

2 Forschungsstand

Zahlreiche Studien unterschiedlicher wissenschaftlicher Disziplinen, wie der Kommunikationsforschung, Soziologie, Geografie oder Politikwissenschaften, beschäftigen sich bereits mit *Twitter*. In den letzten Jahren stieg vor allem das Interesse an der Auswertung geospezifischer Daten, die Rückschlüsse auf Standorte und Bewegungsmuster liefern können (z.B. Gerätestandort, Zeitzone, Sprache). Beispielsweise nutzten Hawelka et al. (2014) die in Tweets gespeicherten Standortdaten zur Erstellung von Bewegungsmustern, indem sie die über einen längeren Zeitraum gesammelten Geo-Daten (GPS-Daten und IP-Adressen) von Nutzerprofilen verknüpften. Jedoch erkannten Graham, Hale und Gaffney (2014) bei einem Vergleich von benutzerdefinierten Profil-Standorten mit GPS-basierten Gerätestandorten und Tweet-spezifischen Zeitzone-Angaben, dass diese Informationen häufig voneinander abweichen. Nutzer/-innen geben in ihrem Profil oft falsche Wohn- beziehungsweise Aufenthaltsorte an.

Cheng, Caverlee und Lee (2010) versuchten eine zuverlässigere Lokalisierung von Tweets, die unabhängig etwaiger Standortangaben durch die Nutzer ist. Die geolinguistische Analyse von Tweets erfolgte hier mithilfe vorhandener Algorithmen (*Google Geocoding API, Yahoo PlaceFinder*), um über den Tweet-Inhalt den momentanen Aufenthaltsort abzuleiten. Letztlich erwies sich jedoch eine rein technische Auswertung im Vergleich zu einer menschlichen, manuellen Codierung als nur bedingt verlässlich.

Carter, Tsagkias und Weerkamp (2011), Gottron und Lipka (2010) sowie Bifet und Frank (2010) verweisen hierbei auf elementare Unterschiede zum normalem Fließtext, auf den jedoch viele Linguistik-Programme ausgelegt sind: Twitter-Meldungen sind auf 140 Zeichen begrenzt, enthalten häufig Abkürzungen, Neologismen, mehrsprachige Inhalte und selten eine klare Satzstruktur, wie dies bei normalen Fließtexten der Fall ist. Dementsprechend können Inhalte leicht fehlgedeutet werden, weshalb eine programmgestützte Inhaltsanalyse immer mit Vorsicht genutzt werden sollte.

Das Problem der eingeschränkten Analysierbarkeit von Tweets durch Programme betrifft auch andere Forschungszweige, die auf eine automatisierte inhaltliche Auswertung angewiesen sind, wie die Sentiment-Forschung. Mithilfe mehrstufiger Filter- und Verarbeitungsprozesse (Bollen, Pepe, & Mao, 2011; Pak & Paroubek, 2010), Einbezug zusätzlicher Informationen wie Nutzerdaten (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010) oder verwandter/ähnlicher Tweets (Jiang, Yu, Zhou, Liu, & Zhao, 2007) sowie der Verwendung lernfähiger Analyseprogramme (Bifet & Frank, 2010; Carter et al., 2011; Tumasjan, Sprenger, Sandner, & Welpe, 2010) können die oben genannten Probleme jedoch einigermaßen kontrolliert werden. Allerdings sollte eine korrekte Einordnung der Stimmung auch den Kontext des Tweets berücksichtigen, was in bisherigen Studien nicht gemacht wurde. Dies könnte an der Datenstruktur der Tweets liegen: Tweets werden einzeln nach Veröffentlichungszeitpunkt und nicht als zusammenhängende Konversationen übermittelt. Folglich müssten Unterhaltungen erst manuell zusammengefügt werden.

