Potential and Limits of Automated Classification of Big Data: A Case Study

Weichbold, Martin; Seymer, Alexander; Aschauer, Wolfgang; Herdin, Thomas

Potential and Limits of Automated Classification of Big Data – A Case Study

Martin Weichbold, Alexander Seymer, Wolfgang Aschauer & Thomas Herdin *

Abstract: *»Potentiale und Grenzen der automatischen Klassifikation von Big Data – Eine Fallstudie«*. This case study highlights the potentials and limits of big-data analyses of media sources compared to conventional, quantitative content analysis. In an FFG-funded multidisciplinary project in Austria (based on the KIRAS security research program), the software tool WebLyzard was used for an automated analysis of online news and social media sources (comments on articles, Facebook postings, and Twitter statements) in order to analyze the media representation of pressing societal issues and citizens' perceptions of security. Frequency and sentiment analyses were carried out by two independent observers in parallel to the automated WebLyzard results. Specific articles on selected key topics like *technology* or *Muslims* in two major online newspapers in Austria (*Der Standard* and *Kronen Zeitung*) were counted, as were user comments, and both were evaluated according to different sentiment categories. The results indicate various weaknesses of the software leading to misinterpretations, and the automated analyses yield substantially different results compared to the sentiment analysis carried out by the two raters, especially for cynical or irrelevant statements. From a social-sciences methodological perspective, the results clearly show that methodology in our discipline should promote theory-based research, should counteract the attraction of superficial analyses of complex social issues, and should emphasize not only the potentials but also the dangers and risks associated with big data.

Keywords: Security perceptions, social media, big data, evaluation study, automated analysis.

* Martin Weichbold, Department of Political Science and Sociology, University of Salzburg, Rudolfskai 42, A-5020 Salzburg, Austria; martin.weichbold@sbg.ac.at.
Alexander Seymer, Department of Political Science and Sociology, University of Salzburg, Rudolfskai 42, A-5020 Salzburg, Austria; alexander.seymer@sbg.ac.at.
Wolfgang Aschauer, Department of Political Science and Sociology, University of Salzburg, Rudolfskai 42, A-5020 Salzburg, Austria; wolfgang.aschauer@sbg.ac.at.
Thomas Herdin, Department of Communication Studies, University of Salzburg, Rudolfskai 42, A-5020 Salzburg, Austria; thomas.herdin@sbg.ac.at.

Historical Social Research 45 (2020) 3, 288-313 │ published by GESIS
DOI: 10.12759/hsr.45.2020.3.288-313

1. Introduction

Public opinion has always been an issue, not only for social sciences but also for the state and the government as well as for businesses and the media. Public discourses shape and reflect the values and attitudes of citizens and affect the personal behavior of voters and consumers as well as how groups or parties behave collectively. A hundred years ago, Charles Horton Cooley characterized public opinion as a process of interaction and mutual influence rather than a state of agreement (Cooley 1918, 378). Influencing public debates and opinion is of great concern for all kinds of political actors, and for a long time classical mass media like newspapers, radio, or television were the arena for such activities. With the rise of social media, the focus of research has shifted towards new opportunities to influence public discourse, but this new media landscape has also thrown up new challenges. On the one hand, it is currently argued that the widespread use of the internet for social networking, blogging, video-sharing, and tweeting fosters participatory democracy. Politicians and media are no longer the only producers of political information; actors from civil society, e.g., citizens and NGOs, are also relevant actors in the field of political discourse. The idea of a potential improvement of democracy is based on the expectation that the increased diversity of digital networks and better accessibility to them will lower the barriers to engage in public discourse. But this optimistic view of a universally accessible, transparent discourse is challenged increasingly by the notion that this very accessibility of digital networks can also facilitate forms of use which threaten democracy (van Dijk and Hacker 2018). It can be clearly observed that governments as well as large private-sector corporations are increasingly using the internet and social media as means to maintaining ever-greater control over citizens and other stakeholders. This results in highly centralized networks in which content is disseminated through just a few influential hubs. A further critical perspective on the structural features of the web relates to the risk that the public sphere may disintegrate into fragmented publics that can no longer connect sufficiently with each other to form a shared world. Internet users are presented with content which is perfectly tailored and filtered to meet their personal preferences within their "filter bubbles" (Pariser 2011), reducing users' need for critical, intellectual reflection, while also taking less and less note of socially relevant discourses. Finally, there are also notable examples of political actors (in collaboration with private firms[1]) as well as of countries or their intelligence services trying to influence debates in other countries, especially during election campaigns.

[1] The most prominent example concerns Cambridge Analytica. Recent articles (e.g., Persily 2016) demonstrate that the firm was used in the run up to the Brexit referendum, and in even more sophisticated fashion during the US presidential elections. Cambridge Analytica

For the social sciences, but also for various kinds of actors (and as a prerequisite for influencing), it is important to know about the salience of various topics and discourses in the public. Research on media coverage and public opinion therefore has a long tradition. The crucial question is how to *evaluate* public discourse and how to *assess* public opinion. Content and discourse analysis, as elements of a qualitative methodology, are often used within communication and political sciences to identify which topics and argumentative strategies are prevalent or successful in public debates. By analyzing data like news coverage or stump speeches, these approaches have the advantage of being nonreactive, yielding results that are not biased by social desirability.

Research by survey is the classical quantitative approach in researching public opinion (Groves 2011; Bardes and Oldendick 2012): through the use of standardized questionnaires, values, attitudes, and opinions across a society can be measured, identifying socio-demographic, regional, or class-related differences and changes over time. Using representative samples, such research provides information about the distribution of views and beliefs in a country or even beyond. As well as national research, numerous international survey programs have emerged in recent decades, Eurobarometer (created in 1974) being the one with the longest series of data.

As social media now provides important platforms for public debates it is important to cover these new channels using new methodologies (see Lomborg 2017). Being part of the phenomenon of "big data," social media can be characterized by a number of "V"s, of which the three classical ones are *velocity*, *volume* and *variety* (see, e.g., Laney 2001; Ekbia et al. 2015; Mayerl and Zweig 2016):

- *Velocity* describes the enormous speed at which data is produced and dispersed. A president's tweet is sent and read around the world within seconds, triggering immediate responses.
- *Volume* refers to the huge amount of data produced by social media: while newspaper or TV reports are limited as they cannot exceed a certain length, social media can be used by everyone (not only journalists or public relations managers) and can host extensive messages.
- *Variety* stands for the multiple data formats being produced, including (user) statistics, text, and visual data.

Khan et al. (2014) list a total of seven terms beginning with V to characterize big data, of which *veracity* is probably the most relevant (Lukoianova and Rubin 2014). Veracity refers to the need for accuracy and trustworthiness of

targeted millions of voters, focusing predominantly on "hidden" Trump voters, who could not be addressed by opinion polls, and on Clinton voters in order to reduce voter turnout. Psychographic profiling methods (especially with regard to Facebook) attracted a lot of attention following these campaigns (Persily 2017, 65-66).

data, as social media data can be inconsistent, incomplete, ambiguous, or biased in some respect (Graeff and Baur 2020, in this issue).

These features are challenges for the analysis of public opinion when assessing sources of big data such as social media. While traditional methods require weeks or even months for data collection and analysis, a quick but valid method of exploration is necessary to handle the constant flow of data. An adequate method must be able to process large amounts of data, capable not only of counting specific occurrences (e.g., the number of messages from a particular sender or within a certain thread), but also of capturing the content of a debate and differentiating between arguments. Furthermore, the method should be able to integrate different forms and sources of data and – where veracity is concerned – to deal with different qualitative aspects of the data being analyzed (Weichbold 2009).

In recent years, much effort has been put into the development of tools for automated text analysis (see, e.g., Brier and Hopp 2011; González-Bailón and Paltoglou 2015; Boumans and Trilling 2016), including machine learning approaches (e.g., Samizade and Mahmoudi Saeid Abad 2018), but the question is whether the tools available are able to meet all these requirements (Trilling and Jonkman 2018). In what follows, we present the findings and conclusions of a research project in which we evaluated a tool for automated social media analysis, comparing its outcomes with a traditional content analysis.

