

Jede Form der Auswertung kann nur so gut und vollständig sein wie ihre Datengrundlage. Die Datengrundlage, die sich aus plattform- und nutzergenerierten Inhalten von Social Media-Plattformen realisieren lässt, ist zumeist durch sehr große Mengen an unstrukturierten Daten aus einer Vielzahl von Quellen charakterisiert. Zur Veranschaulichung verfügten die geschätzt 3,8 Mrd. Internetnutzer im Jahr 2016 (prozentualer Anstieg im Vergleich zu 2014 jeweils in Klammern, hier ca. +25 % gegenüber 2014) weltweit knapp 2,8 Mrd. (+34 %) aktive Accounts auf den aktuell relevantesten Social Media-Plattformen (We Are Social Report 2015 und 2017). Ein Großteil der globalen Social Media-Nutzer ist hierbei in den sozialen Netzwerken Facebook, Qzone (in China) sowie bei deren Messenger-Apps Facebook Messenger, WhatsApp und QQ registriert. Geringfügig kleiner als Qzone folgen Instagram (mit 600 Mio. Nutzern etwa ein Drittel so groß wie Facebook) und Twitter (550 Mio.) sowie deutlich kleineren, aber stark zuletzt immer stärker in den Fokus der Marketeers gerutschten Plattformen wie Snapchat oder sogar LinkedIn.

Neben den an registrierten und aktiven Nutzern gemessenen größten sozialen Netzwerken existiert allerdings noch eine Vielzahl von Blogs, Foren und Videoportalen, die unter Umständen eine deutlich höhere Relevanz bezogen auf die Zielgruppe eines Unternehmens haben können. Folglich sollte bei der Auswahl der für Marketingmaßnahmen zu nutzenden sowie zur Erkenntnisgewinnung zu analysierenden Plattformen die Relevanz für das entsprechende Unternehmen berücksichtigt werden.

Infolgedessen können durch eine Marketingkampagne auf der Plattform Facebook potenziell die meisten Nutzer erreicht werden, ein intensiverer Austausch zu einem spezifischen Thema oder Inhalten erfolgt aber dagegen oftmals in themenspezifischen Foren (z. B. im Automobilbereich). Aus einer analytischen

Perspektive können Daten, die aus Foren gewonnen werden, oft besser für qualitative Analysemethoden herangezogen werden und im Vergleich zu quantitativen Auswertungen auf großen Datenmengen bessere und detailliertere Erkenntnisse liefern (z. B. zu Produktschwächen und Nutzungsszenarien). Auch haben Social Media-Plattformen in verschiedenen Ländern eine unterschiedliche Bedeutung und Relevanz. So verwenden in China durch die von der Regierung durchgesetzten Verbote kaum Nutzer Twitter oder Facebook, sondern deutlich stärker Qzone oder WECHAT, während in Russland die Social Media-Plattform vKontakte weit verbreitet ist (We Are Social Report 2015).

Zusätzlich zur Plattform spielt das verwendete Endgerät eine wichtige Rolle. Auch hier hat sich in den letzten Jahren das Nutzungsverhalten stark verändert und die Nutzung von Desktop wird immer weiter von Mobile in den Hintergrund gedrängt. Hier war zuletzt ein Zuwachs um 51 % von 2014 auf 2016 zu 2,6 Mrd. aktiven Social Media-Nutzern über Mobile zu beobachten. Mit einer anderen Zahl umschrieben: 34 % der Weltbevölkerung, die über Mobile aktiv in Social Media-Plattformen teilnehmen. In der Analyse ist es wichtig, sich den ggf. unterschiedlichen Kontext einer mobilen Endgerätnutzung zu verdeutlichen.

Das Sammeln von Social Media-Daten ist in der Regel für ein Unternehmen nur mit hohem informationstechnologischen Aufwand umsetzbar. Zwar bieten die großen Plattformen verschiedene offene Schnittstellen zur Datensammlung an, ändern aber zugleich fortlaufend technische Spezifikationen dieser Schnittstellen. So hat Facebook seit April 2015 die Freitextsuche über alle öffentlichen Posts eingestellt und liefert diese Information nur noch an das Partnerunternehmen DataSift aus. Dieses bietet wiederum eine kostenpflichtige Schnittstelle für die Daten an. Im Gegensatz zu bisherigen API können aber nicht nur alle öffentlichen Posts durchsucht werden, sondern ebenfalls auf aggregierte private Statusupdates (kein Opt-out möglich) der Nutzer zurückgegriffen werden. Letzteres wird unter dem Begriff „Topic Data“ vertrieben und ermöglicht es Unternehmen, weitreichende Erkenntnisse zu Marken, Themen und Aktivitäten im Kontext von Sentiment, Volumen und Ort zu sammeln. Um die Privatsphäre seiner Nutzer zu schützen, aggregiert Facebook die Daten immer für mindestens 100 unterschiedliche Nutzer und entfernt persönliche Informationen wie z. B. die eigene Adresse.

Plattformen ohne entsprechende Schnittstellen müssen oft aufwendig gecrawlt und die Daten prozessiert werden. Dies hat zumeist den Nachteil, dass Änderungen auf der Seite oder in der Datenstruktur nicht von den vordefinierten Prozessierungsroutinen abgefangen werden können und hierdurch Datenlücken oder im schlimmsten Fall fehlerhafte Werte erzeugt werden.

Neben der Sammlung von Inhalten und Kennzahlen zu deren Reichweite und Frequentierung, ist es aber natürlich auch von großer Bedeutung, den Social

Media zugeordneten Traffic auf der eigenen Webseite (sofern die eigene Social Media-Strategie auf ein dort verankertes Ziel ausgerichtet ist) zu messen. Gerade wenn es darum geht Budget zu allokalieren, um die unterschiedlichen zur Verfügung stehenden Marketingkanäle zu optimieren, ist die Messung des Konversionsbeitrags zu einer Zielstellung elementar.

Dark Social ist ein Begriff, der sich auf das Teilen von digitalen Inhalten außerhalb des messbaren Bereichs von Web Analyse Tools abspielt. Dies geschieht zumeist wenn ein Link per Chat oder E-Mail ausgetauscht wird und nicht über eine Social Media-Plattform, von der man einen Referrer messen kann. Sieht man sich die Einstiegspunkte ohne Referrer genauer auf der eigenen Webseite an, kann man jedoch ein ungefähres Gefühl für den Anteil von Dark Social bekommen, indem man sich auf die Seiten mit längeren, komplizierten URLs konzentriert, da diese eher unwahrscheinlich vom Nutzer manuell eingegeben werden. Berücksichtigt man diesen zusätzlichen Traffic nicht, unterschätzt man u. U. deutlich den eigenen Social Media-Kanal und dessen Einfluss.

Gerade kleinere Unternehmen verzichten mittlerweile immer häufiger auf die traditionelle Webseite und fokussieren ihre Aktivitäten ausschließlich auf ihren Social Media-Auftritt. Auch hier lassen sich entsprechende Kennzahlen definieren und messen. Die Verknüpfung der Inhalte mit dieser Performancemessung ergibt schlussendlich erst ein ganzheitliches Bild des eigenen Social Media-Auftritts.

Neben der Datensammlung ist es zudem eine enorme Herausforderung, die Datenmenge und unterschiedlichen Inhalte (Text, Foto, Video) und deren dynamische Struktur (Kommentare, Retweets, Änderungen) zu speichern. Hierfür bieten sich Big Data-Lösungen wie NoSQL-Datenbanken (bspw. Document Store für text-intensive Daten) an, die hochskalierende Daten verarbeiten können, allerdings keine umfassende Lösung für die zugrunde liegende Datenkomplexität darstellen.

Zwischen Datensammlung, Datenspeicherung und Datenabfrage stehen komplexe ETL-Prozesse (Extraktion, Transformation und Laden) zur Transformation und Bereinigung der Daten. Einfachere Probleme stellen dabei noch das Entfernen von Duplikaten oder die Spracherkennung dar. Eine höhere Komplexität weisen die Identifikation von sog. Bots, die gerade auf Twitter oftmals eingesetzt werden, und das Filtern von Spam bzw. irrelevanten Nachrichten auf. In den letzten Monaten sind gerade Bots immer wieder in den Fokus gerückt, die nicht nur Spam verbreiten, sondern gezielt zur Meinungsbeeinflussung verwendet werden, um ein falsches Stimmungsbild zu suggerieren und eigene Interessen zu verstärken. Der Umgang mit diesen Einflussfaktoren und ggf. geeigneter Gegenmaßnahmen im eigenen Netzwerk ist ein bisher ungelöstes Problem.

Speziell bei breitangelegten Umfeldanalysen auf Basis von Freitexten ist die Filterung entscheidend für den Erfolg der Auswertung. Gerade die Suche nach

im allgemeinen Sprachgebrauch weitverbreiteten Markennamen, wie z. B. MAN oder MINI, stellt kommerzielle Tools wie Analysten oft vor das Problem, entweder eine Vielzahl von potenzieller Beiträge zu verwerfen (z. B. die Suche „Case Sensitive“, also auf Groß-/Kleinschreibung achtend einzustellen) oder mit enormen, manuellen Aufwand Filterregeln zu definieren.

In diesem Beitrag werden exemplarisch Daten aus Facebook und Twitter verwendet, die zwischen Mai 2014 und April 2015 gesammelt wurden. Dieser Datensatz setzt sich aus allen öffentlichen Einträgen auf Facebook und Twitter zum Thema „Tatort“ zusammen. Der Datensatz enthält 550.000 Einträge von über 90.000 Nutzern und wurde auf alle in deutscher Sprache verfassten Nachrichten gefiltert. Die Daten wurden in einer MongoDB gespeichert und mittels R prozessiert und analysiert. Eine Übersicht ist in Tab. 2.1 dargestellt.

Die Verteilung der Geschlechter wurde über einen Abgleich der Profilnamen mit gängigen Namenslexika zu Frauen und Männern berechnet. Nicht zuordenbare Profilnamen wurden als neutral eingestuft und repräsentieren in der Regel Unternehmen und andere Organisationen.

Tab. 2.1 Übersicht zu den gesammelten Daten. Dedupliziert bedeutet, dass mehrfache Posts (z. B. Retweets) entfernt wurden. Die rechte Spalte stellt die Aufteilung der Nutzer in Anteile von Männern, Frauen und neutralen Organisationen (oder nicht zuordenbaren Profilen) dar

Tatort	#Gesamt	#Dedupliziert	#User	% m/f/n
Twitter	398.000	316.000	49.000	34 %/15 %/51 %
Facebook	153.000	153.000	43.000	37 %/26 %/37 %

Social-Media-Analyse – mehr als nur eine Wordcloud

HMD Best Paper Award 2016

Böck, M.; Köbler, F.; Anderl, E.; Le, L.

2017, XI, 28 S. 5 Abb., 3 Abb. in Farbe., Softcover

ISBN: 978-3-658-19801-5