3 Grundlagen

Dieses Kapitel dient der Vermittlung technischer Grundlagen und soll dem Leser einen Einblick in den Mikroblogging-Dienst Twitter. Dafür wird zuerst Twitter vorgestellt (Kapitel 3.1), indem auch die Kommunikation auf Twitter charakterisiert sowie Konventionen der Interaktion und allgemeine Begrifflichkeiten erläutert werden. Anschließend folgen ein Überblick der Datenstruktur und eine Skizzierung der darin enthaltenen wesentlichen Informationen. Da die vorliegende Arbeit die Programmiersprache Python zur Datensammlung und -analyse verwendet, stellt Kapitel 3.2 diese kurz vor und erläutert deren Vorteile hinsichtlich Anwendung und Verständlichkeit.

3.1 Post, Reply, Retweet – der Internet-Dienst Twitter

Twitter ist in erster Linie ein Echtzeit-Internetdienst zum Teilen von auf 140 Zeichen limitierten Text-Nachrichten (*Tweets*) in einem personalisierten, öffentlichen Nachrichtenstrom (Jürgens & Jungherr, 2011, S. 203). Dieser Nachrichtenfeed kann von anderen Twitterern abonniert werden, um dadurch jeder neuen Nachricht eines Nutzers automatisch zu folgen. Der abonnierende Nutzer wird als *Follower* bezeichnet und ist als dieser öffentlich gekennzeichnet (siehe Kapitel 3.1.2). Die Stärke des Dienstes liegt in der schnellen und ungefilterten Verbreitung von Informationen (Parmelee & Bichard, 2012, S. 216). Durch die Begrenzung auf 140 Zeichen muss der Nachrichteninhalt auf das Wesentliche konzentriert werden, die geringe Länge fördert auch eine gute und schnelle Lesbarkeit. Während in der Frühphase der Entwicklung der reine Informationsaustausch im Fokus stand, folgten in mehreren Entwicklungsschritten weitere Funktionen zur sozialen Interaktion. So ermöglicht die Plattform mittlerweile auch das Weiterleiten von Tweets (*Retweet*), eine explizite Nennung und Verknüpfung anderer Nutzer/-innen in Nachrichten (*Mention*), das Teilen von Fotos, Links und Videos sowie das Schreiben privater Nachrichten (*Direct Message*) zu einzelnen Personen oder Gruppen (Stone, 2009; Weil, 2014).

F. Pfaffenberger, *Twitter als Basis wissenschaftlicher Studien*,
DOI 10.1007/978-3-658-14414-2_3

Twitter ist mittlerweile ein weit verbreiteter Kommunikationskanal mit einer Fülle von Anwendungsmöglichkeiten. Beispielweise nutzt die Politik Twitter zur Interaktion mit (potentiellen) Wählern, Journalist/-innen zur Verbreitung von Informationen (wie Eilmeldungen), Fernsehanstalten als weiteren Kommunikationskanal während TV-Sendungen (für Kommentare und Feedback) oder Unternehmen als Werbekanal mit Hinweisen zu Aktionen oder Produkten (Bruns & Stieglitz, 2012; Grant et al., 2010; Jansen, Zhang, Sobel, & Chowdury, 2009; Jungherr, 2015; Lasorsa, Lewis, & Holton, 2012; Mamic & Almaraz, 2014). Hinzu kommt eine vielfältige Anwendung als Kommunikationskanal zwischen sich gegenseitig bekannten oder unbekannten Personen – von der „normalen" Nutzung im Alltag bis zur Interaktion während politischer Krisen, wie der des sogenannten *Arabischen Frühlings* (boyd et al., 2010b; Christensen, 2011; Lotan et al., 2011).

Laut Twitter (2015k) gab es im März 2015 etwa 302 Millionen monatlich aktive Nutzer, was im Vergleich zu März 2010 mit 30 Millionen (Twitter, Inc., 2015b) eine Verzehnfachung bedeutet. Twitter definiert aktive Nutzer/-innen als Personen, die pro Monat mindestens einmal auf der Plattform aktiv waren (z.B. durch Anmelden im Account). Von den aktiven Usern nutzen etwa 80 Prozent den Dienst über das mobile Internet, insgesamt werden pro Tag 500 Millionen Tweets verfasst (Twitter, Inc., 2015k). Wissenschaftler, die Twitter-Daten beziehen, steht somit ein sehr großes potentielles Datenset zur Verfügung.