2. Project Background

The Austrian Research Promotion Agency (FFG), the national funding agency for applied research and development, runs a program for security research issues known as KIRAS. In the context of this program, in 2015/16 a team of sociologists, IT experts, managers, and law consultants together with various Austrian ministries successfully developed an online tool called Foresight Cockpit. Using this platform, Austrian ministries were able to identify developments which might threaten public security by analyzing a broad range of statistical data provided by various authorities and organizations. They were also able to work collaboratively on future scenarios in order to take political or administrative measures. While our pilot study was on migration dynamics and integration challenges, any other topic could have been covered using the tool.

When evaluating Foresight Cockpit, it became clear that although a lot of statistical indicators on regional, national, and international levels had been integrated, it was not possible to look at very recent developments, as it takes time for relevant events (like refugee flows coming to a country) to show up in statistics. Furthermore, although our tool was able to cover official numbers, it could not handle actual sentiments or identify specific opinions expressed in public discussions. That these aspects are of great importance for the perception of security was seen recently: in 2015, a high number of refugees from Syria and Afghanistan came to Austria (and other European countries), where they met an open and welcoming attitude and were supported by many volunteers. After some time, however, and although the number of refugees was already decreasing, the public mood in significant parts of the Austrian population changed. In public debate, the voices against asylum seekers were clearly audible, and the political arguments especially in social media became more controversial and aggressive.

The idea was to complement Foresight Cockpit with an automated tool that enables its users to take media coverage and public sentiment into account, as it is not the absolute number of refugees that explains a country's perception of security and social cohesion, but how people and the public immediately react to certain refugee movements or integration challenges. Therefore, a follow-up project called Forestrat Cockpit was implemented, again funded by the FFG and in collaboration with ministries and researchers from different fields, in order to develop a tool for measuring media coverage and public discourses in different social media. Forestrat Cockpit ran from October 2016 to September 2017.

The project teams for both Foresight Cockpit and Forestrat Cockpit were multidisciplinary, with rather complex relations. Decisions in such teams do not necessarily follow scientific arguments. Rather, they were informed by a consensual aggregation of interests and the expertise and roles of the different project members: state authorities (the Ministries of Defense and of the Interior, as well as the Federal Chancellery) were involved to define the needs and requirements of the research project. Management consultants coordinated research with IT experts, and software engineers implemented the solutions, while law experts were engaged for legal consulting. As a team of sociologists and communication scientists, our role was to develop a theoretical model for security issues and perceptions, prepare its implementation in Forestrat Cockpit as a pilot study, and carry out an evaluation study of the software.

The aim of the project was to automatically capture the main content of discussions according to various thematic subfields in online newspapers and social media channels concerning security threats (or perceptions of them). Our assignment was: (1) to develop a theoretically driven model of (the feeling of) security which should serve as a basis for empirical investigation; (2) to implement the model for specific topics using WebLyzard (Scharl et al. 2013), a software for automated analysis of online media, by defining a list of keywords for retrieval and a selection of relevant social media sources; and (3) to evaluate the results of the automated analysis in terms of plausibility, reliability, and validity.

3. Overview of the Different Stages of the Project

3.1 Establishing a Theoretical Framework: Towards a Model of Security

Our model of highly relevant security issues and perceptions was constructed by referring to the literature that opts for a broad, multi-dimensional concept of security (e.g., Daase 2010). It goes beyond well-documented concepts like fear of becoming a victim of crime (Lüdemann 2006) or threats of social security (for instance unemployment; Hirtenlehner 2009). Objectively, these fears are largely unfounded in Austria, where there is a high level of social security and a low level of crime. In addition, quality of life is generally considered high, although anxiety about social status and pessimism about the future are reported in some surveys (Heitmeyer, 2010; Hadler and Klebel 2019). People see a danger of a societal decline – Bude (2014) refers to a *society of fear* – even if they evaluate their personal situation rather positively.

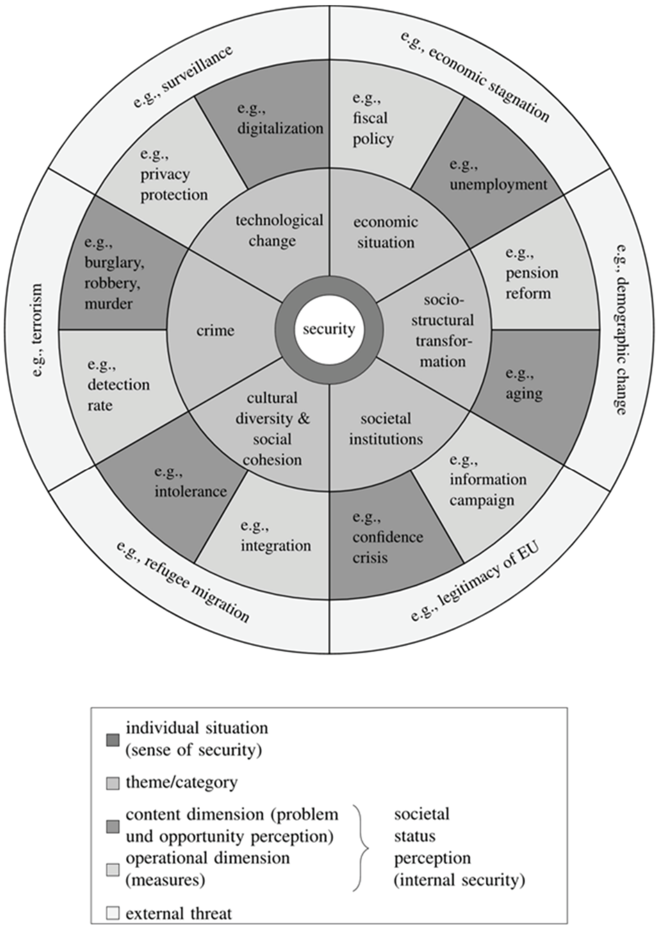
These gaps between objective measures (like crime statistics), and subjective perception, between evaluation of the general societal situation and personal circumstances, should be taken into consideration by the model and different media sources should be included. To conceptualize such a framework, it is necessary to identify central topics which might result in a security threat for the population as a whole or for specific groups. The analysis covers threats along six dimensions:

- *Cultural diversity and social cohesion* focuses on migration dynamics and integration, and their challenges for society. These topics are discussed extensively in public and seem to be a major source of uncertainty and fear.
- Connected with this is *crime*, which covers a broad range from personal risks (like sexual harassment of women by asylum seekers) to general, external threats (for instance fear of terrorism).
- Tensions in current European societies concern not only cultural values but also economic issues. Therefore, a separate *economic dimension* encompasses social inequality in Europe (e.g., Fredriksen 2012), the future of the middle classes (e.g., Burzan and Berger 2010), and increased precarity (Standing 2011).
- *Technological change* is related to economic challenges and addresses specific worries about development. This dimension covers new technological developments which have the potential to accelerate social change, such as the impact of digitalization on the working environment and social relations.
- *Socio-structural transformation* (caused by globalization, aging, or migration) is a challenge for political governance and may promote feelings

of anomy (e.g., Bohle et al. 1997). People may experience rapid social change merely as passive observers, unable to actively engage with the permanently changing environment.

- Consequently, *societal institutions* also fall into crisis. Disenchantment with politics is increasing (see, e.g., Huth 2004), and politicians are not expected to be able to solve the pressing issues. People are overstrained by the complexity of social dynamics and are looking for simple explanations and solution strategies. Out of this, even post-democratic developments may emerge in Western societies (Crouch 2008, Blühdorn 2013).

Figure 1: Theoretical Model of Security Perceptions of the Austrian Population



Source: Authors' own figure.

