# PART II

# Digital Journalism Studies

## Research design

# 6
# CONTENT ANALYSIS OF TWITTER
## Big data, big studies

*Cornelia Brantner and Jürgen Pfeffer*

With billions of users and hundreds of millions of posts and tweets per day, social media's big data have attracted the attention of the social sciences. This opens up unprecedented possibilities, but also necessities, to the introduction of automated analyses in order to extract meaning and gain orientation in this mass of communications. In recent years, Twitter has attracted growing attention from researchers in many disciplines. In their literature review of Twitter research, Williams et al. (2013) found 575 academic papers published between 2007 and 2011 focused on Twitter. In their literature review of journal articles on Twitter published between 2007 and 2012, Zimmer and Proferes (2014) found that of the total 382 papers analyzed, the majority (59%) stemmed from computer science or information science scholars, but communications scholars came third, accounting for 14% of the Twitter publications. And since 2012, the last year captured by the study of Zimmer and Proferes (2014), communication research on Twitter has been further refined, especially in the field of journalism research.

The relevance of Twitter for journalists and news media reflects its growing importance as a source for political information and news, in particular for younger audiences (Bastos, 2015; Newman et al., 2016; Nielsen and Schroder, 2014). According to the *news use across social media* study by Pew in 2017, for example, 11% of U.S. adults used Twitter as a news source (Shearer and Gottfried, 2017). Communications' interest in Twitter, termed an "ambient journalism" network by Hermida (2010), also reflects the important role of Twitter for both the dissemination of news and interaction with the audience, as well as for being a source in the news production process. Of special interest are new agenda-setting dynamics (Russell Neuman et al., 2014).

With respect to the news production process, newsroom observations and interviews and surveys with newsroom directors and journalists focus on how media and journalists use and incorporate Twitter in their daily working routine (e.g., Cision, 2015; El Gody, 2014; Neuberger et al., 2014a, 2014b; Thurman and Walters, 2013). For example, Neuberger et al. (2014a, 2014b) interviewed German newsroom directors on their uses of Twitter. Thurman and Walters (2013) conducted interviews with journalists from the British Guardian.co.uk news site, asking how they use and link to Twitter and content-analyzed blogs on the news site. Verweij and Noort (2014) included qualitative interviews with leading journalists and editors in their Twitter study. Since 2001 Cision, a public relations company, has published international comparisons about Twitter use in selected countries (Cision, 2015). According to the Cision's (2015) *global social*

*journalism study*, most journalists from the surveyed six countries have included social media in their daily working routine. The frequency of social media use is above 92% in all surveyed countries, with Facebook and Twitter leading. Twitter use among journalists was lowest in Germany at 43% and highest in the UK, Australia, and US, where between 71% and 75% are on Twitter. While for U.S. and UK journalists the publication and promotion of their own content is the main reason for the use of social media, journalists in Australia, Finland, Germany, and Sweden state that sourcing is the primary use (Cision, 2015). In their 2010 survey of German newsroom directors, Neuberger et al. (2014a: 349) found almost all German news departments used Twitter to attract readers (97%), for investigative purposes (94%), and for monitoring audience responses (91%). Two-thirds of newsroom directors said that they used Twitter to interact with users (66%) and for live coverage and breaking news (63%). In a follow-up study in 2014 (Neuberger et al., 2014b: 48–67), online newsroom directors ascribed to Twitter the strength of being particularly well suited to real-time interaction with their audience, for investigation, in particular for the continuous observation of prominent sources, the search for experts and the maintenance of expert networks, and for inquiry of facts. For short breaking news and for live reporting, they also favored Twitter.

When Twitter is studied as a news source for journalists (Bennett, 2016; Broersma and Graham, 2013), scholars analyze how tweets are embedded in news reporting. Several studies triangulate methods (Barnard, 2016; Deprez and Leuven, 2017; Revers, 2014; Verweij and Noort, 2014). Revers (2014), for example, combined an observation of reporting practices, interviews, and analysis of tweets to study the adoption of Twitter in the everyday working practices of reporters. In his study of journalistic practice and meta-discourse on Twitter, Barnard (2016) applied a combination of digital ethnography and content analysis. Verweij and Noort (2014) combined qualitative interviews with leading journalists and editors with a network analysis of the top 500 South African journalists on Twitter. Deprez and Leuven (2017) have analyzed the social media sourcing practices of professional health journalists. They combined in-depth interviews with health journalists and a content analysis using digital methods with manual coding.

To determine how news organizations and journalists perform on Twitter, studies use content analysis of Twitter accounts and tweets and manually code tweet content (e.g., Brems et al., 2017; Canter and Brookes, 2016; Coddington et al., 2014; Cozma and Chen, 2013; Engesser and Humprecht, 2015; Golan and Himelboim, 2016; Hanusch and Bruns, 2017; Lawrence et al., 2014; Molyneux et al., 2017; Mourão et al., 2016; Nuernbergk, 2016). Some of these studies additionally use digital methods and measure large amounts of link-, retweet-, like-, mention-, follower-, top-hashtag-structures and/or network-structures of media's and/or journalists' Twitter accounts (e.g., Chorley and Mottershead, 2016; Enli and Simonsen, 2017; Hahn et al., 2015; Larsson and Hallvard, 2015; Majó-Vázquez et al., 2017; Nuernbergk, 2016; Vergeer, 2015).

