: Herbert von Halem.
- van Atteveldt, Wouter; Trilling, Damian; Arcíla Calderón, Carlos (2022): Computational analysis of communication. Hoboken (NJ): Wiley.

6.2 Link- und Ressourcensammlungen

Folgende Linksammlung stellt eine kleine Auswahl von Software- und Lernressourcen sowie Beispielprojekte aus der CSS-Forschung dar.

- https://wiki.digitalmethods.net: Die Digital Methods Initiative (DMI), Amsterdam veranstaltet jährliche Summer- und Winter-Schools und verweist auf zahlreiche relevante Projekte im CSS-Bereich.
- https://github.com/strohne/cm enthält ausführliches Begleitmaterial (R-Skripte) zum Open Access Buch Jünger und Gärtner (2022).
- https://github.com/Tarlanc/CCSL: Eine offene Sammlung von Lernressourcen und Lehrkonzepten der Arbeitsgruppe „Computational Communication Science in der Lehre“ der DGPuK.

- https://smo-wiki.leibniz-hbi.de: Das Wiki des Social Media Observatory am Hans-Bredow-Institut kuratiert Übersichten für CSS-Forschungssoftware und Tutorials für Datenanalysen.
- https://ladal.edu.au: Das Language Technology and Data Analysis Laboratory der University of Queensland pflegt eine umfangreiche Sammlung von R-Tutorials.
- https://tm4ss.github.io enthält Beispielskripte aus einem 5-tägigen Kurs zur Einführung in Text Mining mit R.
- https://sicss.io: Die Summer Institutes in Computational Social Science stellen eine große Sammlung von Videos zum Online-Lernen bereit.
- http://inhaltsanalyse-mit-r.de enthält umfangreiche Tutorials zur Einführung in die automatisierte Inhaltsanalyse.
- https://bookdown.org/joone/ComputationalMethods enthält das vollständige Material des Seminars „Methodische Vertiefung: Computational Methods mit R und RStudio“ (WiSe 2020) von Julian Unkel an der LMU.

6.3 Digitale Beteiligungsformate

Mit Citizen Science Projekten, Hackathons, Datathons und Shared Tasks werden zunehmend Formate digitaler Partizipation in der CSS-Forschung genutzt, die neben anderen Forscher*innen vor allem auch Bürger*innen einbeziehen.

- https://www.buergerschaffenwissen.de/projekte listet aktuelle Projekte, die Bürger*innen in die Datenerhebung, -Kodierung oder Auswertung einbeziehen.

- https://github.com/uhh-lt/news-speaker-attribution-2023: Der Arbeitskreis Computerlinguistik für die Politik- und Sozialwissenschaften lädt im Rahmen der Tagung KONVENS 2023[1] zu einer shared task „Speaker Attribution in Newswire and Parliamentary Debates“ ein, bei der interessierte Forschende an einem Wettbewerb um das bestmögliche System zur automatischen Erkennung von Sprecherkategorien teilnehmen können.
- https://www.bundesregierung.de/breg-de/themen/coronavirus/hackathon-der-bundesregierung-1733632: Die Bundesregierung veranstaltete 2020 unter dem Titel „#WirVersusVirus“ einen Hackathon, bei dem es um die Entwicklung von digitalen Anwendungen und Forschungstools zur Bekämpfung der Corona-Pandemie ging. Ca. 23.000 Menschen beteiligten sich ein Wochenende lang an über 1.500 Projekten.
- https://wirfuerschule.de/hackathon-2021/: Das BMBF veranstaltete 2021 unter dem Titel #WirFürSchule einen weiteren Hackathon mit dem Schwerpunkt digitale Bildungstransformation (alle Links zuletzt geprüft am 24.06.2024).

[1] https://www.thi.de/konvens-2023/, zuletzt geprüft am 24.06.2024.