Zur Betrachtung der Staaten mit den meisten Twitter-Nutzern sollten gewichtete Daten verwendet werden, um Verzerrungen durch die Einwohnerzahl zu vermeiden. Da Twitter keine offiziellen Zahlen zur Herkunft seiner Nutzer/-innen veröffentlicht, führten Mocanu et al. (2013) eine Lokalisierung anhand von Sprache und Standort durch. Nach Anzahl der Accounts je 1000 Einwohner eines Staates ergab sich folgendes Bild: Kuwait (1 Prozent), die Niederlande (0,39 Prozent), Brunei (0,31), Großbritannien (0,3) und die USA (0,25) belegten Platz eins bis fünf. Deutschland wies in etwa einen Anteil von 0,04 Prozent an Twitter-Nutzern auf (Ebda). In absoluten Zahlen weisen die USA zwar den größten Anteil an Nutzern auf – dies bestätigen unter anderem Analysen des Twitter-Volumens (SimilarWeb, 2014) – relativ zur Einwohnerzahl belegt US-Amerika jedoch nur Platz fünf.

Allerdings sind diese Angaben alle nur eingeschränkt zuverlässig: Der Tweet-Standort erlaubt noch keinen verlässlichen Rückschluss auf die Nationalität des Nutzers. So könnten unter anderem Verzerrungen durch Reisen in andere Länder auftreten. Bei der Erhebung durch Mocanu et al. (2013) wurden zumindest Hauptreisezeiten berücksichtigt. Die Problematik der Lokalisierung von Tweets ist folglich auch hier präsent.

Interessant ist auch die Altersstruktur der monatlich aktiven Nutzer: Während bei Facebook die Altersgruppe der 25- bis 34-Jährigen mit 29 Prozent dominiert und knapp 24 Prozent der Nutzer älter als 45 sind (GlobalWebIndex, 2015), ist die Altersverteilung der männlichen und weiblichen Twitter-Nutzer gleichmäßiger (siehe Abbildung 2:). Zwar haben auch hier die 25- bis 34-Jährigen mit 22 Prozent den größten Anteil, jedoch sind die Abstände zu anderen Altersgruppen deutlich geringer. Nutzer ab 45 Jahren haben sogar einen Anteil von 38 Prozent, wovon die knappe Mehrheit über 55 Jahre alt ist (comScore, 2015).

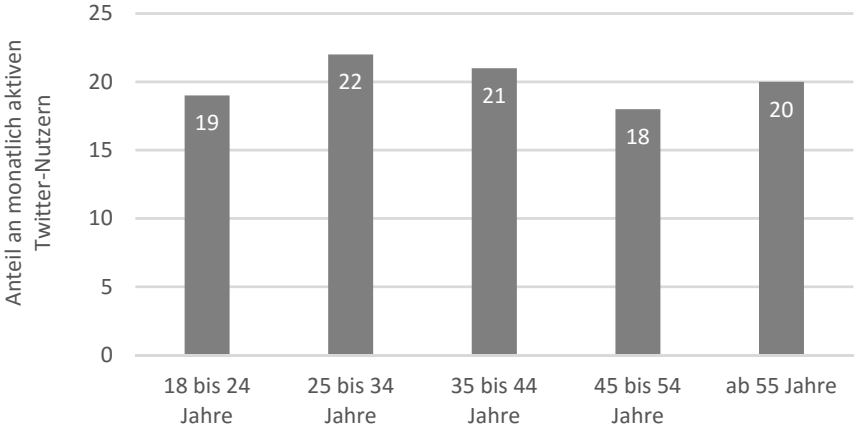

Abbildung 2: Altersverteilung aktiver Twitter-Nutzer im Dezember 2014. Quelle: comScore (2015), eigene Darstellung.

Der Online-Dienst Twitter wird, je nach Perspektive und Nutzung, mal als soziales Netzwerk, mal als reiner Kurznachrichtendienst bezeichnet. Diese Diskussion um die Definition von Twitter soll zunächst in Kapitel 3.1.1 aufgegriffen werden. Dabei wird auch auf aktuelle Statistiken über Nutzer und Nutzung eingegangen. Anschließend folgt eine genauere Betrachtung der Twitter-Nutzung und der Datenstruktur von Tweets.