Whereas the latter-mentioned studies harvest data from preselected media and/or journalists' Twitter handles, other authors (e.g., Groshek and Tandoc, 2017; Kirilenko and Stepchenkova, 2014) harvest and automatically analyze the Twitter discourses on certain topics and/or hashtags and also look at the contribution of news media to these debates. For example, Groshek and Tandoc (2017) and Kirilenko and Stepchenkova (2014) manually code the most important users that contribute to debates, noting whether they are media or journalists or other users in order to analyze the role and importance of professional media and journalists in these debates.

Faris et al. (2016: 5855) utilized a mixed-methods approach combining link analysis with qualitative content analysis in order to analyze the evolution of the net neutrality policy debate and thereby assess the role, reach, and influence of different media sources. However, in their qualitative content analysis, they did not analyze the tweet texts but rather the linked stories. In order to assess the media contributions on Twitter they analyzed the shared links and found that among the top 25 shared stories only three came from mainstream news media. Yet instead of

using automated approaches for attributing whether an account or a URL link is from media or journalists, these studies assign it manually.

Most of the studies mentioned analyze preselected media and journalists' social media handles and harvest their tweets, applying qualitative or quantitative content analysis or digital methods including network analysis, but hardly any of these studies use other automated text-analysis approaches, such as topic modeling, sentiment analysis, or machine learning. Malik and Pfeffer (2016) point out that there have been only a few studies bringing computational analysis to the study of news organizations and journalists' use of social media. Among the exceptions are Zamith and Lewis (2015), who address big data studies in communications and social science, and Flaounas et al. (2013), who apply automated content analysis to digital journalism. Yet these authors do not analyze Twitter text content. The purpose of a study by Guo et al. (2016: 332) was to "evaluate the efficacy and validity of different computer-assisted methods for conducting journalism and mass communication research." They utilize and compare unsupervised topic modeling (LDA analysis, see later in this article) and a dictionary-based analysis (search and annotate predefined words in texts) on 77 million tweets related to the 2012 U.S. presidential elections. The authors extracted topics and studied the association of the two candidates, Obama and Romney, to these topics. They identify the advantages and disadvantages of both techniques and conclude that overall, LDA analysis performed better than the dictionary-based analysis. Yet their analysis focused on the text content but did not examine the tweet authors; thus, no inferences could be made about the contribution of journalists or media organizations to the Twitter debate. Using a list of more than 6,000 pre-identified news media and journalists' Twitter handles, Malik and Pfeffer (2016) studied 1.8 billion tweets and found less than 1% of Twitter content is news-media related and that news organizations mainly use Twitter as a professionalized, one-way communication medium to promote their own reporting. However, as they themselves state, by using a predetermined list they probably underestimate the proportion of news media. Raghuram et al. (2016) suggest an automated solution for the endeavor of tweet author detection: they show how several machine learning algorithms (including the support vectors machine, which will be discussed later) can be deployed for classifying Twitter accounts and categorize them into six user groups, namely politics, entertainment, entrepreneurship, journalism, science and technology, and health care.

In this chapter we showcase some of the automated content analysis approaches for analyzing large-scale Twitter data, namely sentiment analysis, network text analysis, topic modeling, and machine-learning-driven text classification. We also debate the strengths and weaknesses of these methods. First, we discuss briefly the strategies that are used to collect Twitter data, followed by the steps necessary to preprocess and prepare the data for automated content analysis.

## Harvesting Twitter data

The way tweets are gathered in large numbers is by utilizing Twitter's API. This acronym stands for application programming interface, i.e., a direct connection to Twitter's data that can be accessed with programming code.

Tweets can be collected in real time or in retrospect. Two different real-time data collection APIs are available. First, the Sample API provides a random 1% sample of all tweets worldwide – at the time of writing this article, this was about 3.5 million tweets per day. Second, with the Filter API we are able to collect more specialized data by defining search terms. The Filter API can handle user accounts (collect all tweets from these accounts), words (collect all tweets that include at least one of the selected words), and geographic-boundary boxes (collect all tweets sent from within this geographic area). The REST API can collect historic tweets from specified users or the follower/followee lists of users. Filtering tweets based on keywords is only possible in real time.

The characteristics of these data access points drive the data collection strategies of researchers. If the Filter API is used to collect tweets from a list of user accounts, researchers need to predefine that list, such as a list of journalists' Twitter handles. This allows for analysis of how journalists act on Twitter and how their tweets are disseminated (see previous examples, e.g., Chorley and Mottershead, 2016; Majó-Vázquez et al., 2017; Nuernbergk, 2016; Vergeer, 2015). On the other hand, if we want to study the role of journalism in Twitter discourses related to specific topics, tweets from all Twitter users are collected based on predefined lists of keywords and hashtags (e.g., Groshek and Tandoc, 2017; Kirilenko and Stepchenkova, 2014). Finally, the REST API's capability of collecting followers of any user accounts can be used to create follower-networks and to analyze the position of journalists and media outlets in these networks.

All these data access points are free of charge and open to all researchers and practitioners (though other fee-based options are available). The programming code for accessing Twitter data is well developed in many programming languages and tools, so all of the previously described approaches to collect tweets can be implemented with 10–20 lines of code, found easily in the web. This easy and free access is the key reason why Twitter became the predominant data source for social media studies.

All of these approaches of data collection also have limitations typical of social media data (Ruths and Pfeffer, 2014). Morstatter et al. (2013) show that tweets collected with the Filter API do not necessarily represent the overall activity on Twitter, and the proportion of tweets provided is not stable. Searching for tweets in certain geographic areas is biased by the fact that only a fraction of users allows adding geographic information to their tweets.