When analyzing security concerns, it is important go beyond a population's anxiety and negative feelings and to find a reasonable balance between challenges, options, and solution strategies. Therefore our theoretical model also integrates chances as well as measures for the selected topic areas. For instance, in the case of *cultural diversity and social cohesion* this means that the model includes *mutual intolerance* (as a perceived problem), *cross-cultural understanding and integration* (as a chance for the future), and *language* or *vocational training* (as concrete measures).

Needless to say, although the focus of the project is on Austria, national debates are linked to a supranational context. Thus it is necessary to include a spatial dimension and to assess major external threats which may have an impact on security perceptions in certain European countries.

3.2 Implementing the Model: Using WebLyzard

WebLyzard is a software tool based on a search engine for crawled data and provides descriptive tools to process, visualize, and analyze media sources.[2] These data consist of online content extracted according to a predefined set of search terms and retrieved from a project-specific list of sources (Scharl et al. 2013; Scharl et al. 2017).

The search engine is accessible through a dashboard providing advanced search specifications like common expressions or logical operators to narrow down the high volume of data. Thus, localization of relevant information is a multi-step process starting with more general search terms and isolating the subject by adding multiple layers of filters and additional search phrases. The dashboard provides graphical user interfaces to certain standard filters like a calendar, drop-down menus for the sources, and predefined searches covering the categories of the theoretical model. Visualization tools include trend charts, relation trackers, geographical maps, tag clouds, cluster maps, and network maps. Furthermore, the full text and content of the stored data can be accessed including a link to the original source.

Another major feature is the sentiment analysis. The extracted information is compared as full text against two dictionaries – one including positively connoted words, the other negative ones. This dictionary-based sentiment analysis addresses the syntactical basis in a very quick and extensive manner: words, word stem and even complex combinations or distances may be identified very effectively, but it falls short in "understanding" the content semantically. Each word has a weight based on the dictionaries aggregating to the standardized total sentiment score of each text. The sentiment analysis considers emphasizing phrases in the weighting process (Weichselbraun et al. 2013, 2014; Weich-

[2] Since the end of the project, WebLyzard has undergone further developments enhancing features and capabilities. The description here refers to the features in the Foresight Cockpit implemented at the time of our research.

selbraun et al. 2017). The numerical sentiment score can be visualized in a color gradient from red to green, with red representing a negative sentiment, green a positive one, and white a neutral one.

To fulfil our role within the project, we had to find terms to capture the intended content as comprehensively as possible. In a brainstorming workshop with researchers and students from different disciplines, terms for each field were defined, paying attention to a balance between positive, neutral, and negatively connoted words. In addition, various social media channels were explored for terms frequently used when expressing security concerns. Altogether, more than 1,000 words were found. In another workshop with the other partners involved in the project, the list was reduced to about 500. After thematic narrowing, and because of the limitations of the software, these were finally reduced to 241 terms. It is evident that the selection of the key words is a crucial step for the validity of the research, and the reduction of their number is likely to affect the value of the results.

Furthermore, it was clear that specific social media channels had to be selected, particularly because in this project the discourses had to be related to the situation in Austria. As a technical requirement, the discourses had, furthermore, to be public and accessible for scientific use; this is not the case for *WhatsApp*, for instance, which uses encryption, or for private communication on *Facebook*. In the end, two different kinds of sources were covered: as classical media still play an important role in public discussions, we decided to integrate reports published in selected Austrian newspapers and magazines. Unlike printed newspapers, the online versions allow readers to add direct comments on the reports. These comments were also captured. The second major source was social media channels: the application programming interfaces (APIs) of *Facebook* and *Twitter* were used to monitor tweets and posts on a number of accounts. These accounts included those of important actors in political discourses (e.g., politicians, journalists, news agencies, NGOs) and public agencies (e.g., the police or ministries); attention was payed to cover a broad range of opinions. Facebook limits the account-based extraction of information to 100 accounts, while Twitter implements no such limit. Additionally, Twitter provides direct access to the stream of tweets via the API, which could be used to extract relevant tweets based on the search terms. Altogether, five sources were analyzed: newspaper articles, comments on newspaper articles, Facebook accounts, Twitter accounts, and Twitter full-text search of all tweets.[3]

3.3 The Evaluation of the Automated Analysis

Screening these media sources for the keywords and classifying the hits for sentiment should – if the automated analysis works properly – give a picture of

[3] The data will be available at the Austrian Social Science Data Archive in the near future.

the public discourse on security issues in certain media sources in Austria. But is this picture really reliable and valid? It was tested by implementing a parallel manual analysis to obtain insights into the quality of the automated result. However, due to limited resources for the offline analysis, the comparison had to be restricted.[4]

To focus on specific topics, search terms were defined representing three thematic areas covering the content as well as the operational dimension (see Figure 1). The term *technology* covers the theme of technological change, especially digitalization, *job market* refers to the economic situation and fears of unemployment, while the subject *radicalization* is implemented via three terms (*radicalization*, *racism*, and *Muslims*). The decision to include three search terms related to radicalization resulted from the low frequency and the one-dimensionality of individual terms, which turned out to be unable to grasp the manifold forms of radicalization.

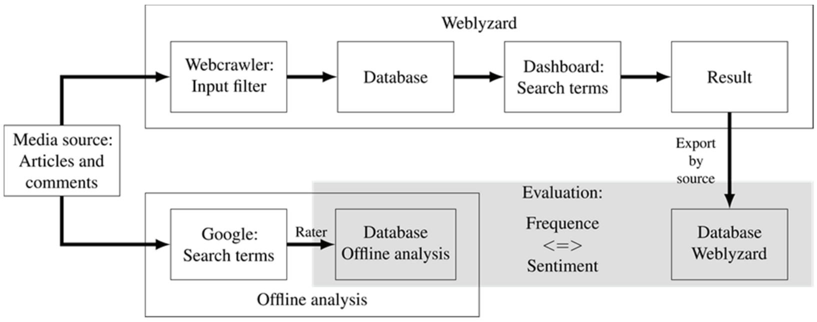
Both the online and the offline analyses were carried out for a two-week period from 29.05.2017 to 11.06.2017. Two newspapers were selected, representing different parts of the Austrian media landscape. *Der Standard* (*derstandard.at)* represents a quality newspaper; the *Kronen Zeitung* (*krone.at)* is the tabloid with the highest print run in Austria (Magin and Stark 2011, 107). The heterogeneity of the two newspapers is demonstrated, among other things, by the use of different styles of reporting, which could have an effect on automated coding. For WebLyzard, news media were crawled by full website mirroring and RSS-feeds to extract the relevant articles and related user comments posted within 24 hours of the publication of the article. For the manual analysis, relevant articles were searched using the search engines of the newspapers' websites, and those of both Google Alerts (www.google.at/alerts) and Google News (news.google.com). The articles as well as their respective comments were downloaded and stored in a database with title, URL, and date stamp. As for the automated analysis, comments posted within 24 hours of the article being published were included; the only limitation compared to the WebLyzard analysis was that for the manual retrieval only the first 200 comments were included, due to the time-consuming nature of the work.[5] Facebook and Twitter comments were not subject to manual retrieval. To control for reliability, manual analysis was carried out by two members of the research team independently, resulting in values (Cohen's kappa) of 0.87 (articles in krone.at), 0.92 (articles in derstandard.at); correlations between sentiment evaluations for the

4 While automated retrieval can process enormous volumes of text in a very short time, in the manual analysis each article and each comment had to be selected, read, and classified by two members of the research team independently.

5 We would like to thank our project assistants Lena Stöllinger and Patric Messer for the tremendous amount of work they carried out in evaluating more than 12,000 comments based on sentiment.

comments are above 0.7 across articles,[6] which can be interpreted as very reliable (see Greve and Ventura 1997, 111).

Figure 2: Diagram for the Database Creation in the Evaluation Study



Source: Authors' own figure.