3.1.1 Einordnung in die Social Media Landschaft

Aufgrund der mittlerweile umfassenden sozialen Kommunikationsmöglichkeiten ist eine klare Einordnung des Dienstes innerhalb der Social Media Landschaft nicht mehr möglich (Parmelee & Bichard, 2012, S. 38). Einerseits teilt Twitter viele Eigenschaften sozialer Netzwerke (wie *Facebook* oder *LinkedIn*): halb-öffentliche Profile, Interaktivität, einen sozialen Charakter der Interaktion, Vernetzung mit Nutzerlisten (boyd & Ellison, 2007). Andererseits wird Twitter auch als Mikroblogging-Plattform gesehen (Ebersbach, Glaser & Heigl, 2008) und ist mit seinen Funktionen und Eigenschaften immer noch näher an Blogs als an sozialen Netzwerken. Ross, Terras, Warwick und Welsh (2011, S. 217) definieren Mikroblogging wie folgt:

> „Microblogging is a variant of blogging, which allows users to quickly post short updates, providing an innovative communication method that can be seen as a hybrid of blogging, instant messaging, social networking and status notifications. The word's origin suggests that it shares the majority of elements with blogging, therefore it can potentially be described using blogging's three key concepts (Karger and Quan, 2004): the contents are short postings, these postings are kept together by a common content author who controls publication, and individual blog entries can be easily aggregated together."

Im Vergleich zu anderen Netzwerken wie *Facebook*, *Google+* oder *MySpace* basiert die soziale Vernetzung/Freundschaft durch Followers nicht auf Reziprozität (Kwak, Lee, Park, & Moon, 2010, S. 591). Der ursprüngliche, informationelle Zweck ist immer noch ein wichtiger Grund für die Twitter-Nutzung. Nach Parmelee und Bichard (2012, S. 64) sind *Information Seeking* und *Guidance*, also im weitere Sinne die Informationssuche zur Erleichterung von Entscheidungen und der Meinungsbildung, neben Unterhaltungs-Aspekten immer noch zentrale Nutzungsmotive.

Dass eine klare Abgrenzung nicht möglich ist, verdeutlicht auch die Tatsache, dass mittlerweile eine eigene Ökosphäre von Zusatzprogrammen und Diensten rund um Twitter entstanden ist (siehe Abbildung 3: auf der nächsten Seite), wie Kurzlink-Generatoren, Storytelling-Plattformen für Tweets oder Aggregatoren. Tweets sind zudem häufig der Ausgangspunkt für weitere Informationen, die über Links, Fotos und Videos vermittelt werden. Dadurch entsteht unter Umständen auch der Charakter eines Content Networks, auf welchem Inhalte geteilt werden.

Dennoch entwickelt sich Twitter immer mehr zu einem sozialen Netzwerk. Dies zeigt sich vor allem in der Implementierung zusätzlicher Funktionen: Während zu Beginn der Plattform nur ein Schreiben reiner, auf 140 Zeichen begrenzter,

Textnachrichten möglich war, wuchs Twitter nach und nach um soziale Funktionen, wie das Beantworten oder Teilen von Tweets. Wie stark sich Twitter von der ursprünglichen Idee der rein öffentlichen Informationsvermittlung distanziert hat, zeigt die jüngste Ankündigung von Twitter, Inc. Seit Juli 2015 sind private Nachrichten (*Direct Messages*) nicht mehr auf 140 Zeichen begrenzt (Twitter, 2015), sodass ausführliche, private Interaktionen ermöglicht werden. Diese und weitere Möglichkeiten sowie Konventionen der Kommunikation auf Twitter soll das folgende Kapitel betrachten.