## Data preprocessing

To illustrate different quantitative content analysis approaches, we utilize data from Malik and Pfeffer's (2016) Egypt case study. The dataset consists of about 105,000 tweets written in English from March to June 2014, including the hashtag "#Egypt." The authors assembled 6,103 news/journalism-related Twitter handles from news media websites, as well as from Twitter-handle white pages, and identified 8% of tweets in this dataset that were news related, i.e., were either tweeted by news media and journalists, mentioned news media, or linked to websites from media outlets. The authors applied LDA topic modeling (Blei et al., 2003) to identify topics and to show media-related tweets were over-proportionally represented in topics related "to journalism and press freedom" (Malik and Pfeffer, 2016: 16). In the following we describe the text preprocessing steps that are applied for most automated content analysis procedures. Specific steps have differing importance depending on language; we focus on texts in English.

*Strip case.* A straightforward first step of text preparation is to make all letters lower case. This is done to make sure that, for instance, "Egypt", "egypt", and "EGYPT" are handled as the same word. Sometimes analysts are particularly interested in proper nouns. Some languages, e.g., German, have many more words with an initial capital letter that might be of interest for the analysis of the texts. In this case, stripping cases should be handled more deliberately.

*Tokenization.* The process of splitting up a given text into a set of words is called tokenization. For short texts like tweets, this includes the removal of punctuation. For longer texts, a sentence or a paragraph can be the entity of analysis. Then, these elements must be preserved. Specialized tokenizers have been developed that handle Twitter data. We used TweetTokenizer from the Natural Language Toolkit (NLTK)[1] for Python (Loper and Bird, 2002) to tokenize tweets and also remove users mentioned in the tweets.

*Delete list.* Twenty-five percent of English texts consist of only 17 words, e.g. "the," "a," "is," etc. When we analyze texts, we are looking for terms that distinguish different texts, words that

occur everywhere impede that effort. Stopword lists include these stopwords as well as discourse markers. We utilize Ranks NL's "default English stopwords list" and its "MySQL stopwords list,"[2] resulting in a delete list of 555 words. We also removed web-links from the tweets and the case-stripped token "#egypt," since this was the search term for our tweet collection and can be found in every tweet. We did not remove the hashtag symbol from any hashtags.

*Term normalization* is the process of unifying words that should be identical but are written (slightly) differently. The most important procedure for normalizing words is called stemming and refers to the removal of inflectional endings of words so that singular and plural as well as different verb forms map to the same term (Porter, 1980). Stemming is not often used for text preprocessing, as reducing words to their stems can result in hard-to-understand terms and can sometimes change meaning. In inflection-rich languages, as in French, term normalization is more important and more complicated.

*Deduplication.* Deduplication removes multiple occurrences of the same text. Deduplication can be performed on texts that are "exactly" identical. While rarely the case for longer texts, it is a reasonable approach for tweets. Deduplication of "almost" identical texts is computationally very expensive and can take days for large numbers of texts, while exact deduplication can be quickly done. For this case study, we apply exact deduplication after the already described preprocessing steps. This removed 40% of news-related tweets and 37% of non-news tweets. In contrast, if we were interested in analyzing the importance of certain users or topics, we would be interested in tweets occurring multiple times resulting from retweets or from multiple references to one online news article. Then, deduplication would be counterproductive.

After performing these preprocessing steps, the remaining tweet corpora consists of 5,101 news-related tweets with 460,000 words, and 60,168 non-news-related tweets with 6.2 million words. A very common and easy-to-create first analysis for obtaining overview and orientation in the text are word clouds. Figure 6.1 shows two word clouds[3] of the top 100 most frequent terms from the two sets. Word clouds resemble visual representations of frequency lists – a larger font for terms occurring more often in the text – and the position of a word is purely to optimize the visual aesthetics. Without discussing any details, the word clouds support Malik and Pfeffer's (2016) observation that news-related tweets were especially concerned with Al Jazeera journalists who were arrested in late December 2013. Two more technical details related to the creation of these word clouds should be mentioned. First, we did not remove numbers, which is a frequently used option for word clouds. Consequently, the term "529", representing the number of Morsi supporters who were sentenced to death, is visible. Second, the tool that we used to create these figures could not visualize Arabic text but instead showed special characters, e.g. "ØµØ".

(a)


*Figure 6.1* Word clouds with top 100 words from (a) 5,101 news-related tweets and (b) 60,168 non-news-related tweets

(b)


*Figure 6.1* (Continued)

*Table 6.1* Sentiment analysis categories comparing news-related tweets with non-news-related tweets

| *LIWC category* | *Positive emotions* | *Negative emotions* | *Anxiety* | *Anger* | *Sadness* |
|---|---|---|---|---|---|
| *Example words* | *happy, good* | *hate, enemy* | *nervous, afraid* | *hate, kill* | *grief, cry, sad* |
| News-related | 3.73 | 4.91 | 0.88 | 2.79 | 0.70 |
| Other tweets | 3.97 | 3.88 | 0.68 | 2.16 | 0.57 |

## Sentiment analysis

Sentiment analysis is frequently used to describe the importance of certain emotional and other word categories in texts (Pang and Lee, 2008). At its core, sentiment analysis is counting the frequency of words that have been annotated with categories. LIWC (pronounced like "Luke"), which stands for linguistic inquiry and word count (Pennebaker et al., 2007), is the most commonly used sentiment dictionary and comes with an easy-to-use tool. We load our two groups of tweets into the LIWC tool to classify the tweets' words.