7 Zusammenfassung

Die digitale Großtransformation der Gesellschaft und erfordert auch Anpassungen der Prozesse zu deren Erforschung. Die Etablierung und Weiterentwicklung von computergestützten Methoden aus dem Bereich der Informatik in der Sozialwissenschaft ist eine notwendige Antwort auf diese Herausforderung. Dabei kann der Einsatz computergestützter Methoden auf unterschiedlich große Datenmengen skaliert und mit herkömmlichen Methoden der empirischen Sozialforschung kombiniert werden. Insbesondere die Verbindung aus digitalen Spurendaten und Befragungsdaten ermöglicht die Beantwortung neuer Forschungsfragen, die mit einem Methodenzweig einzeln nicht zu bearbeiten wären. Für den Einstieg in CSS-Methoden kann auf eine überschaubare Auswahl an ‚off-the-shelf'-Forschungssoftware zurückgegriffen werden, solange die Datenmengen überschaubar und die Analyseanforderungen an Standardabläufen orientiert bleiben. Für sehr große Datenmengen und projektspezifische Analyseanforderungen wird jedoch die Entwicklung eigener Auswertungsskripte und Analyseumgebungen notwendig, die in der Regel fortgeschrittene Programmierkenntnisse erfordern, einen hohen Aufwand und Kosten verursachen. Hierzu ist der Aufbau von Expertise einzelner Forscher*innen oder in einem Team ein mittel- bis langfristiges Vorhaben, bei dem den Beteiligten genug Zeit und Ressourcen für methodische Weiterbildung eingeräumt werden sollte.

Offen ist derzeit die Frage, wie sich der Forschungszugang zu den großen (sozialen) Medien-Plattformen entwickelt. Twitter/X plant, den Zugang zu seinen (Forschungs-)APIs kostenpflichtig zu machen, TikTok kündigt dagegen einen neuen, umfangreichen und freien Forschungsdatenzugang an. Gleichzeitig schreiben die Regulierungsanforderungen durch den Digital Services Act einen Forschungszugang zu den Plattformen rechtsverbindlich vor. Es bleibt abzuwarten, wie diese Forderung künftig in der Praxis umgesetzt wird.

Für die TA liegen die größten Potenziale bei der Nutzung digitaler Methoden in Erforschung von Expert*innendiskursen sowie deren Rezeption in öffentlichen Diskursen redaktionellen Massenmedien und deren Nutzerkommentaren, Foren, Blogs und Webseiten sowie verschiedenen sozialen Medienplattformen. Darüber hinaus eröffnen sich Möglichkeiten für systematische, großangelegte Reviews und Surveys des sich immer schneller entwickelnden wissenschaftlichen Diskurses. Hierbei ist zu erwarten, dass nicht nur KI-Tools in naher Zukunft einen erheblichen Einfluss auf die Transformation von Sozial- und Arbeitsbeziehungen in der Gesellschaft haben werden, sondern diese auch für die Erforschung dieses Wandels nutzbar gemacht werden können. Das zunehmend besser funktionierende Text- und Bildverstehen auf Basis neuronaler Netze ermöglicht den Anschluss von quantitativen Verfahren, welche die Verbreitung von Aussagen im Diskurs messen, an qualitative, hermeneutische Forschungsprozesse, die intersubjektiv nachvollziehbare Interpretationen und ein tiefes Verständnis des semantischen Bedeutungsgehaltes zentraler Aussagen zum Ziel haben (Wiedemann 2013). Durch die Modellierung komplexer semantischer Zusammenhänge in zeitgenössischen Sprachmodellen eignen sich computergestützte Analyseverfahren auch für kleine und mittlere Datenmengen wie Transkripte von Interviews, Gesprächen in Fokusgruppen und offenen Survey-Fragen. Mit Verfahren der Netzwerkanalyse, des Predictive Modeling und der Causal Inference lassen sich Beobachtungen in digitalen Spurendaten über soziale Beziehungen erklären und zu einem gewissen Grad vorhersagen. Besonders spannend, aber noch weitgehend unerschlossen für die TA, könnten Ansätze zur Simulation komplexer Systeme dazu beitragen, verschiedene Szenarien der Auswirkungen neuer Technologien auf Strukturen der Gesellschaft zu modellieren. Gleichzeitig sollten die Potenziale zur Kausalitätsmessung bzw. zur Vorhersage nicht überschätzt werden.

Für den Einstieg in die Nutzung digitaler Methoden ist es möglich entweder Projekte in interdisziplinäre Teams gemeinsam mit Entwickler*innen spezifischer Auswertungsmethoden in der Informatik bzw. Statistik zu entwickeln oder eine Kooperation mit sozialwissenschaftlichen Partner*innen einzugehen, welche selbst bereits Erfahrung im CSS-Bereich mitbringen. Für den Anfang erscheint letzter Ansatz eher empfehlenswert, da der Aufwand in in-

terdisziplinärer Kooperation ein gemeinsames Problemverständnis, eine gemeinsame Sprache und ein füreinander fruchtbares, methodisches Vorgehen zum Gegenstand der Forschung zu finden nicht unterschätzt werden sollte. Die informatische Seite einer Kooperation betrachtet ihre Arbeit häufig als erledigt an einer Stelle, an der es für den sozialwissenschaftlichen Erkenntnisgewinn erst richtig interessant wird. Insofern ersetzt der Einsatz computergestützter Verfahren weder die Notwendigkeit sorgfältiger Validierung noch qualitativer Interpretation der auf Basis sehr großer Datenmengen abgeleiteten Erkenntnisse.