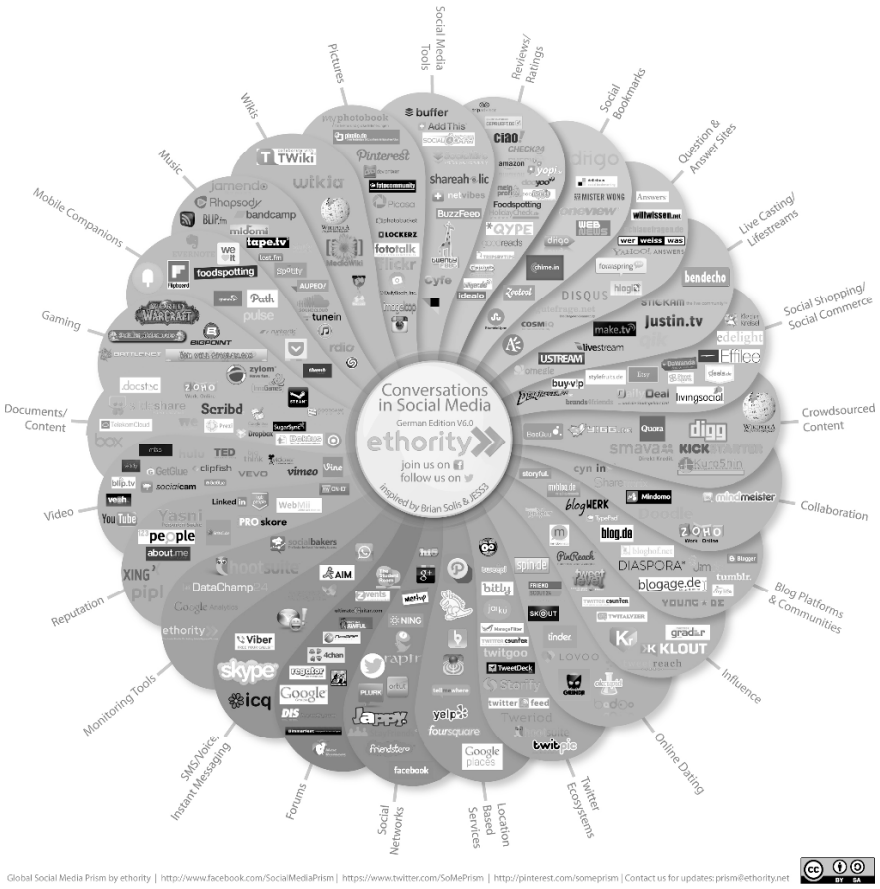

Abbildung 3: Social Media Prisma. Quelle: Ethority (2014).

3.1.2 Konventionen und Struktur der Kommunikation

Twitter bedient sich mehrerer Mechanismen zur Vereinfachung der Kommunikation: Mithilfe eines vorangestellten @ an einem existierenden Benutzernamen können einzelne *Twitter*-Nutzer in einem Tweet direkt adressiert werden. Man spricht hierbei von *Mentions* (*@Username*). Die direkte Beantwortung eines Tweets durch eine andere Nachricht heißt *Reply*. Der Unterschied zu einer reinen Erwähnung in einem Tweet besteht darin, dass bei einem Reply das Mention immer vorangestellt wird (z.B. „@Mustermann: Ich stimme dir zu!"). *Retweets* sind der Kern-Mechanismus auf Twitter: Damit können einzelne Tweets direkt zitiert oder mit anderen Nutzern geteilt werden (Suh, Hong, Pirolli, & Chi, 2010). Ein Retweet ist eine Weiterleitung einer Meldung, früher ersichtlich durch ein „RT @Username" im Fließtext, mittlerweile nur durch eine spezielle Markierung des Tweets (Halavais, 2014, S. 35). Dabei können Nutzer durch Retweets nicht nur Informationen teilen, sondern beispielweise auch Follower unterhalten (durch Teilen unterhaltsamer Tweets) oder mit beigefügten Kommentaren die eigene Zustimmung oder Ablehnung eines Tweets äußern (boyd et al., 2010a). Seit Juni 2014 besteht die Möglichkeit, zu einem Retweet nochmal zusätzlich einen bis zu 140 Zeichen langen Kommentar zu schreiben (Perez, 2014).

Hashtags (*hash*, engl. für „Raute", und *tag*, engl. für „Markierung") sind Wörter oder Abkürzungen, die durch ein vorangestelltes #-Symbol markiert werden. Diese Stichwörter sind nicht moderiert (jeder Nutzer kann eigene Hashtags erstellen) und dienen zur thematischen Vernetzung mit anderen Tweets beziehungsweise gleichen Themen (Parmelee & Bichard, 2012, S. 4). Über die portaleigene Suche oder andere Webdienste können auch nicht registrierte Personen gezielt nach bestimmten Hashtags suchen. Hashtags sind ein nützlicher und sehr wichtiger Mechanismus zur Verbreitung und Verknüpfung von Informationen auf Twitter (Bruns & Moe, 2014, S. 164). Nur so besteht die Möglichkeit, thematisch ähnliche Tweets miteinander zu assoziieren.

Des Weiteren gibt es einen Mechanismus, Tweets von anderen Nutzern zu favorisieren. Diese *Favorites* werden jedoch, im Vergleich zu Retweets seltener eingesetzt (Suh et al., 2010). Die Darstellung der Favorites eines Users erfolgt nicht, wie bei Retweets, auf der eigenen Profilseite im Twitter-Verlauf. Diese sind nur beim jeweiligen favorisierten Tweet aufgelistet. Dennoch ist ein Favorite ein wichtiges Kennzeichen für die Verbreitung einer Nachricht. In der Funktionsweise ist es vergleichbar mit dem *Like* auf Facebook. Tabelle 1 listet nochmal alle Konventionen auf und Abbildung 4 stellt deren Verwendung und Darstellung in einem ausgewählten Tweet dar.