Table 6.1 shows a selection of the results and contrasts news-related tweets with their non-news-related counterparts. Tweets discussing negative emotions seem to be more prevalent in news-related tweets. Here, analysis was not at the tweet level. Instead two large *documents* were analyzed, one consisting of all news-related tweets and one with all the other tweets.

## Network text analysis

Social network analysis is interested in social actors (individuals, organizations, etc.) and their relationships as well as topological structures emerging from these relationships. Key research question ask for important (central) actors, or how the network can be fragmented into groups. Network text analysis is concerned with similar questions, and analyzing text as networks predates the advent of social media (Roberts and Popping, 1996). Networks consist of nodes and edges (Hennig et al., 2012). In text networks, nodes represent words and edges depict connections among words. A connection between two words is created when two words co-occur in

the same sentence, paragraph, or text. With tweets, all words of a single tweet are connected. The advantage of treating words as nodes and co-occurrence as edge is that this allows us to analyze text networks with network analysis tools and methods (Wasserman and Faust, 1994). In other words, we can determine which words are central, which words form groups (topics?), and which words connect those groups.

The 65,000 tweets for this case study create 2.4 million links among words. In general, networks extracted from large numbers of texts tend to be big (many nodes) and dense (many links). These large numbers create their own challenges for handling and analyzing the networks, a discussion of which is beyond the scope of this article. For the sake of presenting the method, we focus on the most important words and connections of our corpora. Figure 6.2 shows the top 20 words and top 30 co-occurrence links of each group of tweets. The size of the nodes shows the number of occurrences in each of the two groups of tweets.

In Figure 6.2, network text analysis provides us with (a) framing information for every term by showing its connections and (b) a structural overview of the content by revealing global positions of terms (central vs. peripheral) as well as groups of terms, which increases the readability of the analysis. For instance, three topics are clearly visible in the news-related figure. Different topological structures of text networks can be used to study and interpret different narratives (Bearman and Stovel, 2000) or reveal how stories from news outlets overlap with those of eyewitness accounts (Martin et al., 2013). We could further analyze different terms, framing of terms, or groups of terms by focusing on different or larger parts of the text network. Other text networks can be extracted from tweets. For instance, a network connecting user accounts to every word that they use would allow us to study user groups created by overlapping sets of words.

## Topic modeling

The goal of topic modeling is to identify topics in large text corpora. Latent Dirichlet allocation (LDA) (Blei et al., 2003) is the most commonly used topic modeling technique. In a nutshell, LDA creates a predefined number of groups as well as a probability for every document and for every word in the corpora to belong to these topics. The groups can be interpreted as topics, and the result of an LDA calculation often shows the top words with highest probability for every topic, which is helpful for interpreting the topics. The algorithm behind calculating LDA topics is mathematically challenging, but several tools as well as packages for programming languages exist, so that LDA can be performed without knowing all the technical details. Based on their exemplary analysis of the U.S. presidential election tweets from 2012, Guo et al. (2016) conclude that LDA topic modeling performs better than more traditional dictionary-based techniques. However, they also stress the necessity of human interventions in LDA analysis to avoid topic allocation errors.

For the topic modeling example and for the machine learning example in the following section, we created a modified dataset. We concatenated all tweets per user account with at least two tweets. This resulted in 922 texts from the news-related tweets, each representing all tweets of a single user. To limit the complexity of these examples, we select a random subsample of the non-news-related users to have two equally sized datasets. We ran two different LDA calculations, one for each dataset of texts, with the tool MALLET,[4] a machine learning toolkit for language. Table 6.2 shows the results. Every topic is represented by words that are strongest in their association with the topic. Interestingly, topics from non-news-related tweets have a larger number of hashtag terms.

Our example can be used to discuss some of the major limitations. First, in traditional LDA, the number of topics must be predefined as a parameter for the algorithms. Normally, we do not know whether five or 20 or any other number of topics is a good representation for our texts.

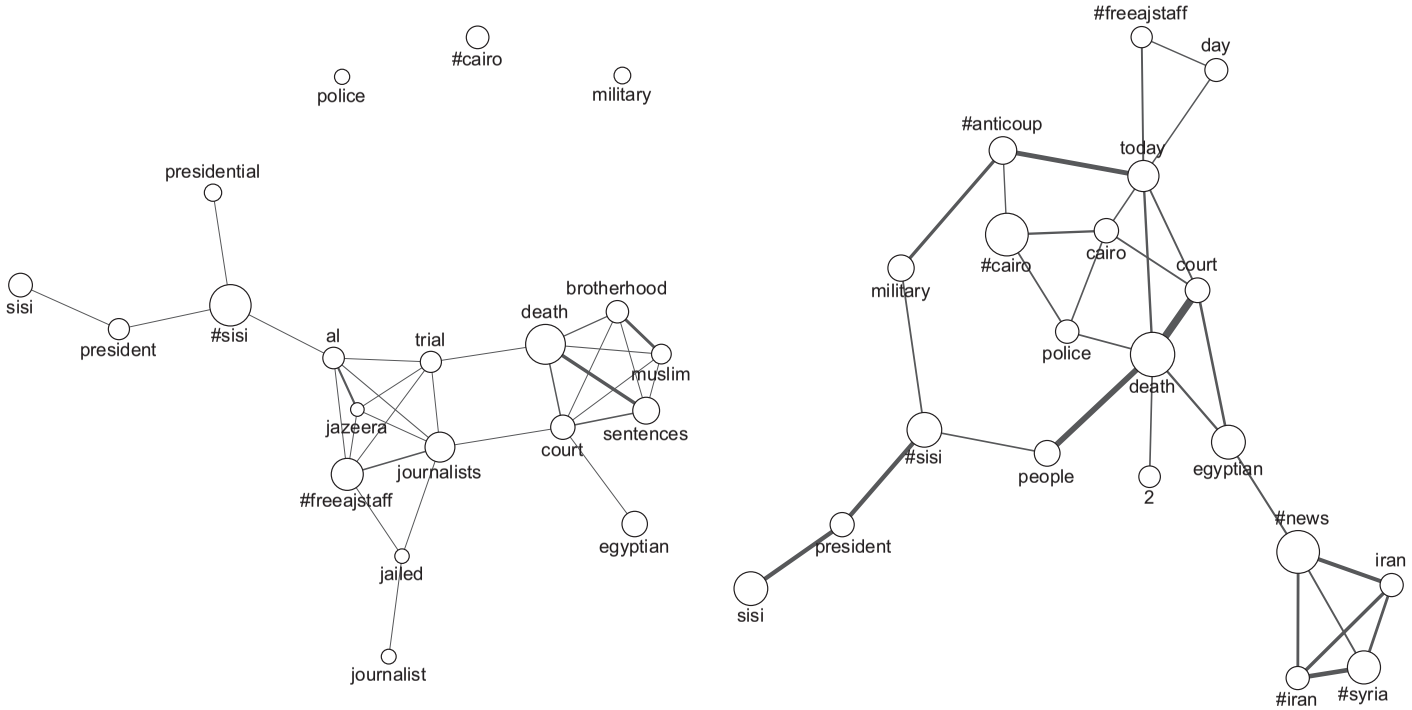


*Figure 6.2* News-related tweets (left); other tweets (right). Network of top 20 word occurrences in both groups of tweets as well as their 30 strongest connections in terms of co-occurrence within tweets.

*Table 6.2* Top 10 words associated with five topics from news-related as well as other tweets

| # | *Topics from news-related tweets* | *Topics from other tweets* |
|---|---|---|
| 1 | #freeajstaff #cairo journalists journalist al trial #مصر jailed jazeera #فن | #sisi egyptian sisi president elections brotherhood presidential #cairo killed muslim |
| 2 | military #libya cairo egyptian #tunisia #westernsahara #morocco #us aid killed | death people military sentenced court years #freeajstaff journalists today coup |
| 3 | journalists #syria #israel deadly torture #iraq military prison nations piece | #uae #kuwait #saudi #ksa #ff free #bahrain watch #qatar #syria |
| 4 | #sisi president presidential election sisi vote elections sexual #breakingnews al | #travel #tourism #photography egypt #art #design #discover_egypt_come #journey #cairo #welcometoegypt |
| 5 | death brotherhood muslim sentences court police sisi supporters mass coup | iran #iran #iraq #maryamrajavi #syria #news #world #cnn #android #iphone |

Second, LDA is a probabilistic algorithm. Different runs with the same dataset and settings will result in (slightly) different results. Third, all words are assigned to all topics; the difference is the order (probability) of words. Consequently, words often occur multiple times in the topics represented by the top words for each topic. Finally, humans are very good at identifying patterns and *Gestalt* (Koffka, 1935), even if there are none.

## Machine learning approaches to text classification

A large number of machine learning approaches can be applied to Twitter text (Raghuram et al., 2016). In general, we distinguish between supervised and unsupervised machine learning approaches. For supervised learning algorithms, the goal is primarily to predict a certain continuous variable (regression) or class (classification). It is called "supervised" because data are available for which the correct solution ("golden truth") is known. Unsupervised learning algorithms do not utilize target variables that should be predicted. Instead, inherent characteristics of the data (e.g. groups, topics) should be revealed. For our data, we know for every tweet whether it is news-related or not, based on Malik and Pfeffer's (2016) definition. Consequently, we can use this to *train* a machine learning model that we can then apply to new tweets to automatically classify them into these two groups. An essential preprocessing step for every machine learning algorithm is to extract *features* (variables) from the dataset that will serve as independent variables. The most straightforward approach is to use every single word in a text corpus as a feature and count how often every word occurs in every text. It is common for machine learning approaches to have more variables than cases.

Support vector machines (SVM) are very popular algorithms for classifying elements because they perform well for many different datasets (Marsland, 2009). Imagine a two-dimensional plot with body weight and body height on the axes and the respective data points from 25 men and 25 women. In this plot, men and women would most likely form two (overlapping) clusters of points. An SVM for this dataset is the best straight line to separate these two groups. The *accuracy* of the SVM is an assessment about how many data points are on the wrong side of the separation line.

Table 6.3 shows a typical result presentation of a machine learning classification model. We used the scikit-learn[5] package in the Python programming language to perform the calculations, and we aggregated all tweets of a single user to one case for our classification model. The accuracy of a prediction model is normally described with two values. *Precision* is the proportion of

*Table 6.3* Accuracy of the Support Vector Machine for predicting whether a user account is news-related or not

| | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|
| News-related | 0.79 | 0.85 | 0.82 |
| Other tweets | 0.84 | 0.78 | 0.80 |
| Average/total | 0.81 | 0.81 | 0.81 |

classified elements for a group that are classified correctly. Tweets that were predicted as being non-news-related are correctly predicted at a higher rate than news-related tweets. *Recall* defines the proportion of elements of a group that can be classified correctly. These accuracy metrics measure different aspects of the prediction and can be combined to the F1-score, which represents the harmonic mean of precision and recall.

In this example, we classified accounts for which we already knew the correct classification to show that the content of these two different groups of accounts is different enough to create these groups. In a next step, we could classify new tweets from new accounts based on the previously trained classification model. Raghuram et al. (2016) use several machine learning algorithms to classify Twitter accounts into six different groups, including journalism. While this allows for classifying large numbers of accounts based on a rather small manually classified training dataset, critical analyses of machine learning for classifying user groups show the quality of the results can vary heavily if the training and the testing dataset show different characteristics. For instance, Cohen and Ruths (2013) demonstrated that classifying the political orientation of Twitter users works well with accounts from politically active users but performs very poorly when non-activists need to be classified.

## Discussion

The dissemination of computational methods for studying news media-related content on Twitter has so far been limited (Malik and Pfeffer, 2016). However, with the prevalence of social media and the growing importance of platforms such as Twitter for the everyday work of journalists and publishers, an increased interest in computational methods for digital journalism studies can be expected. A major reason why analyzing tweets is popular in many scientific fields is that data can be accessed easily and without cost via Twitter's APIs. But while data access is easy, Twitter data, as with data from other social media outlets, bring tremendous challenges and pitfalls that can jeopardize the reliability of research relying on those data sources (Ruths and Pfeffer, 2014).

In this chapter, we showcased and discussed methods that are used to analyze text from millions of tweets. The application of most of these approaches is technically not very challenging. For instance, once the data are preprocessed into the right format, programming the support vector machine for text classification and reporting the precision/recall matrix takes four (!) lines of code. We used a machine learning toolkit that requires few technical skills. A sentiment analysis can be wholly accomplished with an easy-to-use tool, without any coding necessary. A sufficient number of user-friendly tools for social network analysis are also available for free. So, while these methods are relatively easy to use, some of them are algorithmically very complex and almost impossible to comprehend in detail for researchers from most fields. This leads to the biggest issue related to computational methods – researchers deploying methods without considering their limitations or preconditions for the data. For instance, individual tweets are often regarded as too short for useful topic models (Hong and Davison, 2010); nevertheless, the literature is rich in LDA studies based on individual tweets.

A major conceptual issue regarding automated content analysis is the mismatch between what these methods are expected to do (analyze sentiment, find topics) and what they actually do (count adjectives, count word-co-occurrences). Automated content analysis methods are far from *understanding* text. Instead, text is treated as a *bag of words*. Written text is loaded with culture, context, and linguistic complexity, none of which can be studied with the methods described in this article. Additionally, text preprocessing and content analytical methods are optimized for English language texts. Text in other languages with more complex inflections and word composites can require intensive manual tasks, e.g., cleaning text and compiling a dictionary, in order to create meaningful results.

The reduction of messages to bags of words also ignores another important feature of social media communication: images. The increasing trend toward "distributed content" (Newman et al., 2016) in journalism is linked to the increasing importance of visual elements in digital journalism and to increased space dedicated to visuals within journalistic content circulated via social media, such as Twitter. Images can have many roles in messages: they can be rather decorative elements, illustrate what is shown in the verbal text, or convey central arguments of the message visually (Brantner et al., 2011). Automatically analysis of image content is developing but is mainly concentrated on the motifs of visuals. It is a first step, but only partially helpful, because images do not convey messages only by *what* is shown but also by *how* it is shown – features that can contribute to the visual framing of issues and their evaluations (see: Coleman and Wu, 2016). Approaches to automated analysis of image content can be found mainly in computer science, especially in the fields of image processing, pattern recognition, and computer vision, but have recently been used also in communication research (see: Etlinger, 2017; Peng, 2017). Another methodological challenge is that visuals and verbal text are complexly entangled in multimodal media messages and influence each other.

In accordance with Lewis and colleagues (2013; Zamith and Lewis, 2015), we encourage journalism researchers dealing with big data from social media to combine computational and manual methods. That is, computational methods have to complement established research techniques instead of replacing them. For example, a deliberate triangulation of the computational analysis of big data with manually coded smaller samples can bring in the advantages of both methodologies. The combination of proven and innovative social sciences methods with computational methods is indispensable to ensure validity and reliability and to further enhance theoretically and empirically well-founded assessments of the current state of (online) journalism.

## Further reading

For analysis of the advantages of LDA analysis compared to dictionary-based text analysis for Twitter research see Guo et al. (2016); for the utilization of machine learning algorithms to classify Twitter accounts see Raghuram et al. (2016); see Lewis et al. (2013) and Zamith and Lewis (2015) on triangulation of computational and manual methods. Malik and Pfeffer (2016) provide a case study of journalistic agenda setting on Twitter applying LDA topic modeling. Ruths and Pfeffer (2014) summarize challenges arising from using social media data to study human behavior.

## Notes

1 www.nltk.org/
2 www.ranks.nl/stopwords
3 Created with www.wordle.net/
4 http://mallet.cs.umass.edu/
5 http://scikit-learn.org/stable/index.html

## References


Barnard, S. R. (2016) "'Tweet or be sacked': Twitter and the new elements of journalistic practice." *Journalism*, 17(2), 190–207.

Bastos, M. T. (2015) "Shares, pins, and tweets: News readership from daily papers to social media." *Journalism Studies*, 16(3), 305–325.

Bearman, P. S. and Stovel, K. (2000) "Becoming a Nazi: A model for narrative networks." *Poetics*, 27(2), 69–90.

Bennett, D. (2016) "Sourcing the BBC's live online coverage of terror attacks." *Digital Journalism*, 4(7), 861–874.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) "Latent dirichlet allocation." *The Journal of Machine Learning Research*, 3, 993–1022.

Brantner, C., Lobinger, K. and Wetzstein, I. (2011) "Effects of visual framing on emotional responses and evaluations of news stories about the Gaza conflict 2009." *Journalism & Mass Communication Quarterly*, 88(3), 523–540.

Brems, C., Temmerman, M., Graham, T. and Broersma, M. (2017) "Personal branding on Twitter." *Digital Journalism*, 5(4), 443–459.

Broersma, M. and Graham, T. (2013) "Twitter as a news source: How Dutch and British newspapers used tweets in their news coverage, 2007–2011." *Journalism Practice*, 7(4), 446–464.

Canter, L. and Brookes, D. (2016) "Twitter as a flexible tool." *Digital Journalism*, 4(7), 875–885.

Chorley, M. J. and Mottershead, G. (2016) "Are you talking to me?" *Journalism Practice*, 10(7), 856–867.

Cision. (2015) *2015 Global Social Journalism Study*. Retrieved from www.cision.com/us/resources/white-papers/2015-global-social-journalism-study/

Coddington, M., Molyneux, L. and Lawrence, R. G. (2014) "Fact checking the campaign: How political reporters use Twitter to set the record straight (or not)." *The International Journal of Press/Politics*, 19(4), 391–409.

Cohen, R. and Ruths, D. (2013) "Classifying political orientation on Twitter: It's not easy!" *Proceedings of International AAAI Conference on Weblogs and Social Media* (pp. 91–99).

Coleman, R. and Wu, D. (2016) *Image and Emotion in Voter Decisions: The Affect Agenda*. Lanham, MD: Lexington Books.

Cozma, R. and Chen, K.-J. (2013) "What's in a tweet? Foreign correspondents' use of social media." *Journalism Practice*, 7(1), 33–46.

Deprez, A. and Leuven, S. V. (2017) "About pseudo quarrels and trustworthiness." *Journalism Studies*, 1–18.

El Gody, A. (2014) "The use of information and communication technologies in three Egyptian newsrooms." *Digital Journalism*, 2(1), 77–97.

Engesser, S. and Humprecht, E. (2015) "Frequency or skillfulness." *Journalism Studies*, 16(4), 513–529.

Enli, G. and Simonsen, C.-A. (2017) "'Social media logic' meets professional norms: Twitter hashtags usage by journalists and politicians." *Information, Communication & Society*, 1–16.

Etlinger, S. (2017) "Language of the eye: How computer vision is remaking social media." Presented at the *Annual Meeting of the International Communication Association*, San Diego, CA.

Faris, R., Roberts, H., Etling, B., Othman, D. and Benkler, Y. (2016) "The role of the networked public sphere in the U.S. net neutrality policy debate." *International Journal of Communication*, 10, 5839–5864.

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J. and Cristianini, N. (2013) "Research methods in the age of digital journalism: Massive-scale automated analysis of news-content – Topics, style and gender." *Digital Journalism*, 1(1), 102–116.

Golan, G. J. and Himelboim, I. (2016) "Can world system theory predict news flow on Twitter? The case of government-sponsored broadcasting." *Information, Communication & Society*, 19(8), 1150–1170.

Groshek, J. and Tandoc, E. (2017) "The affordance effect: Gatekeeping and (non)reciprocal journalism on Twitter." *Computers in Human Behavior*, 66, 201–210.

Guo, L., Vargo, C. J., Pan, Z., Ding, W. and Ishwar, P. (2016) "Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling." *Journalism & Mass Communication Quarterly*, 93(2), 332–359.

Hahn, K. S., Ryu, S. and Park, S. (2015) "Fragmentation in the Twitter following of news outlets: The representation of South Korean users' ideological and generational cleavage." *Journalism & Mass Communication Quarterly*, 92(1), 56–76.

Hanusch, F. and Bruns, A. (2017) "Journalistic branding on Twitter." *Digital Journalism*, 5(1), 26–43.

Hennig, M., Brandes, U., Pfeffer, J. and Mergel, I. (2012) *Studying Social Networks: A Guide to Empirical Research*. Frankfurt: Campus Verlag.

Hermida, A. (2010) "Twittering the news: The emergence of ambient journalism." *Journalism Practice*, 4(3), 297–308.

Hong, L. and Davison, B. D. (2010) "Empirical study of topic modeling in Twitter." *Proceedings of Workshop on Social Media Analytics* (pp. 80–88).

Kirilenko, A. P. and Stepchenkova, S. O. (2014) "Public microblogging on climate change: One year of Twitter worldwide." *Global Environmental Change*, 26, 171–182.

Koffka, K. (1935) *Principles of Gestalt Psychology*. New York, NY: Harcourt Brace & World.

Larsson, A. O. and Hallvard, M. (2015) "Bots or journalists? News sharing on Twitter." *Communications*, 40(3), 361–370.

Lawrence, R. G., Molyneux, L., Coddington, M. and Holton, A. (2014) "Tweeting conventions: Political journalists' use of Twitter to cover the 2012 presidential campaign." *Journalism Studies*, 15(6), 789–806.

Lewis, S. C., Zamith, R. and Hermida, A. (2013) "Content analysis in an era of big data: A hybrid approach to computational and manual methods." *Journal of Broadcasting & Electronic Media*, 57(1), 34–52.