Studien der Sentiment-Forschung untersuchten beispielsweise den Zusammenhang der aggregierten Stimmung beobachteter Nutzer (= Sentiment) mit Kurschwankungen, Ölpreisen und medialer Großereignissen (Bollen, Pepe et al., 2011), die Möglichkeit, mithilfe des aktuellen Stimmungsbildes auf Twitter Aktienkurse vorherzusagen (Bollen, Mao, & Zeng, 2011; Zhang, Fuehres, & Gloor, 2011).

Echtzeit-Daten aus Twitter dienen zuletzt auch als Grundlage zur Erkennung von Epidemien, wie Influenza (Signorini, Segre, & Polgreen, 2011), beziehungsweise des allgemeinen Gesundheitszustandes der Bevölkerung (Paul & Dredze, 2011; Scanfeld, Scanfeld, & Larson, 2010). Aramaki, Maskawa und Morita (2011) zeigten, dass mithilfe automatisierter Erhebung und Analyseverfahren, wie dem Natural Language Processing, Vorhersagen über das Auftreten und den Verlauf von Grippe-Wellen machen lassen. Des Weiteren wurden Tweets zur Früherkennung von Erdbeben (Earle, Bowden, & Guy, 2012; Sakai, Okazaki, & Matsuo, 2010), oder zur Analyse der Kommunikation während Krisen verarbeitet (Acar & Muraki, 2011; Heverin & Zach, 2010; Mendoza, Poblete, & Castillo, 2010; Vieweg, Hughes, Starbird, & Palen, 2010). Jedoch besteht auch hier das Problem, dass Tweets aufgrund der unkonventionellen Sprache nur schwer automatisiert analysiert werden können. Ähnlich, wie bei der Sentiment-Erkennung, gehen durch den fehlenden Miteinbezug des Twitter-Kontextes womöglich viele relevante Informationen verloren.

Ein weiterer Schwerpunkt der Forschung liegt in der Analyse der politischen Kommunikation auf beziehungsweise über Twitter. Dabei wurden sowohl Tweets

über Politiker/-innen als auch deren Nachrichten und Interaktion mit anderen Nutzern auf Twitter betrachtet. So dienten Echtzeitdaten für eine Bewertung von Politikern während TV-Debatten (Diakopoulos & Shamma, 2010) oder zur Analyse der Stimmung während der US-Präsidentschaftswahl 2012 (Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012).

Umstritten ist die Möglichkeit, mit Hilfe von Twitter-Daten den Ausgang von Wahlen zu prognostizieren. Es gibt einige Kontroversen hinsichtlich der Aussagekraft von Tweets und der Zuverlässigkeit der Schätzung. Tumasjan et al. (2010) sowie Sang und Bos (2012) sehen in Twitter trotz der eingeschränkten Repräsentativität der Daten ein relativ zuverlässiges Instrument zur Wahlprognose. Jung-herr, Jürgens und Schoen (2012) zeigen jedoch, dass sowohl das Datenmaterial, als auch Erhebungszeitpunkt und Auswahl der Parteien keine valide Erhebungsmethode darstellen und somit keine verlässlichen Rückschlüsse auf Wahlergebnisse erlauben. Aufgrund der unterschiedlichen Wähler-Zielgruppen gibt es eine Verzerrung hinsichtlich Twitter-Nutzungsverhalten und somit der Tweet-Häufigkeit der jeweiligen Partei-Anhängerschaft. Beispielsweise liegt es nahe, dass Unterstützer der Piraten-Partei deutlich häufiger twittern als Anhänger der großen Volksparteien.

Conover et al. (2011) beobachteten während der US-amerikanischen *Midterm Elections* 2010 eine hohe Polarisierung der politisch aktiven Twitter-Nutzer zwischen linkem und rechten Lager und einer geringen Interaktion zwischen diesen Gruppen. Weitere Studien ergaben, dass der Grad politischer Inhalte auf Twitter stark von medialen Ereignissen, wie Diskussionsrunden oder Wahlveranstaltungen, abhängt (Dusch et al., 2015; Larsson & Moe, 2012). Auch ist der Interaktionsgrad zwischen Politikern und „normalen“ Usern eher gering: Politiker verwenden Twitter meist eher als Werbepattform für politische Veranstaltungen (Dusch et al., 2015; Thimm, Einspänner, & Dang-Anh, 2012) beziehungsweise zur Verbreitung ihrer Standpunkte, als zur direkten Interaktion mit ihren Kontakten (Elter, 2013; Grant, Moon, & Busby Grant, 2010; Parmelee & Bichard, 2012).