Literatur

Adamic, Lada (1999): „The small world web." In: Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science 1696, hg. v. Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Serge Abiteboul und Anne-Marie Vercoustre, S. 443–452. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-48155-9_27

Albrecht, Steffen (2024): „ChatGPT als doppelte Herausforderung für die Wissenschaft. Eine Reflexion aus der Perspektive der Technikfolgenabschätzung." Gerhard Schreiber und Lukas Ohly (Hg.): KI:Text. Diskurse über KI-Textgeneratoren. Berlin: De Gruyter, S. 13–28. https://doi.org/10.1515/9783111351490-003

Amaya, Ashley; Bach, Ruben; Kreuter, Frauke; Keusch, Florian (2020): „Measuring the strength of attitudes in social media data." In: Big Data Meets Survey Science, hg. v. Craig Hill et al., S. 163–192. Hoboken (NJ): Wiley. https://doi.org/10.1002/9781118976357.ch5

Andres, Raphaela; Slivko, Olga (2021): „Combating online hate speech: The impact of legislation on Twitter." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4013662

Araujo, Theo et al. (2022): „Osd2f: An open-source data donation framework." Computational Communication Research 4 (2), S. 372–387. https://doi.org/10.5117/CCR2022.2.001.ARAU

Araujo, Theo; Lock, Irina; van de Velde, Bob (2020): „Automated visual content analysis (AVCA) in communication research: A protocol for large scale image classification with pre-trained computer vision models." Communication Methods and Measures 14 (4), S. 239–265. https://doi.org/10.1080/19312458.2020.1810648

Barbaresi, Adrien (2021): „Trafilatura: A web scraping library and command-line tool for text discovery and extraction." In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, hg. v. Heng Ji, Jong C. Park und Rui Xia, 122–31. Stroudsburg (PA): Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-demo.15

Bastian, Mathieu; Heymann, Sebastien; Jacomy, Mathieu (2009): „Gephi: An open source software for exploring and manipulating networks." In: Proceedings of the International AAAI Conference on Web and Social Media 3 (1), S. 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Benoit, Kenneth et al. (2018): „Quanteda: An R package for the quantitative analysis of textual data." Journal of Open Source Software 3 (30), S. 774. https://doi.org/10.21105/joss.00774

Ben-Zeev, Dror; Scherer, Emily; Wang, Rui; Xie, Haiyi; Campbell, Andrew (2015): „Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health." Psychiatric Rehabilitation Journal 38 (3), S. 218–226. https://doi.org/10.1037/prj0000130

Biemann, Chris; Heyer, Gerhard; Quasthoff, Uwe (2022): Wissensrohstoff Text. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-35969-0

Bird, Steven; Klein, Ewan; Loper, Edward (2009): Natural language processing with python: Analyzing text with the natural language toolkit. Beijing, Köln: O'Reilly.

Blaette, Andreas; Leonhardt, Christoph (2022): „GermaParl corpus of plenary protocols." Zenodo. https://doi.org/10.5281/ZENODO.6539967

Boberg, Svenja; Quandt, Thorsten; Schatto-Eckrodt, Tim; Frischlich, Lena (2020): „Pandemic populism: Facebook pages of alternative news media and the Corona crisis. A computational content analysis." arXiv. https://arxiv.org/abs/2004.02566, zuletzt geprüft am 04.06.2024.

Boeschoten, Laura; Ausloos, Jef; Möller, Judith; Araujo, Theo; Oberski, Daniel (2022): „A framework for privacy preserving digital trace data collection through data donation." Computational Communication Research 4 (2), S. 388–423. https://doi.org/10.5117/CCr2022.2.002.BoEs

Böschen, Stefan; Grunwald, Armin; Krings, Bettina-Johanna; Rösch, Christine (2021): Technikfolgenabschätzung: Handbuch für Wissenschaft und Praxis. Baden-Baden: Nomos.

Böschen, Stefan; Grunwald, Armin; Krings, Bettina-Johanna; Rösch, Christine (2021): „Technikfolgenabschätzung – neue Zeiten, neue Aufgaben." In: Technikfolgenabschätzung. Handbuch für Wissenschaft und Praxis, hg. v. Stefan Böschen, Armin Grunwald, Bettina-Johanna Krings und Christine Rösch. Baden-Baden: Nomos, S. 15-40.

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

Castro, Juana et al. (2020): „A review of agent-based modeling of climate-energy policy.“ WIREs Climate Change 11 (4), S. e647. https://doi.org/10.1002/wcc.647

Deutscher Ethikrat (2023): „Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz. Stellungnahme.“ Online verfügbar unter https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf, zuletzt geprüft am 13.03.2024.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2019): „BERT: Pre-training of deep bidirectional transformers for language understanding.“ In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, S. 4171–4186. Minneapolis (MN): Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Djanatliev, Anatoli; German, Reinhard; Kolominsky-Rabas, Peter; Hofmann, Bernd (2012): „Hybrid simulation with loosely coupled system dynamics and agent-based models for prospective health technology assessments.“ In: Proceedings of the 2012 Winter Simulation Conference, S. 1–12. Berlin: IEEE. https://doi.org/10.1109/WSC.2012.6465024

Dubovi, Ilana; Iris Tabak (2020): „An empirical analysis of knowledge co-construction in YouTube comments.“ Computers & Education 156, S. 103939. https://doi.org/10.1016/j.compedu.2020.103939

Eck, Detlev; Hensel, Gerd; Kappei, Günther (2020): Wissensmanagement von Altdokumenten aus Forschung, Verwaltung Und Betrieb: Schlussbericht. Neuherberg: Helmholtz Zentrum München.

Elevelt, Anne; Lugtig, Peter; Toepoel, Vera (2019): „Doing a time use survey on smartphones only: What factors predict nonresponse at different stages of the survey process?“ Survey Research Methods 13 (2), S. 195–213. https://doi.org/10.18148/srm/2019.v13i2.7385

Engel, Uwe (2021): „Causal and predictive modeling in computational social science.“ In: Handbook of Computational Social Science, hg. v. Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu und Lars Lyberg, S. 131–149. London: Routledge. https://doi.org/10.4324/9781003024583-10

Engl, Elisabeth (2020): „OCR-d Kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative.“ Bibliothek Forschung und Praxis 44 (2), S. 218–230. https://doi.org/10.1515/bfp-2020-0024

Evers, Jeanine; Caprioli, Mauro; Nöst, Stefan; Wiedemann, Gregor (2020): „What Is the REFI-QDA Standard: Experimenting with the transfer of analyzed research projects between QDA software.“ Forum Qualitative Sozialforschung 21 (2), S. 1—37. https://doi.org/10.17169/fqs-21.2.3439

Fedtke, Cornelia; Wiedemann, Gregor (2020): „Hass- und Gegenrede in der Kommentierung Massenmedialer Berichterstattung auf Facebook: Eine Computergestützte kritische Diskursanalyse.“ In: Soziale Medien? Interdisziplinäre Zugänge Zur Onlinekommunikation, hg. v. Peter Klimczak, Christer Petersen und Samuel Breidenbach, 91–120. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-30702-8_5

Gilardi, Fabrizio; Alizadeh, Meysam; Kubli, Maël (2023): „ChatGPT outperforms crowd-workers for text-annotation tasks.“ In: Proceedings of the National Academy of Sciences 120 (30), S. e2305016120. https://doi.org/10.1073/pnas.2305016120

Gilbert, Nigel (2004): „Agent-based social simulation: dealing with complexity.“ https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e372a468ad80e71b0cac1dd9ea9bf20c3fedd6e3, zuletzt geprüft am 06.06.2024.

Gold, Nicolas (2009): „Service-oriented software in the humanities: A software engineering perspective.“ Digital Humanities Quarterly 3 (4). http://www.digitalhumanities.org/dhq/vol/3/4/000072/000072.html, zuletzt geprüft am 06.06.2024.

Grimmer, Justin (2010): „A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases.“ Political Analysis 18 (1), S. 1–35. https://doi.org/10.1093/pan/mpp034

Grimmer, Justin; Roberts, Margaret; Stewart, Brandon (2022): Text as data: A new framework for machine learning and the social sciences. Princeton; Oxford: Princeton University Press.

Grimmer, Justin; Stewart, Brandon (2013): „Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political Analysis 21 (3), S. 267–97. https://doi.org/10.1093/pan/mps028

Grootendorst, Maarten (2022): „BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv. https://arxiv.org/pdf/2203.05794, zuletzt geprüft am 06.06.2024.

Guhr, Oliver; Schumann, Anne-Kathrin; Bahrmann, Frank; Böhme, Hans Joachim (2020): „Training a broad-coverage German sentiment classification model for dialog systems." In: Proceedings of the Twelfth Language Resources and Evaluation Conference, 1627–32. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.202, zuletzt geprüft am 06.06.2024.

Guo, Lei et al. (2023): „Proposing an open-sourced tool for computational framing analysis of multilingual data." Digital Journalism 11 (2), S. 276–297. https://doi.org/10.1080/21670811.2022.2031241

Haim, Mario; Jungblut, Marc (2021): „Politicians' self-depiction and their news portrayal: Evidence from 28 countries using visual computational analysis." Political Communication 38 (1-2), S. 55–74. https://doi.org/10.1080/10584609.2020.1753869

Haunss, Sebastian et al. (2020): „Integrating manual and automatic annotation for the creation of discourse network data sets." Politics and Governance 8 (2), S. 326–339. https://doi.org/10.17645/pag.v8i2.2591

Heaton, Dan; Clos, Jeremie; Nichele, Elena; Fischer, Joel (2024): „'The ChatGPT bot is causing panic now – but it'll soon be as mundane a tool as excel': Analysing topics, sentiment and emotions relating to ChatGPT on Twitter." Personal and Ubiquitous Computing, May. https://doi.org/10.1007/s00779-024-01811-x

Hepp, Andreas (2022): „Jenseits der Disruption: Zum Lebenszyklus von Pioniergemeinschaften und ihrer Rolle beim Entstehen einer digitalen Gesellschaft." KZfSS – Kölner Zeitschrift für Soziologie und Sozialpsychologie 74 (S1), S. 231–255. https://doi.org/10.1007/s11577-022-00835-6

Horton, John; Rand, David; Zeckhauser, Richard (2011): „The online laboratory: Conducting experiments in a real labor market." Experimental Economics 14 (3), S. 399–425. https://doi.org/10.1007/s10683-011-9273-9

Howison, James; Wiggins, Andrea; Crowston, Kevin (2011): „Validity issues in the use of social network analysis with digital trace data." Journal of the Association for Information Systems 12 (12), S. 767–797. https://doi.org/10.17705/1jais.00282

Jünger, Jakob (2021): „A brief history of APIs." In: Handbook of Computational Social Science 2, hg. v. Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu und Lars Lyberg, 17–32. London: Routledge. https://doi.org/10.4324/9781003025245-3

Jünger, Jakob; Gärtner, Chantal (2022): Computational Methods für die Sozial- Und Geisteswissenschaften. Wiesbaden: Springer VS.

Keusch, Florian; Kreuter, Frauke (2021): „Digital trace data." In: Handbook of Computational Social Science 1, hg. v. Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu und Lars Lyberg, 100–118. London: Routledge. https://doi.org/10.4324/9781003024583-8

Kieslich, Kimon; Došenović, Pero; Marcinkowski, Frank (2022): „Alles, nur kaum Science-Fiction: Eine Themenanalyse der Deutschen Medienberichterstattung über künstliche Intelligenz." https://www.cais-research.de/wp-content/uploads/Factsheet-7-Medienberichterstattung.pdf, zuletzt geprüft am 06.06.2024.

Kiess, Johannes; Nissen, Sophie; Wetze, Gideon; Winkler, Benjamin (2022): „EFBI digital report #0: Pilotausgabe." https://efbi.de/files/efbi/pdfs/Digital%20Reports/2022-0-EFBI_DigitalReport_final.pdf, zuletzt geprüft am 06.06.2024.

Koch, Wolfgang (2022): „Reichweiten von Social-Media-Plattformen und Messengern: Ergebnisse der ARD/ZDF-Onlinestudie 2022." Media Perspektiven (10), S. 471–478. https://www.ard-zdf-onlinestudie.de/files/2022/2210_Koch.pdf, zuletzt geprüft am 06.06.2024.

Larsson, Anders (2021): „Picture-perfect populism: Tracing the rise of European populist parties on Facebook." New Media & Society 24 (1), S. 227-245. https://doi.org/10.1177/1461444820963777

Lazer, David et al. (2009): „Computational social science." Science 323 (5915), S. 721–723. https://doi.org/10.1126/science.1167742

Lazer, David et al. (2021): „Meaningful measures of human society in the twenty-first century." Nature 595 (7866), S. 189–196. https://doi.org/10.1038/s41586-021-03660-7

Leiter, Christoph et al. (2024): „ChatGPT: A meta-analysis after 2.5 months." Machine Learning with Applications, S. 100541. https://doi.org/10.1016/j.mlwa.2024.100541

Maas, Martina (2016): „Ungleichheitsdeutungen im Medialen Bildungsdiskurs. Eine Analyse des PISA-Diskurses in Deutschland." In: Text Mining in den Sozialwissenschaften, hg. v. Matthias Lemke und Gregor Wiedemann, S. 227–256. Wiesbaden: Springer VS.

Mahrt, Merja; Scharkow, Michael (2014): „Der Wert von Big Data für die Erforschung digitaler Medien." In: Big Data, hg. v. Ramón Reichert, S. 221–238. Bielefeld: transcript.

Maier, Daniel et al. (2018): „Applying LDA topic modeling in communication research: Toward a valid and reliable methodology." Communication Methods and Measures 12 (2-3), S. 93–118. https://doi.org/10.1080/19312458.2018.1430754.

Meister, Jan Christoph et al. (2019): „CATMA: Computer assisted text markup and analysis." https://catma.de, zuletzt geprüft am 04.06.2024.

Meyer, Thomas (2014): „Tagungsbericht: Digital Humanities. Methodischer Brückenschlag oder ‚feindliche Übernahme'? Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik." In: H-Soz-Kult. https://www.hsozkult.de/conferencereport/id/fdkn-124008, zuletzt geprüft am 04.06.2024.

Munzert, Simon; Rubba, Christian; Meißner, Peter; Nyhuis, Dominic (2015): Automated data collection with R: A practical guide to web scraping and text mining. Chichester: Wiley.

Münch, Felix; Thies, Ben; Puschmann, Cornelius; Bruns, Axel (2021): „Walking through Twitter: Sampling a language-based follow network of influential Twitter accounts." Social Media & Society 7 (1), S. 205630512098447. https://doi.org/10.1177/2056305120984475

Nentwich, Michael (2023): „Wissenschaftliche Technikfolgenabschätzung: Begriff, Typologie und Konsequenzen für die Praxis.“ Manu:scripts ITA-23-02, Wien: Österreichische Akademie der Wissenschaften. https://epub.oeaw.ac.at/0xc1aa5576_0x003eb001.pdf, zuletzt geprüft am 06.06.2024.

Nicholas, Josh (2023): „Elon Musk drove more than a million people to Mastodon – but many aren’t sticking around.“ 07.01.2023, https://www.theguardian.com/news/datablog/2023/jan/08/elon-musk-drove-more-than-a-million-people-to-mastodon-but-many-arent-sticking-around, zuletzt geprüft am 04.06.2024.

Niekler, Andreas et al. (2018): „iLCM: A virtual research infrastructure for large-scale qualitative data.“ In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), S. 1313–1319. Miyazaki: ELRA.

Nyhuis, Dominic; Ringwald, Tobias; Rittmann, Oliver; Gschwend, Thomas; Stiefelhagen, Rainer (2021): „Automated video analysis for social science research.“ In: Handbook of Computational Social Science, S. 386–398. London: Routledge. https://doi.org/10.4324/9781003025245-26

OpenAI; Achiam, Josh et al. (2024): „GPT-4 technical report.“ arXiv. http://arxiv.org/abs/2303.08774, zuletzt geprüft am 06.06.2024.

Papachristos, George (2019): „System dynamics modelling and simulation for sociotechnical transitions research.“ Environmental Innovation and Societal Transitions 31: 248–61. https://doi.org/10.1016/j.eist.2018.10.001

Pekárek, Aleš; Willems, Marieke (2012): „The Europeana newspapers: A gateway to European newspapers online.“ In: Progress in Cultural Heritage Preservation 7616, hg. v. David Hutchison et al., S. 654–659. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-34234-9_68

Pfetsch, Barbara, Daniel Maier, Peter Miltner; Annie Waldherr (2016): „Challenger networks of food policy on the Internet.“ International Journal of E-Politics 7 (1), S. 16–36. https://doi.org/10.4018/IJEP.2016010102

Porter, Nathaniel; Verdery, Ashton; Gaddis, Michael (2020): „Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities.“ PloS One 15 (6), S. e0233154. https://doi.org/10.1371/journal.pone.0233154

Pournaki, Armin; Gaisbauer, Felix; Banisch, Sven; Olbrich, Eckehard (2021): „Twitter explorer.“ Journal of Digital Social Research 3 (1). https://doi.org/10.33621/jdsr.v3i1.64

Rauh, Christian; Schwalbach, Jan (2020): „The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.“ Harvard Dataverse. https://doi.org/10.7910/DVN/L4OAKN

Rädiker, Stefan; Kuckartz, Udo (2019): Analyse qualitativer Daten mit MAXQDA. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-22095-2

Roberts, Margaret et al. (2014): „Structural topic models for open-ended survey responses.“ American Journal of Political Science 58 (4), S. 1064–1082. https://doi.org/10.1111/ajps.12103

Rüdiger, Jan Oliver (2021): „CorpusExplorer.“ PhD thesis, Universität Kassel. https://doi.org/10.17170/kobra-202202085725