Tabelle 1: Konventionen/Begriffe der Kommunikation auf Twitter

KONVENTION, BEGRIFF	BESCHREIBUNG	BEISPIEL/HINWEIS
TWEET	Kurznachricht auf Twitter, limitiert auf 140 Zeichen. Kann Links, Fotos und Videos enthalten.	Wann scheint endlich die #Sonne! Dann eben #kino… http://t.co/123458abc
MENTION	Erwähnung eines Nutzers in einem Tweet, bzw. Verknüpfung einer Nachricht mit einem Twitter-Nutzer. Vorangestelltes „@“-Zeichen bei Benutzernamen.	Im #Kino mit @musteruser :-)
REPLY	Direkte Antwort auf einen Tweet. Beginnt mit Nennung des kommentierten Nutzers.	@musteruser: Viel Spaß im Kino!
RETWEET	Teilen eines fremden Tweets durch den eignen Nutzeraccount. Nachricht enthält in der Regel „RT@username“.	RT@musteruser: Im #Kino mit musterfrau :-)
HASHTAG	Wörter oder Abkürzungen, die durch ein vorangestelltes „#“-Zeichen markiert werden. Hashtags können gesucht werden und dienen zur Verknüpfung von Themen.	Hätte Lust auf #kino #zeitvertreib #langeweile
FAVORITE	Markierung eines Tweets durch einen Nutzer, dass ihm der Tweet gefällt. Entspricht dem „Like“ auf Facebook.	*Zahl der Favorites wird unterhalb eines Tweets angezeigt (Zahl neben dem Sternchen).*
FOLLOWER	Twitter-Nutzer, der alle Tweets eines anderen Nutzers abonniert hat.	*Follower werden in der Account-Übersicht angezeigt.*
FOLLOWEE	Twitter-Nutzer, dem gefolgt wird/der abonniert wurde.	
FRIEND	Reziproke Follower-Followee-Beziehung.	*Zwei Nutzer sind gegenseitige Follower.*
DIRECT MESSAGE	Private Nachricht, die an eine Person oder Gruppe geschickt wird. Direct Messages werden nicht öffentlich angezeigt.	
LIST	Durch Nutzer verwaltete, öffentliche Liste, von anderen Accounts. Kann abonniert werden.	*Liste mit Accounts von Nachrichtenagenturen.*

Abbildung 4: Konventionen auf Twitter anhand eines Tweets durch den Regierungssprecher. Steffen Seibert (@RegSprecher) retweetet am 11. März 2015 eine Nachricht des Auswärtigen Amtes (@GermanyDiplo) anlässlich des Jahrestags der Naturkatastrophe in Japan 2011. Verwendet werden unter anderem die Hashtags #Japan, #Quake und #Tsunami. Zum Zeitpunkt der Erhebung hatte dieser Tweet 14 Retweets und 32 Favorites. Quelle: Seifert (2015).

Die Interaktion auf Twitter kann unterschiedlich typisiert werden: Anhand der Kommunikationsrichtung, der Kommunikationsebene und Kommunikationsbeziehung. Hinsichtlich der *Richtung* findet bei Twitter sowohl eine unidirektionale, als auch eine bidirektionale Kommunikation statt. Ursprünglich war der Mikroblogging-Dienst primär als Verteiler von Informationen/Neuigkeiten konstruiert (Rogers, 2014), indem Wissen unidirektional von einem Nutzer zu anderen vermittelt und multipliziert werden sollte. Abonniert ein Nutzer beispielsweise einen anderen Nutzer, werden diesem Follower nun automatisch alle Tweets des abonnierten *Followees* angezeigt. Durch die spätere Implementierung weiterer sozialer Interaktions-Funktionen haben sich die Kommunikationsmöglichkeiten jedoch ausgeweitet. Wird ein Tweet kommentiert oder eine private (direkte) Nachricht verschickt, findet eine zweiseitige Kommunikation statt.

Des Weiteren lässt sich die Kommunikation auf Twitter nach Bruns (2014) in drei Ebenen einordnen. Auf der *Mikroebene* findet die auf zwei Nutzer begrenzte, interpersonellen Kommunikation statt: Replies, Mentions und Direct Messages, wobei letztere als einzige Interaktionsform nicht öffentlich ist und somit der Mikroebene am ehesten entspricht. Die *Mesoebene* bildet alle Interaktionen zwischen einem Followee und dessen Followern ab. Diese Kommunikation ist somit auf eine spezifische, relativ konstante und abgrenzbare Nutzer-Gruppe ausgerichtet. Bruns (2014, S. 16) argumentiert, dass Tweets primär von den eigenen Followern gelesen, kommentiert und geteilt würden. Es entstünde somit eine „personal public" (Ebda, S. 17), also persönliche Öffentlichkeit eines Followees. Dieser Effekt der Zielgruppenbegrenzung verstärkt sich durch die Tatsache, dass pro Sekunde durchschnittlich etwa 5.800 Tweets veröffentlicht werden (Twitter, Inc., 2015k) und somit die Wahrscheinlichkeit gering ist, dass der Tweet von Nutzern gelesen wird, die den Verfasser nicht abonniert haben. Bei großen medialen Ereignissen, wie dem Finale der Fußball-Weltmeisterschaft am 13. Juli 2014 mit insgesamt 31,2 Millionen Tweets zum Finale, können es mehr als 600.000 Tweets pro Minute sein (Wiltshire, 2014). Zu der *Makroebene* gehört der Großteil der Kommunikation auf Twitter. Da grundsätzlich jeder Tweet durch die Öffentlichkeit gelesen, durch Hashtags gezielt gesucht und mit Themen verknüpft werden könne, seien Tweets meist Teil eines großen Kommunikationsflusses von in der Popularität schnell steigenden und fallenden Themen/Begriffen (Bruns, 2014, S. 19-20).

Die drei genannten Ebenen sollten nicht als isolierte Strukturen der Twitter-Kommunikation betrachtet werden, sondern als sich teils kreuzende oder überschneidende Kommunikationsstränge: Replies und Retweets können beispielsweise Teil einer übergeordneten Ad-hoc-Diskussion bezüglich eines Themas (verknüpft durch ein gemeinsam genutztes Hashtag) sein.

Schließlich kann die Twitter-Kommunikation noch, wie Tabelle 2 dargestellt, anhand der Beziehung typisiert werden. In Anlehnung an Konert und Hermanns (2002, S. 416) wird einerseits eine Einordnung anhand der Anzahl und Organisation der beteiligten Akteure vorgenommen (von *One-to-One* bis *Many-to-Many*), andererseits nach der Chronologie der Kommunikation (synchron oder asynchron). Die Interaktion auf Twitter erfolgt in der Regel nicht zeitgleich (synchron), wie bei einer Unterhaltung oder einem Telefonat, sondern wird unter Umständen stark zeitverzögert (asynchron) fortgesetzt. Twitter-User können zu einem beliebigen Zeitpunkt Tweets versenden, Nachrichten anderer Nutzer teilen oder kommentieren. Aufgrund der unterschiedlichen Interaktionsmöglichkeiten auf Twitter

sind auch mehrere Interaktionsbeziehungen möglich: *One-to-One* (private Nachrichten, direkte Antworten), *One-to-Few* (private Nachricht an Gruppe) beziehungsweise *One-to-Many* (normaler Tweet) sowie *Many-to-Many* (Tweet innerhalb einer per Hashtag verknüpften Ad-hoc-Öffentlichkeit) möglich.

Tabelle 2: Typisierung der Kommunikations-Beziehungen im Internet, in Anlehnung an Konert und Hermanns (2002, S. 416).

	SYNCHRONE KOMMUNIKATION (NAHEZU SIMULTAN)	**ASYNCHRONE KOMMUNIKATION (ZEITUNABHÄNGIG, VERZÖGERT)**
ONE-TO-ONE	Private Chats/Instant Messaging, Video-Chat (z.B. *Skype*)	E-Mails, Twitter: Direct Message, Reply
ONE-TO-FEW/MANY, FEW/MANY-TO-ONE	Live-Streaming, Newsticker	Webseiten, Blogs, E-Mails Twitter: Direct Message an Gruppe, Tweets an Follower
MANY-TO-MANY	Video-Konferenzen (z.B. *Google+ Hangout*), öffentliche Chat-Rooms	Foren Twitter: Hashtag-verknüpfte Unterhaltung

Charakteristisch für Online-Plattformen ist die non-konforme Textstruktur von Tweets. Wie bei Chatnachrichten oder SMS achten viele (vor allem nicht-kommerziell ausgerichtete) Twitter-Nutzer selten auf Grammatik oder Rechtschreibung. Häufig werden nur Kleinbuchstaben, Abkürzungen oder Neologismen verwendet. Auch Dialekte oder die Vermischung von Sprachen, wie zum Beispiel deutscher Fließtext mit englischen Hashtags, sowie eine fehlende Interpunktion oder eine unkonventionelle Verwendung von Sonderzeichen erschweren die Analyse von Tweets. Hashtags, Mentions und Links werden teilweise in die Satzstruktur integriert.

Abbildung 5 auf der folgenden Seite zeigt beispielhaft Tweets, die für die spätere Analyse (Kapitel 4.3) erfasst wurden. Bei einer automatisierten Analyse durch Computer müssen diese Besonderheiten berücksichtigt und − sofern möglich − bereinigt werden. Mit diesem Problem beschäftigt sich Kapitel 4.3.1 im hinteren Teil dieser Arbeit.

Abbildung 5: Typische Sprache auf Twitter anhand zweier Tweets von Jokolove (2015) und Fahrstuhlprofi (2015).

3.1.3 Datenstruktur von Tweets

Jeder Tweet besteht nicht nur aus dem ersichtlichen Tweet-Text, sondern aus einem Bündel an Meta-Daten, die sich hinsichtlich Inhalt und Umfang nach Tweet und Nutzer unterscheiden[5]. Twitter verwendet hierfür eine ungeordnete Datenstruktur im *JSON*-Format (*JavaScript Object Notation*), welches sich durch eine kompakte, leicht lesbare und schnell zu verarbeitende Textform auszeichnet. Jede Abfrage liefert Datensätze in diesem Format. Die verschachtelte Struktur ermöglicht eine einfache Zuordnung spezifischer Werte zu übergeordneten Wertegruppen. Die einzelnen Informationen werden relativ unsortiert übermittelt, sind jedoch in vier logische Objektgruppen gegliedert. Diese bündeln jeweils spezifische Informationen über User, Tweet, Informationsobjekte und – sofern angegeben – den Ort. Anhang A beschreibt die wichtigsten Felder der einzelnen Objekte.

Abbildung 6 auf der übernächsten Seite liefert einen beispielhaften Überblick über die Datenstruktur eines Tweets. Ein typischer Datensatz bezieht sich immer auf einen einzelnen Tweet, unabhängig ob es ein originärer Tweet, Retweet oder ein Reply ist. Je nach Typ werden dabei unterschiedliche Informationen zur Verfügung gestellt. Ein originärer Tweet, wie in Abbildung 4, umfasst Daten über den Tweet-Inhalt, Sprache, Zeit (und Ort) sowie den Verfasser. Bei Retweets ist zusätzlich der weitergeleitete Tweet, dessen Metadaten (wie unter anderem Favorites

[5] Anmerkung: Häufig wird irrtümlich die reine Textnachricht als Tweet bezeichnet. Streng genommen ist die Nachricht aber nur eines von vielen Merkmalen eines Tweets.

und Retweets) sowie dessen Verfasser ersichtlich, wogegen bei Replies der Umfang begrenzter ist: Hier sind nur die IDs des beantworteten Tweets und des damit verbundenen Twitter-Nutzers einsehbar.

Für inhaltliche Analysen sind vor allem die extrahierten Entities interessant. Hashtags, Hashtag-Trends, Mentions, URLs, Symbole, Bilder und Videos werden automatisch erkannt und in diesem Daten-Array aufgelistet. Zusätzlich sind Informationen über die genaue Position eines Objektes innerhalb des Tweets ersichtlich: Indizes liefern Werte über die Position des ersten und letzten Zeichens eines Objektes und damit auch über die Länge. Bei dem in Abbildung 6 dargestellten Tweet beginnt das Hashtag „#merkel“ mit Zeichen 10 und endet bei Zeichen 17, während die URL bei Zeichen 11 beginnt.

Der hohe Informationsgehalt und die Mehrebenen-Struktur von Twitter-Daten sind mit einfachen Daten-Verwaltungsprogrammen, wie *Microsoft Excel* und *Access* nur schwer zu bewältigen, weshalb Datenbank-Systeme wie SQL oder NoSQL sinnvoll sind. Vor der Datenspeicherung oder spätestens vor der eigentlichen Analyse sollten die gesammelten Daten für eine bessere Übersicht umstrukturiert werden, indem unwichtige Parameter gefiltert und bedeutende Bestandteile umcodiert werden. Es bedarf somit an Programmen oder Programmiersprachen, die die jeweiligen Operationen zur Restrukturierung des Datensatzes ermöglichen und die aufbereiteten Daten dann in Datenbanken schreiben. Eine sehr geeignete, da einfache und übersichtliche, Sprache ist Python, welche das folgende Kapitel kurz vorstellt.

Abbildung 6: Datenstruktur eines Tweets. Eigene, gekürzte Darstellung in Anlehnung an Krikorian (2010).