Eine deutlich aktivere politische Partizipation und Kommunikation auf Twitter findet dagegen während politischer Proteste und Aufständen statt: Besonders bei der Ägyptischen Revolution sowie der *Grünen Revolution* im Iran spielte Twitter eine zentrale Rolle bei der Organisation und Informationsweitergabe (Bruns et al., 2013). Bei Ereignissen, in denen klassische Massenmedien aufgrund der Rasanz der Entwicklungen oder Abwesenheit von Journalisten nicht zeitnah reagieren konnten, etablierte sich Twitter als wichtige Plattform für die Produktion und Distribution von Nachrichten (Papacharissi & de Fatima Oliveira, 2012). Der Kurznachrichtendienst diente, aus Mangel an zuverlässigen staatlichen Medien und

aufgrund der Unterdrückung freier, kritischer Meinungsäußerungen, hierbei auch als Nachrichtenquelle für westliche Medien (Khondker, 2011; Lotan et al., 2011). Schließlich befassten sich auch zahlreiche Arbeiten mit der Kommunikation auf Twitter an sich: Chen (2011) begründete die Art und Stärke des Nutzungsverhalten von Twitter-Usern mit dem *Uses and Gratification* Ansatz. Je länger die Nutzung (hinsichtlich des Zeitraums), desto belohnender wird eine Vernetzung mit anderen Nutzern wahrgenommen, wobei die Zahl eigener Tweets und Replies die Stärke des Effekts beeinflusst. Liu, Cheung und Lee (2010) sehen vier Dimensionen innerhalb des Ansatzes, die belohnend auf die Nutzung von Twitter wirken: *Content* (Möglichkeit zur Informationsaufnahme- und Verbreitung), *Technology* (bequeme und unmittelbare Kommunikation), *Social* (Soziale Interaktion und Vernetzung) sowie *Process* (Unterhaltung, Zeitvertreib), wobei die letzten beiden eine geringere Bedeutung haben. Dies wird damit erklärt, dass Twitter ursprünglich nur zum Austausch von Informationen konzipiert war und soziale Interaktionsmöglichkeiten erst später implementiert wurden. Auch haben sich Kommunikationsmöglichkeiten erst mithilfe von Konventionen, wie Retweets oder direkte User-Verweise in Tweets (@username) durchgesetzt (boyd, Golder, & Lotan, 2010; Honeycutt & Herring, 2009).

Cha, Haddadi, Benevenuto und Gummadi (2010) betrachteten die soziale und informationelle Komponente hinsichtlich des Einflusses populärer Twitterer. Die Wahrscheinlichkeit, dass ein Tweet eine hohe Aufmerksamkeit erhält ist demnach weniger von der Vernetzung eines Users, im Sinne von Followern, sondern vom Inhalt der Nachricht abhängig. Diese These, dass Twitter eher zum Informationsaustausch verwendet wird, als zum Knüpfen sozialer Kontakte, unterstützen auch Huberman, Romero und Wu (2008), Java, Song, Finin und Tseng (2007) sowie Johnson und Yang (2009).

So unterschiedlich die Forschungsabsichten und Verwendungszwecke bezüglich Twitter sind, so verschieden sind auch die Methoden zur Datengewinnung und Auswertung. Die Erfassung von Twitter-Daten lässt sich dabei in drei Komplexe zusammenfassen: Abfragen historischer Daten über die Programmschnittstelle *Search API*, Erhebungen von gesampelten Echtzeitdaten über die *Streaming API* und die Verwendung von Programmen und Datensätzen Dritter (siehe Kapitel 4.1). Die hier erwähnten Studien gingen wie folgt vor: boyd, Golder und Lotan (2010), Cheng et al. (2010), Diakopoulos und Shamma (2010), Grant et al. (2010), Sakai et al. (2010) und Vieweg et al. (2010) nutzen die frei zugängliche Search API von Twitter, um über Suchabfragen historische Daten eines eingeschränkten Zeitraums zu erhalten. Bifet und Frank (2010), Graham et al. (2014), Hawelka et al. (2014), Sang und Bos (2012) sowie Signorini et al. (2011) griffen dagegen mit Hilfe der

Streaming API (gesampelte) Daten in Echtzeit ab. Neben diesen beiden populären, da kostenlosen, Methoden der Datenerhebung auf Twitter, gibt es noch eine Vielzahl von Drittanbietern, die Twitter-Daten gebührenpflichtig oder gratis zur Verfügung stellen. Conover et al. (2011) erhielten Zugriff auf die sogenannte *Gardenhose*⁴ wogegen Wang et al. (2012) Daten vom Dienstleister *Gnip* kauften und Dusch et al. (2015) sowie Larsson und Moe (2012) Online-Dienste zur Datenerfassung nutzen.

Hinsichtlich der Auswertung von Twitter-Daten ergibt sich ein ähnlich differenziertes Bild: Twitter-bezogene Studien fokussieren sich nicht nur auf reine Inhaltsanalysen, sondern beziehen sich (zusätzlich) auch auf Befragungen oder Experimente. Dennoch ist die Inhaltsanalyse nach einer Meta-Studie von Williams, Terras und Warwick (2013) eine dominierende Methode, was sich auch auf das umfangreiche Datenangebot durch Twitter zurückführen lässt. Die Nutzung der bereits vorhandenen, leicht zugänglichen, stark strukturierten und ausführlichen Daten ist einfacher und schneller als die Durchführung von Befragungen oder Experimenten. Interviews werden häufig nur ergänzend durchgeführt, etwa, um Ergebnisse der Inhaltsanalyse durch ermittelte Einstellungen und Verhalten der Nutzer zu erklären.

Dennoch fehlen methodische Standards, da die Twitter-Forschung noch sehr jung ist (Bruns & Liang, 2012). Ein weiterer Forschungsschwerpunkt liegt deshalb in der Konzeption neuer Methoden und Algorithmen zur Analyse der Daten (Williams et al., 2013). Einige Forschende entwickeln für ihre Forschungszwecke eigene Ansätze beziehungsweise Programme zur Twitter-Analyse. Das eigentliche methodische Vorgehen, insbesondere die Datengewinnung, wird dabei selten detailliert präsentiert (Weller, 2014). Trotz der hier dargestellten großen Bandbreite an Ansätzen und Verfahren der Twitter-Analyse, gibt es kaum wissenschaftliche Arbeiten, die sich mit den Methoden der Datengewinnung und Auswertung befassen. Wenn überhaupt, wurden nur einzelne Vorgehensweisen angesprochen.

So zeigen Perera, Anand, Subbalakshmi und Chandramouli (2010), wie mit der Programmiersprache *Python* Twitter-Daten gesammelt und in einem *MySQL*-Datensystem verarbeitet werden können. Tugores und Colet (2013) vergleichen für eine Mobilitätsanalyse zwei Varianten von Datenbanksystemen (*SQL* und *noSQL*) im Kontext einer Twitter-Analyse mit *Python*. Bruns und Liang (2012) präsentierten mehrere Programme zum Erfassen und Analysieren von Tweets

⁴ Die Daten der öffentlichen Streaming API und REST APIs sind hinsichtlich Datenvolumen und Abfragehäufigkeit limitiert. Innerhalb der Streaming API bietet die Gardenhose einen größeren Datenumfang als der allgemeine Datenzugang *Spritzer*, wogegen die *Firehose* einen Echtzeit-Zugriff auf alle Daten ermöglicht (siehe Kapitel 4.1 für eine ausführliche Erläuterung).

während Naturkatastrophen. Dennoch findet sich nirgends eine detaillierte Bewertung der Ansätze. Aufgrund der unterschiedlichen Disziplinen und somit auch Forschungsschwerpunkte fehlte auch die spezifische Bewertung dieser Möglichkeiten im Hinblick auf die Analyse der Twitter-Kommunikation.