3.3.1 The Comparison of Search Results

Using the same search terms for the same newspapers and the same time period, one might expect very similar results. In total, the manual search queries resulted in 180 articles in *Der Standard* and 55 articles in the *Kronen Zeitung*; after eliminating irrelevant and duplicate articles, 138 or 44 articles in the two newspapers respectively remained for further evaluation.

Within the defined period of 24 hours from the publication of an article, a total of 26,440 comments were identified for the 182 articles. As only the first 200 postings for each article were rated, 12,195 comments remained for further classification.

[6] Both raters used the same articles and corresponding comments, resulting in two sentiments per article and two aggregated counts of sentiments for the comments. Due to the huge number of comments, the comments were not matched by an ID across raters. Raters reported only a sum for each sentiment category per article. The correlation is based on these counts.

Table 1: Manually Identified Articles and their Respective Comments per Search Term, by Source

	Der Standard		Kronen Zeitung		Both	
	Articles	Comments on articles	Articles	Comments on articles	Articles	Comments on articles
Technology	48	1,772	13	17	61	1,789
Job market	27	2,472	3	281	30	2,753
Radicalization	5	349	4	337	9	686
Racism	21	1,412	3	0	24	1,412
Muslims	37	4,239	21	1,368	58	5,607
Total	138	10,244	44	2,003	182	12,247

Source: Authors' own table.

In a manner similar to the offline analysis, search queries were performed by WebLyzard on derstandard.at and krone.at. But whereas Weblyzard retrieved additional information from Facebook accounts and Twitter accounts and through a Twitter full text search separately, the comparison to the offline analysis is limited to the newspaper articles and comments. Facebook and Twitter results are reported to allow for a comparison of the sources within WebLyzard.

Table 2: Identified Articles, Comments and Messages in Twitter and Facebook Extracted from WebLyzard

	Twitter accounts	Facebook accounts	Twitter full text	Articles	Comments on articles
Technology	4	14	251	61	4
Job market	17	18	1,812	35	11
Radicalization	4	8	274	15	6
Racism	3	23	4,778	22	8
Muslims	17	80	16,267	42	38
Total (search result)	45	143	23,382	175	67
Total (period)	14,143	12,121	966,748	37,423 [a]	7.185 [b]
Search / Period	0.3 %	1.2 %	2.4 %	0.5 %	0.9 %

Notes: [a]2,021 articles were extracted from *Der Standard*, and 1,931 from *Kronen Zeitung*. [b]4,624 comments were extracted from *Der Standard*, and 986 from *Kronen Zeitung*.

The number of units identified by WebLyzard seems to be very different for the five media sources, which can be explained by technical reasons of data processing. Twitter can be accessed directly via an API: using the predefined search terms, data represent a direct search query in Twitter's database, which consists of nearly one million tweets. Within *Twitter accounts*, only tweets from 259 selected Twitter accounts (of news media, institutions, and people of public interest) are included. Therefore, the number of messages identified is much lower. For Facebook, the postings of 185 accounts are considered on condition that the owner's privacy settings allow access. Compared to Twitter, the data extraction from Facebook is thus clearly limited.

Newspaper articles are captured using two parallel procedures. On the one hand, there is a regular full mirroring of the news source's website and, on the other hand, the RSS feed is read several times a day. This procedure may cause problems, as some articles are counted twice, which is also significant where the following analyses of relevance and sentiment are concerned.[7] Comprising 37,423 documents from news media, of which about 4,000 are from derstandard.at and krone.at, the media landscape is generally well covered. Comments are selected based on the articles they belong to: since any comments are tapped along with the articles, the comments are extracted subsequently (Scharl and Weichselbraun 2008; Pollach, Scharl, and Weichselbraun 2009). Although a reasonable number of articles could be identified, this is not the case for associated comments: as the total number of comments is much lower than the number of articles, this can only be explained by technical problems of the software.

Another noteworthy point is the discrepancy between Twitter and Facebook accounts on the one hand, and Twitter full-text search results on the other. Since only a limited number of individual Twitter and Facebook accounts are read, the results are of limited validity if we are claiming them as evidence for "public debates," and conclusions based on them should be taken with a degree of caution.

Comparing the online and offline analyses, the overall number of newspaper articles is quite similar (182 compared to 175), although there are greater differences in the respective topics (see 3.3.2 below). The comparison of the comments extracted from the newspapers is hardly feasible, because the results from WebLyzard have only very limited validity as a comparison criterion due to their low frequency.

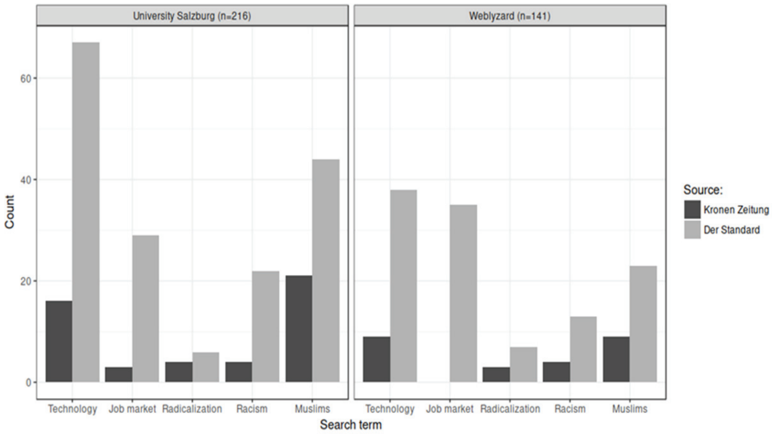
3.3.2 The Comparison of Frequencies

In the next step, a detailed comparative frequency analysis of the extracted articles and comments compared to the manual analysis was carried out. Although the overall number of articles is similar for the online and offline analyses, it is evident that the articles extracted were not the same ones. Searching for *Muslims* and *technology* in derstandard.at results in a much lower number in WebLyzard, and only half as many articles were found for *racism* compared to the manual retrieval. For krone.at, we found an astonishingly consistent number of articles at first sight, but it turned out that in the raw data of WebLyzard almost half of the 50 articles were duplicates. After deleting

7 As these doublets only occurred at krone.at, they are probably caused by a technical shortcoming.

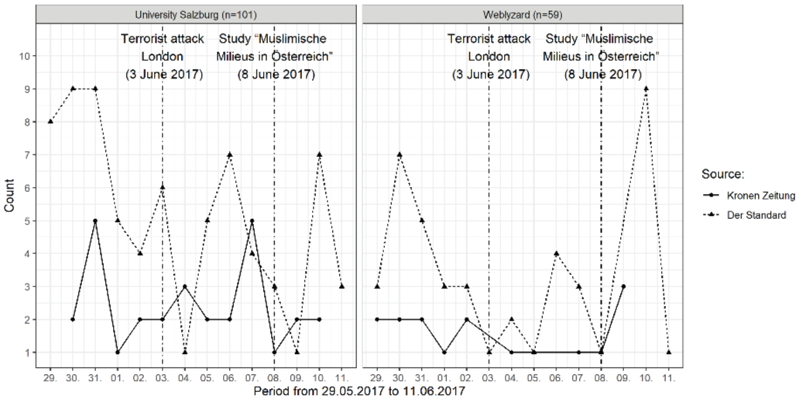
these duplicates, the numbers in the relevant categories were much lower. Figure 2 shows the number of articles for the three topics: according to WebLyzard, derstandard.at reported on all three subject areas equally often, but the offline analysis shows that the job market was covered only half as often as radicalization or technology. Differences can also be observed for krone.at.