3.2 Programmiersprache Python

Entwickelt im Jahr 1990 durch Guido van Rossum, etablierte sich die Programmiersprache *Python* mittlerweile als Standard für deskriptive, computergestützte Studien (Millman & Aivazis, 2011, S. 9). Python ist eine interpretierte, interaktive, objekt-orientierte Programmiersprache, die eine sehr einfache und übersichtliche Syntax aufweist (Sanner, 1999, S. 3). Während die Sprache ursprünglich nicht für wissenschaftliche Zwecke gestaltet war, entstanden im Lauf der Zeit mit zunehmendem Interesse durch die Wissenschaft mehrere spezialisierte Module, wie *SciPy*, *matplotlib* und *NumPy*. Diese Pakete beinhalten etwa Funktionen zur Darstellung von Plots oder zur Ausführung einfacher, numerischer Funktionen bis hin zu komplexen Berechnungen (Millman & Aivazis, 2011, S. 10). Eines dieser Pakete ist auch *Tweepy*.

Tweepy[6] ist ein Python-Modul, das speziell zur Interaktion mit den Twitter APIs entwickelt wurde. Es unterstützt Anwender bei der Autorisierung und Durchführung von Abfragen. Zusätzlich zu den normalen Abfrage-Methoden über die REST und Streaming API stehen weitere nützliche Funktionen zur Verfügung. So berücksichtigt Tweepy bei Bedarf die Bandbreiten-Limitierung der REST API und plant beziehungsweise pausiert die definierten Requests. Tweepy wird in dieser Arbeit zur Datensammlung verwendet (siehe Listing 8, Kapitel 4.1.2).

Python weist mehrere Eigenschaften auf, die für eine wissenschaftliche Nutzung ohne vorhandene, fortgeschrittene Programmierkenntnisse von Vorteil sind: Eine intuitive, klar strukturierte Syntax, die eine gute Lesbarkeit[7] und somit auch einfachere Programmierung fördert (Russell, 2013, S. xv). Die Lernkurve ist dadurch hoch und die Einarbeitungszeit kurz. Der Programmcode ist plattformunabhängig und kann auf nahezu jedem Betriebssystem (wie *Windows*, *Mac OS* oder *Linux*) ausgeführt werden.

[6] http://www.tweepy.org/
[7] Der Programmcode liest sich wie ein Text.

Listing 1: Beispiel für ein Python-Skript

```
# Laden von Modulen
import tweepy
from slistener import tweepylistener

#Definition von Parametern
list_terms = ["#obama"]
listen = tweepylistener(api)
stream = tweepy.Stream(auth, listen, timeout=600.0)

#Definition einer Suchabfrage
while True:
    try:
        stream.filter(track=list_terms, async=False)
        break
    #Vorgehen bei Fehlern: Abbruch
    except Exception, e:
        sys.exit(1)
```

Listing 1 zeigt ein sehr einfaches Skript, das mit Python zum Sammeln von Tweets zu einem bestimmten Begriff geschrieben wurde. Das Programm bedient sich dabei an vordefinierten Klassen, die hier bereits im Modul `Tweepy` integriert sind oder vorher manuell erstellt wurden (Klasse `slistener`). Aufgrund dieser modularen Struktur können auf einfache und übersichtliche Weise Befehle ausgeführt werden, für die vorher nur wenige Parameter definiert werden müssen. In diesem Fall wurde mit `list_terms` nur ein Suchterm angegeben.

Der geschriebene Programmcode wird unmittelbar interpretiert, es entfällt also das aufwändige Kompilieren vor der Ausführung. Als objekt-orientierte Sprache erlaubt Python eine Modularisierung von Programmteilen, die zu einem anderen Zeitpunkt im Code wiederverwendet werden können, als dynamische Sprache können Paramater während der Ausführung hinzugefügt oder geändert werden (Bird, Klein, & Loper, 2009, S. xiii). Python ist zudem, im Vergleich zu ähnlichen Sprachen wie *MatLab*, als Open Source lizenziert und damit kostenlos. Idealerweise ist die Kern-Datenstruktur von Python im JSON-Format – dem gleichen Format, in dem auch Twitter-Daten zur Verfügung gestellt werden. Es gibt mittlerweile ein sehr großes Ökosystem mit vielen Programmpaketen für unterschiedlichste Zwecke – von Schnittstellen zu Datenbanken oder anderen Programmen, bis hin zu eigenständigen Bibliotheken zur Visualisierung von Daten. Besonders aufgrund der hohen Lesbarkeit und der großen Popularität bei Twitter-basierten Analysen, wird Python in dieser Arbeit als Programmiersprache verwendet.