Kumar, Morstatter und Liu (2014) liefern bisher die umfassendste Übersicht, mit welchen Ansätzen Twitter-Daten gesammelt und verarbeitet werden können. Jedoch werden auch hier nur ausgewählte Aspekte der Datensammlung und -analyse betrachtet und die einzelnen Methoden weder miteinander verglichen, noch hinsichtlich ihrer Praktikabilität bei wissenschaftlichen Erhebungen bewertet. Eine ähnliche Zielsetzung verfolgt Russell (2013): Anhand zahlreicher fallspezifischer Beispiele erhält der Leser einen guten Überblick über die Möglichkeiten der (nicht nur) Twitter-bezogenen Datenerhebung und Auswertung mittels Python, MongoDB und NLTK. Jedoch ist auch hier eine vergleichende und wertende Betrachtung – besonders im Hinblick auf die Nützlichkeit für die Forschung – nicht vorhanden.

Es gibt folglich bereits eine Vielzahl an Studien, die sich mit der Kommunikation auf Twitter beziehungsweise der wissenschaftlichen Auswertung der generierten Nutzer-Daten befassen haben. Was fehlt, ist ein vergleichender Überblick über Verfahren der Twitter-Analyse für die Sozialwissenschaften. Williams et al. (2013) befanden in ihrer Meta-Analyse, dass sich etwa 80% der analysierten Beiträge auf den Inhalt der Tweets sowie die Nutzer und deren Kommunikationsweise konzentrierten. Dabei wurde eine Vielzahl unterschiedlichster Methoden zur Erfassung und Analyse von Twitter-Daten angewendet, oftmals sogar mehrere Ansätze in einer Arbeit. Demgegenüber war die rein technische Betrachtung von Twitter am stärksten unterrepräsentiert. Die Autoren verwiesen hier nicht nur auf eine geringere Beimessung an Bedeutung, sondern auch auf mögliche technische Barrieren und Verständnisprobleme (Williams et al., 2013, S. 402).

Aktuelle Verfahren zur Messung der Nutzung von Twitter (und anderen sozialen Medien) sind weder standardisiert, noch unabhängig bestätigt, sondern funktionieren eher als eine Art „Black Box“ (Weller, Bruns, Burgess, Mahrt, & Puschmann, 2014, S. xxxii), deren Ergebnisse Forschende vertrauen müssen. Deshalb sollen nach einer theoretischen Einführung in den Dienst Twitter und die hier verwendete Programmiersprache Python einige gängige Ansätze genauer betrachtet und anhand von Fallbeispielen hinsichtlich Praktikabilität und Anwendungsweise verglichen werden.

Open Access Dieses Kapitel wird unter der Creative Commons Namensnennung - Nicht kommerziell 4.0 International Lizenz (<http://creativecommons.org/licenses/by-nc/4.0/deed.de>) veröffentlicht, welche für nicht kommerzielle Zwecke die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Etwasige Abbildungen oder sonstiges Drittmateriale unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende oder der Quellreferenz nichts anderes ergibt. Sofern solches Drittmateriale nicht unter der genannten Creative Commons Lizenz steht, ist eine Vervielfältigung, Bearbeitung oder öffentliche Wiedergabe nur mit vorheriger Zustimmung des betreffenden Rechteinhabers oder auf der Grundlage einschlägiger gesetzlicher Erlaubnisvorschriften zulässig.