Figure 3: Frequency of Articles by Subject Area, Source, and Analysis (Excluding Duplicates)



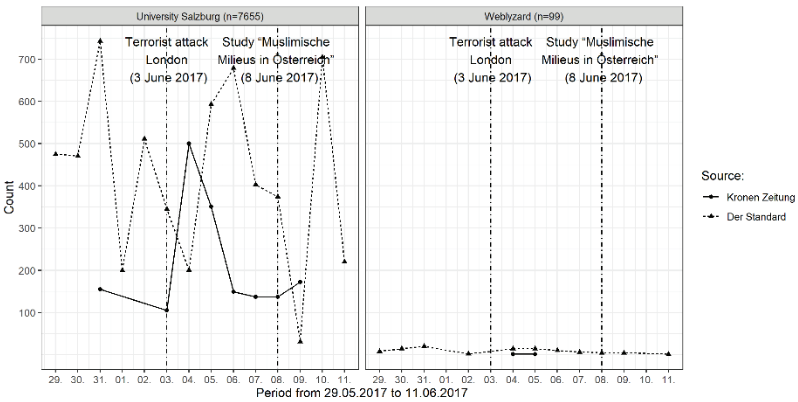
Reporting on radicalization can serve as a specific example for differences between on- and offline analyses, as two key events occurred during the investigation period, namely a terrorist attack in London 3 June 2017, and the presentation of a study on "Muslim milieus in Austria" on 8 June. Particularly striking is the lack of coverage of the terrorist attack in London in WebLyzard, although it is evident (and visible in the manual retrieval) that the attack committed by Islamic fundamentalists had been reported by the media. On the other hand, the presentation of the study generated intense discussion among readers of derstandard.at, with 1,180 comments posted within the first 24 hours. Comparing the frequency of the articles over time, the curve is similar for derstandard.at, with slight differences at the beginning and end of the study period. For krone.at, there are two distinct peaks in the offline analysis, which are not shown by WebLyzard.

Figure 4: Frequency of Articles over Time for the Topic of Radicalization



Overall, it can be stated that WebLyzard performs comparatively well in terms of the frequency of the articles, but differences across the subject areas are evident. For the comments, however, Weblyzard was unable to provide valuable conclusions.[8] While 26,440 comments were identified in the offline analysis, the WebLyzard database comprised only 67 comments for the same period (61 from derstandard.at, and 6 from krone.at), which makes a reliable temporal comparison of the frequency of the postings impossible.

Figure 5: Frequency of the Comments Over Time for the Topic of Radicalization

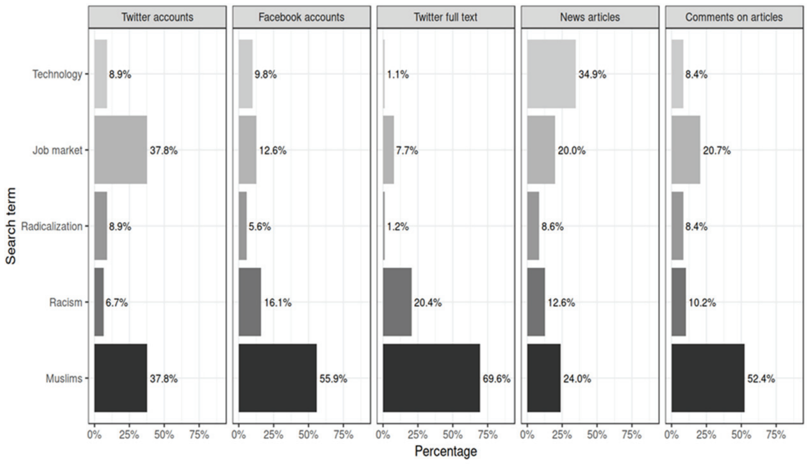


[8] We should state that we retrieved the information on 12 July and our project partner considered the extraction of comments as work in progress. The low frequency of comments may be due to mainly technical reasons. There is room for improvement.

In addition to the articles and the comments, WebLyzard comprises three further data sources that were not the subject of the offline analysis. Figure 5 shows the percentage share of search terms by data source. It should be noticed that the absolute number of messages is extremely different for the five sources, ranging from 45 in Twitter accounts to 23,382 for the Twitter full-text database (for details, see Table 2 above).

Comparing the media sources, one might get the impression that newspapers report technology extensively, and that issues such as Muslims and the job market receive considerably less coverage. However, one has to be mindful of whether such a frequency analysis is able to capture the salience of topics in the news media and in the public in a valid way: it is quite plausible that job market issues and integration challenges (among Muslims) generate more public debate than technological issues. The comparison of Twitter accounts and the Twitter full-text search makes clear that public debate does not necessarily reflect topics set by traditional opinion leaders (from politics and the media). The majority of the Twitter community seems to discuss Muslims and racism and, to a lesser extent, job market issues, while technology remains unimportant.

Figure 6: Relative Frequencies of the Search Terms by Data Source for WebLyzard



Source: Authors' own figure.

3.3.3 Comparison of Sentiment Ratings

It is important to be aware of which topics are being discussed in the media and by the public, but the frequency of issues is just one side of the coin: frequency does not necessarily tell us anything about what people think about security issues. To capture the societal climate with regard to crucial issues, it is necessary to analyze the content of relevant comments and statements, at least to see whether they follow a positive, negative, or neutral tendency. Table 3 shows the results of the sentiment ratings for the articles, divided according to subject area and media source. In the offline analysis, it turned out that three categories are not enough to capture the essential message of a text. Some texts turned out to be irrelevant for the topic although they comprised one of the keywords and were therefore selected. "Irrelevant" identifies articles which Weblyzard ideally filters after the crawling process and are therefore not supposed to be in the WebLyzard database. These articles thus have minor relevance for our analysis. "Ambiguous" is an indicator for minor difficulties in identifying the sentiment in articles. In some cases, the two reviewers came to a different evaluation – obviously it was not clear whether the general statement of a text was positive or negative. In these cases, we decided to classify them as ambiguous. Neither "irrelevant" nor "ambiguous" are categories provided in WebLyzard.

The comparison of the classifications of WebLyzard and our reviewers reveal that the automated evaluation rates the articles of derstandard.at much more positively. Meanwhile, for krone.at the difference between the automated sentiment analysis and the manual analysis by the raters is less clear.

Table 3: Sentiment Evaluations of Articles, by Source and Search Term

		Der Standard				*Kronen Zeitung*			
	Sentiment	University Salzburg		Weblyzard		University Salzburg		Weblyzard	
Technology	positive	7.5%	(5)	68.4%	(26)	12.5%	(2)	33.3%	(3)
	neutral	34.3%	(23)	5.3%	(2)	43.8%	(7)	22.2%	(2)
	negative	25.4%	(17)	26.3%	(10)	12.5%	(2)	44.4%	(4)
	ambiguous	4.5%	(3)	0.0%	(0)	12.5%	(2)	0.0%	(0)
	irrelevant	28.4%	(19)	0.0%	(0)	18.8%	(3)	0.0%	(0)
Job market	positive	10.3%	(3)	65.7%	(23)	33.3%	(1)	33.3%	(1)
	neutral	55.2%	(16)	20.0%	(7)	33.3%	(1)	33.3%	(1)
	negative	27.6%	(8)	14.3%	(5)	33.3%	(1)	33.3%	(1)
	ambiguous	0.0%	(0)	0.0%	(0)	0.0%	(0)	0.0%	(0)
	irrelevant	6.9%	(2)	0.0%	(0)	0.0%	(0)	0.0%	(0)
Radicalization	positive	0.0%	(0)	0.0%	(0)	0.0%	(0)	0.0%	(0)
	neutral	16.7%	(1)	28.6%	(2)	25.0%	(1)	12.5%	(1)
	negative	66.7%	(4)	71.4%	(5)	75.0%	(3)	87.5%	(7)
	ambiguous	0.0%	(0)	0.0%	(0)	0.0%	(0)	0.0%	(0)
	irrelevant	16.7%	(1)	0.0%	(0)	0.0%	(0)	0.0%	(0)
Racism	positive	4.5%	(1)	30.8%	(4)	0.0%	(0)	44.4%	(4)
	neutral	22.7%	(5)	30.8%	(4)	50.0%	(2)	0.0%	(0)
	negative	54.5%	(12)	38.5%	(5)	0.0%	(0)	55.6%	(5)