Loper, E. and Bird, S. (2002) "NLTK: The Natural Language Toolkit." *Proceedings of Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics – Volume 1* (pp. 63–70).

Majó-Vázquez, S., Zhao, J. and Nielsen, R. K. (2017) "The digital-born and legacy news media on Twitter during the French presidential election." *Reuters Institute for the Study of Journalism*. Retrieved from http://reutersinstitute.politics.ox.ac.uk/our-research/digital-born-and-legacy-news-media-twitter-during-french-presidential-elections

Malik, M. M. and Pfeffer, J. (2016) "A macroscopic analysis of news content in Twitter." *Digital Journalism*, 4(8), 955–979.

Marsland, S. (2009) *Machine Learning: An Algorithmic Perspective*. Boca Raton, FL: Taylor & Francis Ltd.

Martin, M. K., Pfeffer, J. and Carley, K. M. (2013) "Network text analysis of conceptual overlap in interviews, newspaper articles and keywords." *Social Network Analysis and Mining*, 3(4), 1165–1177.

Molyneux, L., Holton, A. and Lewis, S. C. (2017) "How journalists engage in branding on Twitter: Individual, organizational, and institutional levels." *Information, Communication & Society*, 1–16.

Morstatter, F., Pfeffer, J., Liu, H. and Carley, K. M. (2013) "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose." *Proceedings of International Conference on Weblogs and Social Media* (pp. 400–408).

Mourão, R., Diehl, T. and Vasudevan, K. (2016) "I love big bird. How journalists tweeted humor during the 2012 presidential debates." *Digital Journalism*, 4(2), 211–228.

Neuberger, C., Jo vom Hofe, H. and Nuernbergk, C. (2014a) "The use of Twitter by professional journalists. Results of a newsroom survey in Germany." In K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang (pp. 345–357).

Neuberger, C., Langenohl, S. and Nuernbergk, C. (2014b) "Redaktionsbefragung." In C. Neuberger, S. Langenohl and C. Nuernbergk (eds.), *Social Media und Journalismus*. Düsseldorf: LfM (pp. 34–91).

Newman, N., Fletcher, R., Levy, D. A. L. and Nielsen, R. K. (2016) "Reuters Institute digital news report 2016." *Reuters Institute for the Study of Journalism*. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital-News-Report-2016.pdf

Nielsen, R. K. and Schroder, K. C. (2014) "The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison." *Digital Journalism*, 2(4), 472–489.

Nuernbergk, C. (2016) "Political journalists' interaction networks." *Journalism Practice*, 10(7), 868–879.

Pang, B. and Lee, L. (2008) "Opinion mining and sentiment analysis." *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.

Peng, Y. (2017) "When images meet codes: Applying computer vision methods in communication research." Presented at the *Annual Meeting of the International Communication Association*, San Diego, CA.

Pennebaker, J. W., Booth, R. J. and Francis, M. E. (2007) *Linguistic Inquiry and Word Count: LIWC* [Computer software]. Austin, TX: Liwc. Net.

Porter, M. F. (1980) "An algorithm for suffix stripping." *Program*, 14(3), 130–137.

Raghuram, M. A., Akshay, K. and Chandrasekaran, K. (2016) "Efficient user profiling in Twitter social network using traditional classifiers." In S. Berretti, S. M. Thampi and S. Dasgupta (eds.), *Intelligent Systems Technologies and Applications* (Vol. 2). Berlin: Springer-Verlag Berlin (pp. 399–411).

Revers, M. (2014) "The twitterization of news making: Transparency and journalistic professionalism." *Journal of Communication*, 64(5), 806–826.

Roberts, C. W. and Popping, R. (1996) "Themes, syntax and other necessary steps in the network analysis of texts: A research paper." *Social Science Information*, 35(4), 657–665.

Russell Neuman, W., Guggenheim, L., Mo Jang, S. and Bae, S. Y. (2014) "The dynamics of public attention: Agenda-setting theory meets big data." *Journal of Communication*, 64(2), 193–214.

Ruths, D. and Pfeffer, J. (2014) "Social media for large studies of behavior." *Science*, 346(6213), 1063–1064.
Shearer, E. and Gottfried, J. (2017, September 7) *News Use Across Social Media Platforms 2017*. Retrieved from www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/
Thurman, N. and Walters, A. (2013) "Live blogging? Digital journalisms pivotal platform? A case study of the production, consumption, and form of Live Blogs at Guardian.co.uk." *Digital Journalism*, 1(1), 82–101.
Vergeer, M. (2015) "Peers and sources as social capital in the production of news: Online social networks as communities of journalists." *Social Science Computer Review*, 33(3), 277–297.
Verweij, P. and Noort, E. van. (2014) "Journalists' Twitter networks, public debates and relationships in South Africa." *Digital Journalism*, 2(1), 98–114.
Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*. Cambridge, MA: Cambridge University Press.
Williams, S. A., Terras, M. M. and Warwick, C. (2013) "What do people study when they study Twitter? Classifying Twitter related academic papers." *Journal of Documentation*, 69(3), 384–410.
Zamith, R. and Lewis, S. C. (2015) "Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis." *The ANNALS of the American Academy of Political and Social Science*, 659(1), 307–318.
Zimmer, M. and Proferes, N. J. (2014) "A topology of Twitter research: Disciplines, methods, and ethics." *Aslib Journal of Information Management*, 66(3), 250–261.