Schaal, Gary; Heyer, Gerhard; Wiedemann, Gregor; Niekler, Andreas; Dumm, Sebastian (2016): „Postdemokratie und Neoliberalismus: Zur Nutzung neoliberaler Argumentationen in der bundesdeutschen Politik 1949-2011. Gemeinsamer Sach- und Schlussbericht für das Verbundprojekt, Berichtszeitraum Mai 2012-August 2015.“ Helmut-Schmidt-Universität; Universität der Bundeswehr Hamburg; Fakultät für Wirtschafts- und Sozialwissenschaften; Institut für Politikwissenschaft. https://doi.org/10.2314/GBV:871451344

Schmidt, Jan-Hinrik; Kessling, Philipp; Rau, Jan; Linnekugel, Clara; Moradi, Jasmina; Nasser, Fred (2022): „Twitter- und Facebook-accounts der Kandidierenden zur Bundestagswahl 2021.“ Open Science Framework. https://doi.org/10.17605/OSF.IO/WN48Y

Schmidt, Jan-Hinrik (2021): „Facebook- und Twitter-Nutzung der Kandidierenden zur Bundestagswahl 2021.“ Media Perspektiven 12, S. 639–53. https://www.ard-media.de/fileadmin/user_upload/media-perspektiven/pdf/2021/2112_Schmidt.pdf, zuletzt geprüft am 06.06.2024.

Sen, Indira; Flöck, Fabian; Weller, Katrin; Weiß, Bernd; Wagner, Claudia (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

Silber, Henning et al. (2022): „Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour." Journal of the Royal Statistical Society 185 (2), S. 387—407. https://doi.org/10.1111/rssa.12954

Sinclair, Stéfan; Rockwell, Geoffrey (2012): „Voyant tools." Web application. http://voyant-tools.org, zuletzt geprüft am 04.06.2024.

Stier, Sebastian; Mangold, Frank; Scharkow, Michael; Breuer, Johannes (2022): „Post post-broadcast democracy? News exposure in the age of online intermediaries." American Political Science Review 116 (2), S. 768–774. https://doi.org/10.1017/S0003055421001222

Stokel-Walker, Chris (2024): „Under Elon Musk, X is denying API access to academics who study misinformation." Fastcompany, 02.07.2024. https://www.fastcompany.com/91040397/under-elon-musk-x-is-denying-api-access-to-academics-who-study-misinformation, zuletzt geprüft am 04.06.2024.

Su, Leona Yi-Fan; Xenos, Michael; Rose, Kathleen; Wirz, Christopher; Scheufele, Dietram; Brossard, Dominique (2018): „Uncivil and personal? Comparing patterns of incivility in comments on the facebook pages of news outlets." New Media & Society 20 (10), S. 3678–3699. https://doi.org/10.1177/1461444818757205

Thiele, Daniel (2022): dictvectoR: Word vectors for dictionaries. https://thieled.github.io/dictvectoR/, zuletzt geprüft am 04.06.2024.

Thiele, Daniel; Turnšek, Tjaša (2022): „How Right-Wing Populist Comments Affect Online Deliberation on News Media Facebook Pages." Media & Communication 10 (4), S. 141—154. https://doi.org/10.17645/mac.v10i4.5690

Törnberg, Petter (2023): „ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning." arXiv. http://arxiv.org/abs/2304.06588

van Zundert, Joris (2012): „If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities.“ Historische Sozialforschung 37 (3): 165–86. http://www.jstor.org/stable/41636603

Venturini, Tommaso; Rogers, Richard (2019): „API-based research or how can digital sociology and journalism studies learn from the facebook and Cambridge analytica data breach.“ Digital Journalism 7 (4), S. 532–540. https://doi.org/10.1080/21670811.2019.1591927

Vosen, Simeon; Schmidt, Torsten (2011): „Forecasting private consumption: Survey-based indicators vs. Google trends.“ Journal of Forecasting 30 (6), S. 565–578. https://doi.org/10.1002/for.1213

Vosoughi, Soroush; Roy, Deb; Aral, Sinan (2018): „The spread of true and false news online.“ Science 359 (6380), S. 1146–1151. https://doi.org/10.1126/science.aap9559

Waldherr, Annie (2014): „Emergence of news waves: A social simulation approach.“ Journal of Communication 64 (5), S. 852–73. https://doi.org/10.1111/jcom.12117

Waldherr, Annie; Geise, Stephanie; Mahrt, Merja; Katzenbach, Christian; Nuernbergk, Christian (2021): „Toward a stronger theoretical grounding of computational communication science.“ Computational Communication Research 3 (2), S. 1–28. https://doi.org/10.5117/CCR2021.02.002.WALD

Wang, Xiaohui; Song, Yunya; Su, Youzhen (2022): „Less fragmented but highly centralized: A bibliometric analysis of research in computational social science.“ Social Science Computer Review 41 (3), S. 946-966. https://doi.org/10.1177/08944393211058112

Weinberg, Jill; Freese, Jeremy; McElhattan, David (2014): „Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample.“ Sociological Science 1, S. 292–310. https://doi.org/10.15195/v1.a19

Wiedemann, Gregor (2013): „Opening up to big data: Computer-assisted analysis of textual data in social sciences.“ Historical Social Research 38 (4), S. 332–357.

Wiedemann, Gregor (2016): Text mining for qualitative data analysis in the social sciences. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-15309-0

Wiedemann, Gregor; Fedtke, Cornelia (2021): „From frequency counts to contextualized word embeddings: The Saussurean turn in automatic content analysis.“ In: Handbook of Computational Social Science 2, hg. v. Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu und Lars Lyberg, 366–385. London: Routledge. https://doi.org/10.4324/9781003025245-25

Wiedemann, Gregor; Niekler, Andreas (2017): „Hands-on: A five day text mining course for humanists and social scientists in R.“ In Proceedings of the Workshop on Teaching NLP for Digital Humanities, S. 57–65. http://ceur-ws.org/Vol-1918/wiedemann.pdf, zuletzt geprüft am 04.06.2024.

Wiedemann, Gregor; Münch, Felix; Rau, Jan; Kessling, Phillip; Schmidt, Jan-Hinrik (2023): „Concept and challenges of a social media observatory as a DIY research infrastructure.“ Publizistik 68 (2–3), S. 201–223. https://doi.org/10.1007/s11616-023-00807-6

Wiedemann, Gregor; Schmidt, Jan-Hinrik; Rau, Jan; Münch, Felix; Kessling, Philipp (2023): „Telegram in der politischen Öffentlichkeit. Zur Einführung in das Themenheft.“ Medien & Kommunikationswissenschaft 71 (3–4), S. 207–211. https://doi.org/10.5771/1615-634X-2023-3-4-207

Wijffels, Jan (2022): „Udpipe: tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP toolkit: R package.“ https://CRAN.R-project.org/package=udpipe, 04.06.2024.

Wilkinson, Mark et al. (2016): „The FAIR guiding principles for scientific data management and stewardship.“ Scientific Data 3 (1), S. 160018. https://doi.org/10.1038/SDATA.2016.18

Willaert, Tom; Banisch, Sven; van Eecke, Paul; Beuls, Katrien (2022): „Tracking causal relations in the news: Data, tools, and models for the analysis of argumentative statements in online media.“ Digital Scholarship in the Humanities 37 (4), S. 1358–1375. https://doi.org/10.1093/llc/fqab107

Zallot, Camilla; Paolacci, Gabriele; Chandler, Jesse; Sisso, Itay (2021): „Crowdsourcing in observational and experimental research.“ In: Handbook of Computational Social Science 2, hg. v. Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu und Lars Lyberg, S. 140–157. London: Routledge. https://doi.org/10.4324/9781003025245-12

Zhang, Han; Pan, Jennifer (2019): „CASM: A deep-learning approach for identifying collective action events with text and image data from social media.“ Sociological Methodology 49 (1), S. 1–57. https://doi.org/10.1177/0081175019860244

Ziems, Caleb; Held, William; Shaikh, Omar; Chen, Jiaao; Zhang, Zhehao; Yang, Diyi (2024): „Can large language models transform computational social science?“ Computational Linguistics 50 (1), S. 237–291. https://doi.org/10.1162/coli_a_00502

Die digitale Transformation ist nicht nur auf gesellschaftlicher Ebene relevant, sondern verändert auch Forschungsprozesse. Für das Feld der Technikfolgenabschätzung (TA) zeigt sich dies u.a. in der steigenden Relevanz von digitalen Forschungsmethoden. Die vorliegende Studie gibt einen Überblick über die Computational Social Sciences und stellt für die TA relevante Anwendungen vor. Mit der „Toolbox für Einsteiger“ sowie dem „Fallbeispiel Metaverse“ werden forschungspraktische Handreichungen für die TA konzipiert, die einen Zugang zu digitalen Methoden bieten und die gesellschaftliche Relevanz der Ergebnisse aufzeigen.

ISBN 978-3-7315-1391-9
9 783731 513919

Gedruckt auf FSC-zertifiziertem Papier