		Der Standard				*Kronen Zeitung*			
	Sentiment	University Salzburg		Weblyzard		University Salzburg		Weblyzard	
	ambiguous	13.6%	(3)	0.0%	(0)	25.0%	(1)	0.0%	(0)
	irrelevant	4.5%	(1)	0.0%	(0)	25.0%	(1)	0.0%	(0)
Muslims	positive	0.0%	(0)	17.4%	(4)	4.3%	(1)	15.8%	(3)
	neutral	29.2%	(14)	21.7%	(5)	17.4%	(4)	10.5%	(2)
	negative	45.8%	(22)	60.9%	(14)	65.2%	(15)	73.7%	(14)
	ambiguous	2.1%	(1)	0.0%	(0)	4.3%	(1)	0.0%	(0)
	irrelevant	22.9%	(11)	0.0%	(0)	8.7%	(2)	0.0%	(0)

Figure 7 shows the automated sentiment ratings by search term, for all data sources. For the most debated topic, "Muslims," the sentiment distribution seems fairly similar across most data sources, except for news articles being most negative.[9] For "racism" one can observe large differences between the news articles and the related comments on the one hand, and social media messages on the other. An explanation could be that news editors are very sensitive to racism and extreme comments are removed quickly, while in social media there are no such interventions – but that should be true also for the topic "Muslims." The results for "job market" seem plausible, as news articles reflect fears less than the social media. That the comments are very positive too, however, does not fit with the general picture. The sentiment shares for radicalization and technology cannot be interpreted meaningfully as both search terms have a low prevalence in the social media.

[9] For example, Hafez and Richter (2007) show by a media analysis of the image of Islam on two public-service German TV channels (ARD and ZDF) that even here 81% of all contributions on Muslims deal with negative aspects.

Figure 7: Sentiment Evaluation by Data Source and Search Term for WebLyzard

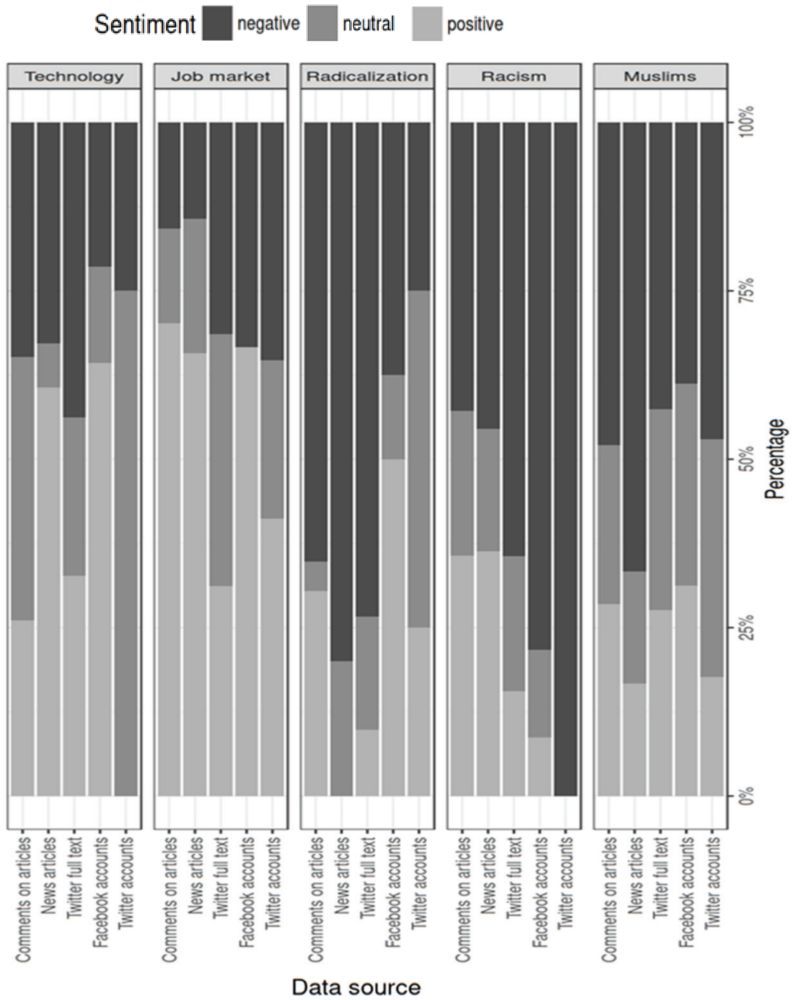


Table 4 reveals the main findings from the manual sentiment evaluation compared to WebLyzard. Of course, the results concerning *Kronen Zeitung* are not comparable, because only three comments were extracted by WebLyzard. The manual analysis is far more meaningful, therefore, leading to two significant results. Firstly, the data indicate the high prevalence of cynical statements, which cannot be recognized by the automated software. Secondly, negative comments are far more common in the tabloid than in the quality newspaper, *Der Standard*.

Table 4: Comparison of Sentiment Evaluation of Comments (Manual vs. Automated Analysis)

		Der Standard				Kronen Zeitung			
	Sentiment	University Salzburg		WebLyzard		University Salzburg		WebLyzard	
Technology	positive	5.6%	(100)	0.0%	(0)	0.0%	(0)	0.0%	(0)
	neutral	38.1%	(675)	50.0%	(1)	29.4%	(5)	50.0%	(1)
	negative	23.1%	(410)	50.0%	(1)	41.2%	(7)	50.0%	(1)
	cynical	20.9%	(371)	0.0%	(0)	29.4%	(5)	0.0%)	(0)
	irrelevant	12.2%	(216)	0.0%	(0)	0.0%	(0)	0.0%	(0)
Job market	positive	4.1%	(101)	80.0%	(8)	5.3%	(15)	100.0%	(1)
	neutral	35.8%	(884)	10.0%	(1)	12.5%	(35)	0.0%	(0)
	negative	34.2%	(844)	10.0%	(1)	52.3%	(147)	0.0%	(0)
	cynical	19.1%	(473)	0.0%	(0)	28.8%	(81)	0.0%	(0)
	irrelevant	6.8%	(168)	0.0%	(0)	1.1%	(3)	0.0%	(0)
Radicalization	positive	4.2%	(248)	30.6%	(15)	6.7%	(114)	0.0%	(0)
	neutral	36.5%	(2172)	18.4%	(9)	17.0%	(290)	0.0%	(0)
	negative	30.8%	(1831)	51.0%	(25)	55.3%	(944)	0.0%	(0)
	cynical	20.7%	(1234)	0.0%	(0)	19.8%	(337)	0.0%	(0)
	irrelevant	7.8%	(464)	0.0%	(0)	1.2%	(21)	0.0%	(0)

Note: Radicalization groups together results for all three search terms: Radicalization, racism and Muslims.

Comparative conclusions between the manual and the automated analyses can be drawn only by looking at the topic *radicalization* in *Der Standard*. Although we have only 49 comments extracted by WebLyzard, we can see that positive and negative results are disproportionately high compared to the manual ratings. Notably, more than 20% of the comments were rated as cynical and nearly 8% of the comments as irrelevant.[10] It is thus highly plausible that due to the difficulty in identifying cynical statements, the results of the automated software may lead to false negative or false positive values.

The comparison between offline analysis and WebLyzard shows that an automated classification of sentiment has clear limitations and the validity of the results is questionable. Besides the technical insufficiencies experienced within the project, sentiment analysis seems to be still in its infancy. Although there is doubtless scope for some technical improvements, it is debatable whether

[10] Thus calculating the proportion of ambiguous or irrelevant statements is crucial to ascertaining the number of postings which cannot easily be automatically judged by sentiment.

counting words with positive and negative connotation captures the meaning of a text. The Weblyzard technology failed in filtering out irrelevant statements and lacks the capabilities of identifying cynicism or sarcasm, which characterize a significant amount of text in social media or comments below news articles.

4. Conclusion: Some Methodological Reflections on Big Data Based on This Case Study

In the introduction, we identified velocity, volume, variety, and veracity as the main characteristics of big data analysis, characteristics which challenge social research in obtaining reliable and valid outcomes from the analysis of big data. With regard to our own results, at first sight one might argue that our attempt at an automated measurement of public opinion using a tool to analyze media coverage and social media reporting failed largely. However, such a conclusion would be too hasty and superficial. The problems we faced were diverse in nature. Some *technical deficiencies* should be easy to overcome. The low number of comments or doublets which were included in the automated analysis is an example of a technical problem which can easily be fixed. Other problems resulted from *actual restrictions and insufficient capacities*, and it should be possible to overcome these, too, in the future. More complex algorithms would allow for more reliable and valid classification; progress in this area will certainly go on. But there are also problems that demand our attention as critical methodologists. These mainly concern *semantics.* Although recent developments in language processing are impressive, computers will not be able to "understand" the meaning of a text in a semantic way. Indexicality (in semiotics), or the problem or recognizing irony or sarcasm can serve as examples.

Despite these limitations, the automated analysis of big data has an enormous potential. Referring to the four "V"s mentioned in the introduction, it is evident that automated classification is very fast (velocity). It allows for immediate conclusions, which is important because discourse in social media is fast-moving, and topics as well as sentiments can change within short periods of time. To capture these changes, it is necessary to use methods which are able to analyze this continuous flow of data.

The same potential is true regarding the mass of data (volume): traditional methods have to limit themselves to a small data sample due to restricted capabilities. Although the selection of data (in our case: social media sources) is also relevant here, the amount of data which can be processed is very much higher and will increase further with technological developments. Variety was less relevant in our study, as we used only text, although the texts were produced by different authors, including journalists (in the case of newspapers), leaders of public opinion (politicians, or agents of public institutions in the case of Facebook and Twitter accounts), or "regular" citizens commenting on newspaper articles. Nevertheless, this caused some problems on a technical level (notably the difficulties extracting comments).

One might control for the extent of media channels' influence by using indicators such as number of subscribers or readers, but dimensions like trustworthiness (veracity) are much more difficult to define and to measure. For our project, in order to account for the fact that the impact of the different data sources may vary, we carried out an expert rating for all media sources with respect to the extent of their influence, striving for seriousness and adequacy, and covering a broad range of opinions (see Aschauer et al. 2019). It is hard to predict whether future developments will allow for an automated evaluation, but at least there are efforts to detect bots in social media (Oentaryo et al. 2016), research which has the potential to influence discussions and sentiments in a relevant way.

The first three "V"s are related to formal and technical aspects of big data; the last one is related to its content and effect. This brings us back to the classical quality criteria of objectivity, reliability, and validity: above all other challenges and measures, these, too, must be met by automated analysis. At first sight, one might not consider there to be any problem with objectivity as analysis is performed "by a machine," without the subjective influence of those who execute the program. But this is not the whole picture. There is some subjective influence – the influence of those who wrote the program and coded the algorithms. For people who are not experts in coding, it is often hard to assess "what the computer does," how exactly a query or analysis is performed, and consequently what exactly a result means and what its limitations are. User-friendly operator interfaces and nicely edited results tempt the user to perform various kinds of analyses without knowing what is really happening in the background; reliability seems to be a minor problem as long as stable algorithms are used; applying the same procedures on the same data should reproduce identical results. But things change when it comes to machine learning algorithms, as procedures can be adapted with every run, making it harder and harder to reconstruct the process of analysis and to interpret the outcomes. What, at the end of the day, are we actually measuring, and have we included all relevant groups of society with our sources? The questions of validity and representativeness are central, but they often take a back seat due to nicely designed visualizations.

Big data provides an enormous potential for empirical social research which cannot be ignored, either as a huge reservoir of information about social phenomena, or as a social phenomenon itself. Tools provided for analysis are tempting, but they run the risk of leading one to work primarily exploratively and inductively, while theory-oriented, deductive social research falls behind. Users who do not really know what they are doing (or what is going on in the background) run a high risk of misinterpreting results and of drawing wrong conclusions. The automated analysis of large amounts of data carries the further risk that far-reaching measures will be delegated to Artificial Intelligence, resulting ultimately in loss of reflection and control. Established methodologies in social sciences should thus always be in the fore, in order to detect the strengths and weaknesses of current big data approaches.

References

Aschauer, Wolfgang, Alexander Seymer, and Martin Weichbold. 2019. Lässt sich das Sicherheitsgefühl der Bevölkerung automatisiert erfassen? Eine Fallstudie zur (künftigen) Rolle sozialwissenschaftlicher Methodologie im Zeitalter von Big Data. In *Digitalisierung und Gesellschaft. Sonderheft der Österreichische Zeitschrift für Soziologie*, ed. Christoph Musik and Alexander Bogner (forthcoming)

Bardes, Barbara A., and Robert W. Oldendick. 2012. *Public Opinion: Measuring the American Mind.* Lanham, MD: Rowman & Littlefield Publishers.

Blühdorn, Ingolfur. 2013. *Simulative Demokratie: Politik nach der postdemokratischen Wende.* Berlin: Suhrkamp.

Boccia Artieri, Giovanni. 2017. Social Media and the Challenge of Big Data/Deep Data Approach. In *Data Science and Social Research: Studies in Classification, Data Analysis, and Knowledge Organization*, ed. N. Carlo Lauro, Enrica Amaturo, Maria Gabriella Grassia, Biagio Aragona and Marina Marino, 57-66. Cham: Springer.

Bohle, Hans Hartwig, Wilhelm Heitmeyer, Wolfgang Kühnel, and Uwe Sander. 1997. Anomie in der modernen Gesellschaft: Bestandsaufnahme und Kritik eines klassischen Ansatzes soziologischer Analyse. In *Was treibt die Gesellschaft auseinander?,* ed. Wilhelm Heitmeyer, 29-68. Frankfurt am Main: Suhrkamp.

Boumans, Jelle W., and Damian Trilling. 2016. Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism* 4: 8-23.

Brier, Alan, and Bruno Hopp. 2011. Computer assisted text analysis in the social sciences. *Quality & Quantity* 45: 103-128.

Bude, Heinz. 2014. *Gesellschaft der Angst.* Hamburg: Hamburger Edition.

Burzan, Nicole, and Peter A. Berger, eds. 2010. *Dynamiken (in) der gesellschaftlichen Mitte*. Wiesbaden: Springer VS Verlag für Sozialwissenschaften.

Cooley, Charles H. 1918. *Social process*. New York: C. Scribner's Sons.

Crouch, Colin. 2008. *Postdemokratie*. Frankfurt am Main: Suhrkamp.

Daase, Christopher. 2010. National, social and human security: on the transformation of political language. *Historical Social Research* 35 (4): 22-37. doi: 10.12759/hsr.35.2010.4.22-37.

Ekbia, Hamid, Michael Mattioli, Inna, Kouper, G. Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R. Sugimoto. 2015. Big Data, Bigger Dilemmas: A Critical Review. *Journal of the Association for Information Science and Technology* 66: 1523-1545.

Fredriksen, Kaja Bonesmo. 2012. Income Inequality in the European Union. *OECD Economics Department Working Papers* 952. doi: 10.1787/5k9bdt47q5zt-en.

González-Bailón, Sandra, and Georgios Paltoglou. 2015. Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science* 659: 95-107.
Groves, Robert M. 2011. Three Eras of Survey Research. *Public Opinion Quarterly* 75: 861-871.
Graeff, Peter, and Nina Baur. 2020. Digital Data, Administrative Data, and Survey Compared: Updating the Classical Toolbox for Assessing Data Quality of Big Data, Exemplified by the Generation of Corruption Data. *Historical Social Research* 45 (3): 244-269. doi: 10.12759/hsr.45.2020.3.244-269.
Hadler, Markus, and Thomas Klebel. 2019. Einkommensungleichheit, Lebensstandard und soziale Position im Zeitvergleich. In *Sozialstruktur und Wertewandel in Österreich: Trends 1986-2016*, ed. Johann Bacher, Alfred Grausgruber, Max Haller, Franz Höllinger, Dimitri Prandner and Roland Verwiebe, 115-130. Wiesbaden: Springer VS Verlag für Sozialwissenschaften.
Hafez, Karola and Cay Richter. 2007. Das Islambild von ARD und ZDF. In *Aus Politik und Zeitgeschichte, Beilage zur Wochenzeitung Das Parlament*, 26-27: 40-46.
Heitmeyer, Wilhelm. 2010. Disparate Entwicklungen in Krisenzeiten, Entsolidarisierung und Gruppenbezogene Menschenfeindlichkeit. In *Deutsche Zustände: Folge 9*, ed. Wilhelm Heitmeyer, 13-33. Frankfurt: Suhrkamp.
Hirtenlehner, Helmut. 2009. Kriminalitätsangst – klar abgrenzbare Furcht vor Straftaten oder Projektionsfläche sozialer Unsicherheitslagen? *Journal für Rechtspolitik* 17: 13-22.
Humphreys, Ashlee, and Rebecca Jen-Hui Wang. 2017. Automated Text Analysis for Consumer Research. *Journal of Consumer Research* 44: 1274-1306.
Huth, Iris. 2004. *Politische Verdrossenheit: Erscheinungsformen und Ursachen als Herausforderungen für das politische System und die politische Kultur der Bundesrepublik Deutschland im 21. Jahrhundert.* Münster: LIT.
Khan, M. Ali-ud-din, Muhammad Fahim Uddin and Navarun Gupta. 2014. Seven V's of Big Data understanding Big Data to extract value. *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*. doi: 10.1109/ASEEZone1.2014.6820689.
Laney, Doug. 2001. 3D data management: Controlling data volume, velocity and variety. META Group Inc. 949. <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (Accessed January 07, 2019).
Lomborg, Stine. 2016. A state of flux: Histories of social media research. *European Journal of Communication 32:* 6-15.
Lomborg, Stine. 2017. A state of flux: Histories of social media research. *European Journal of Communication* 32 (1): 6-15. doi: 10.1177/0267323116682807.
Lüdemann, Christian. 2006. Kriminalitätsfurcht im urbanen Raum: Eine Mehrebenenanalyse zu individuellen und sozialräumlichen Determinanten verschiedener Dimensionen von Kriminalitätsfurcht. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 58: 285-306.
Lukoianova, Tatiana, and Victoria L. Rubin. 2014. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Advances in Classification Research Online* 24 (1). doi: 10.7152/acro.v24i1.14671.

Magin, Melanie and Birgit Stark. 2011. Österreich – Land ohne Leuchttürme? Qualitätszeitungen im Spannungsfeld zwischen publizistischer Leistung und strukturellen Zwängen. In: Blum, Roger, Bonfadelli, Heinz, Imhof, Kurt and Otfried Jarren. *Krise der Leuchttürme öffentlicher Kommunikation*; 97-114. Heidelberg: VS Verlag für Sozialwissenschaften.
Mayerl, Jochen, and Katharina Anna Zweig. 2016. Digitale Gesellschaft und Big Data: Thesen zur Zukunft der Soziologie. In *Big Data als Theorieersatz*, ed. Gregor Ritschel and Thomas Müller, 77-83. Berlin: Berliner Debatte Initial.
Oentaryo, Richard J., Arinto Murdopo, Philips K. Prasetyo, and Ee Peng Lim. 2016. On profiling bots in social media. *Social informatics: 8th International Conference*: 92-109.
Pariser, Eli. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. London, New York: Penguin.
Persily, Nathaniel. 2017. The 2016 US Election: Can Democracy Survive the Internet? *Journal of Democracy* 28 (2): 63-77.
Pollach, Irene, Arno Scharl, and Albert Weichselbraun. 2009. Web Content Mining for Comparing Corporate and Third-Party Online Reporting: A Case Study on Solid Waste Management. *Business Strategy and the Environment* 18: 137-148.
Ruoto, Antonio, Vito Santarcangelo, Davide Liga, Giuseppe Oddo, Massimiliano Giacalone, and Eugenio Iorio. 2017. The Sentiment of the Infosphere: A Sentiment Analysis Approach for the Big Conversation on the Net. In *Data Science and Social Research: Studies in Classification, Data Analysis, and Knowledge Organization*, ed. N. Carlo Lauro, Enrica Amaturo, Maria Gabriella Grassia, Biagio Aragona and Marina Marino, 215-222. Cham: Springer.
Samizade, Reza, and. 2018. The Application of Machine Learning Algorithms for Text Mining based on Sentiment Analysis Approach. *Journal of Information Technology Management* 10: 309-330.
Scharl, Arno, David Herring, Walter Rafelsberger, Alexander Hubmann-Haidvogel, Ruslan Kamolov, Daniel Fischl, Michael Föls, and Albert Weichselbraun. 2017. Semantic Systems and Visual Tools to Support Environmental Communication. *IEEE Systems Journal* 11: 762-771.
Scharl, Arno, Alexander Hubmann-Haidvogel, Marta Sabou, Albert Weichselbraun, and Heinz-Peter Lang. 2013. From Web Intelligence to Knowledge Co-Creation – A Platform to Analyze and Support Stakeholder Communication. *IEEE Computer Society* 17: 21-29.
Scharl, Arno, and Albert Weichselbraun. 2008. An Automated Approach to Investigating the Online Media Coverage of US Presidential Elections. *Journal of Information Technology & Politics* 5: 121-132.
Standing, Guy. 2011. *The Precariat: The new dangerous class*. London, New York: Bloomsbury.
Trilling, Damian, and Jeroen G. F. Jonkman. 2018. Scaling up Content Analysis. *Communication Methods and Measures* 12: 158-174.
Van Dijk, Jan A.G.M, and Kenneth L. Hacker. 2018. *Internet and Democracy in the Network Society*. London: Routledge
Vijver, Fons van de, and Norbert K. Tanzer. 2004. Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology* 54: 119-135.

Weichbold, Martin. 2009. Zur Bestimmung und Sicherung der „Qualität“ von Umfragen. In *Herausforderungen und Grenzen der Umfrageforschung*, ed. Martin Weichbold, Johann Bacher and Christof Wolf, 553-570. Wiesbaden: VS Verlag für Sozialwissenschaften.

Weichselbraun, Albert, Stefan Gindl, and Arno Scharl. 2013. Extracting and Grounding Contextualized Sentiment Lexicons. *IEEE Intelligent Systems* 28: 39-46.

Weichselbraun, Albert, Stefan Gindl, and Arno Scharl. 2014. Enriching Semantic Knowledge Bases for Opinion Mining in Big Data Applications. *Knowledge-Based Systems* 69: 78-85.

Weichselbraun, Albert, Stefan Gindl, Fabian Fischer, Svitlana Vakulenko, and Arno Scharl. 2017. Aspect-Based Extraction and Analysis of Affective Knowledge from Social Media Streams. *IEEE Intelligent Systems* 32: 80-88.

Wiedemann, Gregor. 2013. Opening up to big data: computer-assisted analysis of textual data in social sciences. *Historical Social Research* 38: 332-358. doi: 10.12759/hsr.38.2013.4.332-358.

Witzel, Andreas, Irena Medjedovic, and Susanne Kretzer. 2008. Sekundäranalyse qualitativer Daten: zum gegenwärtigen Stand einer neuen Forschungsstrategie. *Historical Social Research* 33: 10-32. doi: 10.12759/hsr.33.2008.3.10-32.

Historical Social Research
Historische Sozialforschung

All articles published in this Forum: