International Journal of Communication 20(2026), 456–480 1932–8036/20260005

Cultural Motifs of Big Data in User-Generated Content: A Semiautomated Analysis of 10 Years of Discourse

CHARLOTTE KNORR[1,2]
Leipzig University, Germany
LMU Munich, Germany

ANDREAS NIEKLER
CHRISTIAN PENTZOLD
Leipzig University, Germany

This article examines the sensemaking around big data in user-generated content on Reddit, Facebook, and Twitter/X. Big data is not only a technology but also an issue of public concern. However, given that the term "big data" has been around for more than a decade, little is known about how the cultural motifs used to make sense of it have changed over time or how public discussions reverberate with—or dispute—elite framings in news and high-profile publications. To fill this gap, we use a semiautomated content analysis of discourse around big data over a period of 10 years. We focus on cultural motifs that anchor big data frames in sensemaking. Our analysis, which integrates manual annotation with automatic classification using a transformer-based language model, revealed three predominant cultural motifs of seven examined. These recurring motifs primarily emphasize themes of profit and prediction, the evolving datafication of society, or the pursuit of innovations for societal advancement. The motifs encapsulate a range of topics that point to a diversity of adjoining discourses on datafication writ large.

Keywords: big data, datafication, frames, cultural motifs, classification models, Bidirectional Encoder Representations from Transformers (BERT)

Charlotte Knorr: charlotte.knorr@ifkw.lmu.de
Andreas Niekler: aniekler@informatik.uni-leipzig.de
Christian Pentzold: christian.pentzold@uni-leipzig.de
Date submitted: 2024-09-26

[1] Acknowledgments: We would like to extend our gratitude to our student assistants for their support during data analysis.
[2] The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 447465824/PE2436/3-1.

During the past decade, big data and ensuing processes of datafication and automation have generated significant attention, and they still do. The term "big data" itself has inspired a wealth of debate around the power and potential of large quantities of data and analytical procedures that are believed to reshuffle all kinds of sectors, from business and government, education and research to war, health, and personal relationships. The aspirations and fears associated with the notion are important for shaping the public opinion on big data, which again intertwines with regulation, funding, and investments (Jasanoff, 2015; Knorr & Pentzold, 2025). Discourses, in that sense, are vital to assess the opportunities and risks of an innovation (Scheufele & Lewenstein, 2005); ponder contextual influences (Scannell & Gifford, 2013); reflect on scientific, technical, or economic progress (Nisbet et al., 2002); gauge uncertainties (Guenther, Bischoff, Löwe, Marzinkowski, & Voigt, 2019); or spell out political interests (Bolsen, Druckman, & Cook, 2014).

In this capacity, the term "big data" denotes a technology—large troves of digital data and advanced behavioral or predictive analytics—as much as a social issue. Given the stakes involved, there is no shared understanding of what big data entails. Rather, it seems reasonable to assume many, even conflicting perspectives on big data, yet although the term has been around for more than a decade, there is a dearth of knowledge about how the motifs used to make sense of the notion have changed over time. Little is known, too, about how big data has featured in user-generated content (UGC) that could equally reverberate or dispute the views established in elite forums of news and high-profile publications.

This article reconstructs the cultural motifs in the discourse around big data in UGC on Reddit, Facebook, and Twitter/X. More precisely, we research how the expert discourse that gave structure to the mutable and multifaceted conversation is reflected in UGC. To capture the horizon of sensemaking that is mobilized to understand big data, we use a constructivist approach introduced by Gamson and Modigliani (1989) and further developed by Van Gorp (2010). This enables us to analyze the cultural motifs that embed big data in broader frameworks of meaning and cultural norms. These cultural motifs undergird frames and encapsulate a potentially larger number of topics (Knorr, Niekler, Behret, & Pentzold, 2023; Knorr & Pentzold, 2025). We ask:

RQ1: Which cultural motifs associated with big data from expert publications are reflected in user-generated communication?

RQ2: How do these cultural motifs evolve over time?

RQ3: What topics are captured by the cultural motifs?

Following these questions, we contribute to a growing strand of scholarship that places technological innovations in a context of discourse and cultural imaginaries (Jasanoff, 2015; Mager & Katzenbach, 2021; Nye, 1996; Streeter, 2011). It assumes the public understanding of emerging technologies, which includes the analysis of aggregated data sets, prefigures their sociocultural perception and evaluation as well as political decision making (Bolsen et al., 2014; Markham & Tiidenberg, 2020). Rather than putting technological innovations first and cultural appropriation second, we investigate the vital entanglement of sensemaking and technological development during 10 years of discourse. Despite

https://doi.org/10.65476/xgpt3649

the crucial role big data technologies and discourses have played in paving the way toward today's pervasive datafication, empirical research is still in its infancy.

Background: Framing Big Data in User-Generated Content

As a keyword, "big data" is surrounded by ideas of large amounts of information harboring unprecedented insights for those in the position to handle enormous quantities of digital data. This prospect is highly ambivalent: What seems beneficial and revolutionary for some is daunting and a threat for others. Within these quite opposite expectations, big data became the focus of public commentators and evangelists who all contributed to publicly making sense of big data technology, analysis, and mythology (boyd & Crawford, 2012; Pentzold & Knorr, 2023). They thread together various references that may, or may not at all, hinge on a shared understanding of the keyword "big data" and the notion of it.

Big Data Discourse in Constructionist Perspective

In terms of big data, the past two decades have seen a twin development: On the one hand, platforms like Meta or Twitter/X facilitate datafication by generating huge amounts of data and submitting them to increasingly sophisticated analytical models. On the other hand, these platforms have become forums where users discuss their thoughts about increasing social datafication (Paganoni, 2019; Rieder, 2018). Speaking of big data discourse can mean two things: discourses found on and representing big data and discourses reflecting on the concept of big data itself. Stretching the boundaries of big data practically, and questioning its reach and ramifications discursively, does not happen in a vacuum, but is embedded in cultural norms, belief systems, and values that assess what big data means for society and which challenges and risks it involves (Couldry & Yu, 2018; Knorr et al., 2023; van Dijck, 2014; Wyatt, 2021). Its semantics surface, for instance, in metaphors that surround the notion of big data, inviting us to view it as a flood or the new oil (Beer, 2018; Nolin, 2019; Portmess & Tower, 2015; Puschmann & Burgess, 2014).

Following Kitchin and McArdle's (2016) definition, the term "big data" denotes processes of data harvesting and analysis in varied data environments. Its ideology of dataism has been described as the "widespread belief in the objective quantification and potential tracking of all kinds of human behavior and sociality" (van Dijck, 2014, p. 198). While intricately interwoven with daily practices (Burgess, Albury, McCosker, & Wilken, 2022), big data has from the start been a matter of cultural curiosity and critical inquiry. It has been spearheaded by tech evangelists adumbrating a new age of data-driven insights, as well as business gurus and campaign managers heralding predictive analytics (Lohr, 2016; Mayer-Schönberger & Cukier, 2013). In turn, data scandals and whistleblowers raised awareness of dataveillance.

To better understand big data's connotations and capture the zeitgeist behind its frames, we adopt a culturalist approach. Following Van Gorp (2007), we note that cultural motifs are the cultural theme embedded in a text or post. Moreover, a frame anchors in a cultural motif as a leitmotif (Gamson & Modigliani, 1989). Individual big data frames that are the focus of other studies (Paganoni, 2019; Pentzold & Fischer, 2017) ground in cultural motifs that meaningfully connect their elements: the problem, causes, consequences, moral values involved, and possible solutions (Entman, 1993). The motifs condensing the cultural sensemaking around the moniker "big data" in user-generated communication between 2011 and

https://doi.org/10.65476/xgpt3649

2020 form the basis for framing processes. As a leitmotif, they encapsulate notions of what big data is and what should be done about it. Often normatively toned, they help us "better understand positive and negative portrayals of these technologies" (Cools, Van Gorp, & Opgenhaffen, 2024, p. 4). Accordingly, in this study, we look at cultural motifs as an element of user-generated discourse in mass-self communication that embeds big data within broader norms and systems of meaning.

Cultural motifs reflect cultural themes and are connected to social values and discourse patterns. They may encapsulate different topics, as they relate to "a set of discourses that interact in complex ways" (Gamson & Modigliani, 1989, p. 2). They can remain implicit because they belong to cultural beliefs and ways of thinking that are often taken for granted (Ryan & Gamson, 2006), such as metaphors, catchphrases, and stereotypes (Van Gorp & Vercruysse, 2012, p. 1275). Nevertheless, they are used to "define an issue" (Van Gorp, 2010, p. 92), limiting interpretations by emphasizing some aspects of a topic and ignoring others (Entman, 1993). Importantly, cultural motifs take shape as linguistic patterns extracted from statements, making them suitable for the study of large amounts of discursive material (Knorr et al., 2023; Pentzold & Fraas, 2023).

Cultural Motifs of Big Data

We suppose that the different frames articulated in user-generated communication ground in cultural motifs that are of wider significance and usually exhibit some cultural inertia that affords their long-term tracking. What is not known is the number and character of these cultural motifs and the topics they may encapsulate. Some indication about their thematic orientation is given by existing studies, though (Beer, 2018; Nolin, 2019; Portmess & Tower, 2015; Puschmann & Burgess, 2014). These have stressed the prominence of metaphors from the area of natural resources applied to data that are ready to be "harvested" or "mined." This suggests big data is not generated but found so that they can be appropriated at will. Another salient reference was to water and "floods," thus alluding to an overwhelming abundance that needs to be mastered.

The cultural motifs given shape by these metaphors and others can be traced in the available data. To prepare and inform our large-scale analysis, we inductively developed a set of cultural motifs found in 17 bestselling books using "big data" in their titles that catalyzed the expert discourse (Pentzold & Knorr, 2023). Our sampling relied on book review sections in newspapers and on bestseller lists, and we sourced the social cataloging sites Goodreads, LibraryThing, and StoryGraph. The aim was to survey contributions that were giving meaning to public sensemaking around big data. The books came out between 2013 and 2017.

Our preliminary analysis of the manually extracted text examples yielded seven cultural motifs (Table 1). This work happened in a team of two with regular team meetings to discuss possible candidates for motifs until all semantic aspects found in the books were satisfactorily captured. According to Van Gorp (2010), such a stepwise process of interpretation is necessary to minimize subjective readings. The subsequent analysis employs these established motifs and their contextual data as a heuristic framework to detect their occurrence within UGC.

https://doi.org/10.65476/xgpt3649

Table 1. Cultural Motifs.

Cultural motif	Description
1. Innovations for societal progress	Big data is key for an efficient and strong society. There is an underlying assumption of progress because of technological innovations in all social sectors, from politics and journalism to economy, education, and health.
2. Shift in datafying society	Big data is revolutionary. Big data is associated with a historical turning point, notions of change. A new society is to be built and maintained with the help of big data, where people are divided into groups or patterns. Data are facts and replace gut feeling.
3. Preventing wrongs	Big data can prevent crimes and wars and predict risks. The deployment of technologies improves police operations and renders policy actions more effective.
4. Low-profile surveillance	Big data is making surveillance omnipresent. Data are adopted by political, military, and corporate/economic oversight and intelligence. Addressing the legitimacy of it is fundamental because it may threaten democracy.
5. Profits and prediction	Big data is an economic driver. The value of the data is to be exploited and sold. It affords targeting groups in commercial and election campaigns.
6. Civic agency	Big data is a threat to privacy that requires counteraction that can be guided by a transparent data policy and ethics. This motif works on a deeper normative level. Core values are privacy and the protection of private data with an emphasis on collective empowerment and responsibility (bottom-up and grassroots NGOs).
7. Negative consequences	Big data is debated in its possible negative consequences and encourages individual empowerment and personal initiative, suggesting that individuals play a crucial role in shaping data practices and policies. Key values are privacy and the protection of private data plus people's own initiative and empowerment to act both together as publics and on a microlevel.

The seven motifs we found in our preliminary analysis are innovations for societal progress, shift in datafying society, preventing wrongs, low-profile surveillance, profits and prediction, civic agency, and negative consequences. Each of them refers to a distinct cultural motif associated with big data that is not subject to sudden changes but evolves in the long run (Van Gorp, 2010). That way, it enables a diachronic analysis of discursive evolution.

The cultural motifs "innovations for societal progress" and "profits and prediction" are affirmative in character and suggest exploiting big data for moving society forward or creating economic value. The motif "a shift in surveying society" also promises an efficient and strong society that hinges on observation. Instead, the motifs "low-profile surveillance" and "civic agency" refer to norms and values that become a matter of renegotiation, threatening democracy and people's privacy. The motif "preventing wrongs" legitimizes preemptive politics to protect people by preventing crimes and wars. Some cultural motifs allow the discussion of big data for political legitimization ("preventing wrongs," "low-profile surveillance"), while others afford discussion of societal changes coming with new technologies ("shift in surveying society," "civic agency," "negative consequences"). Furthermore, some are related to economic value creation processes

https://doi.org/10.65476/xgpt3649

promoted by big data ("innovations for societal progress," "profits and prediction"). As a set of initial cultural motifs, they provided the base for the topic modeling.

Data and Methods: Big Data on Reddit, Facebook, and Twitter/X

In our study, we approached the analysis from a constructionist perspective using classification models. The goal of it was to examine if the cultural motifs around big data from expert discourse are reflected in UGC. Moreover, next to the elementary register of cultural motifs, we were also able to consider appendant topics whose connection to big data may not have been straightforward, but that nevertheless formed part of the wider discourse on data technology and data-driven analytics.

From a methodological point, manual content analyses are, on the one hand, too cumbersome to analyze large data corpora. Even if it would be possible, the question remains how subjectivity in frame identification can be reduced (Van Gorp, 2007). On the other hand, there is also controversy about the extent to which computer-based approaches are capable of mapping frames accurately (Eisele, Heidenreich, Litvyak, & Boomgaarden, 2023). Therefore, one aspect of the study was to develop and test procedures for large textual data from different platforms over a longer period of time and at the same time explore how the keyword "big data" can be investigated with a combination of approaches from constructionist analysis, semiautomated classification, and automated topic modeling.

We carried out the data analysis in a two-step procedure: First, we collected the data with the help of different data providers. Second, we classified the material for cultural motifs and used topic modeling for the set of documents associated with a cultural motif. Hereby, we developed a mixed-method design of manual annotation plus transformer-based language model classification and topic modeling to analyze the data and reflect the platform-specific prevalence of motifs. The categories used for the classifier were based on the inductively developed cultural motifs (Table 1) that served as a starting point for the semiautomated content analysis using an active-learning approach. We traced the keyword "big data" and derivations on a long-term scale of 10 years (2011–2020) and on three platforms: Twitter/X, Facebook Sites and Facebook Pages, and Reddit. In the years after 2010, Twitter/X was intricately linked to news making, promotion, and political communication. Facebook was the platform with the largest and broadest user base, while Reddit attracted more niche groups, some of them with a stronger IT or countercultural orientation.

Our classification approach utilizes an active-learning framework to augment a small, manually annotated corpus of book passages with a large volume of unlabeled UGC. This methodology trains a robust classifier by strategically selecting the most informative unlabeled data for manual annotation, thereby overcoming the limitations of a sparsely labeled initial data set.

Collecting Texts From Three Platforms

The acquisition of data from the platforms Twitter/X, Facebook, and Reddit was done in a comparable way in accordance with the specifics from each platform. Data acquisition for this study was conducted systematically to ensure a representative and relevant selection of data in English and German. The process involved several steps. We did not capture the entire 10-year time frame but chose to focus

https://doi.org/10.65476/xgpt3649

on phases of intensive discursive activity. For each platform, the two phases per year with the highest occurrence of the relevant keywords were identified. For the analysis, a time frame of −5 to +15 days around these peaks was selected to capture the heightened discursive activity surrounding them. This is in line with evidence from online issue attention cycle studies (David, Ong, & Legara, 2016; Jünger & Gärtner, 2021).

On Twitter/X, data collection was performed using the R package academictwitteR (Barrie & Ho, 2021) with the keywords "big AND data," "big data," and "#bigdata" to identify relevant tweets. The total sample amounted to *N* = 2,052,388 posts. On Reddit, data were collected using Facebook's now-defunct CrowdTangle, with the keywords "#bigdata" and "big data" (*N* = 3,186 threads). Since CrowdTangle could not extract Reddit comments, Python and the PRAW library were used to extract and analyze posts and their comments. Posts and comments were scraped between 2013 and 2020, as earlier data were unavailable. On Facebook, data collection was also conducted through CrowdTangle, examining Facebook Pages (*N* = 44,410 posts) and Facebook Groups (*N* = 33,691 posts) separately, also with respect to their peaks per year. The same keywords ("#bigdata" and "big data") were used. This systematic approach facilitated a comprehensive and well-founded data collection, forming a solid basis for further analysis by integrating multiple platforms and considering the most relevant periods. The raw data were cleaned using a consistent process across all platforms to ensure uniformity and comparability. The data sets were sampled in subsets (one per platform) with R or Python and saved in CSV format. Each post and thread was given a time tag and date tag, but not a location tag. The classification data are provided as CSV on Open Science Framework.[3]

The same cleaning method applied to Twitter/X was adapted to all data sources. First, all posts containing "RT" were removed, as retweets were captured despite being excluded from the search query. Next, links within the tweets were taken out, but only the links themselves, not the entire tweets. Then all duplicates were eliminated, particularly posts that appeared frequently and differed only by a number in the link, which had been removed in the previous step. Finally, all numbers, special characters, and any remaining links were removed from the data. This systematic approach ensured thorough cleaning of the raw data, providing a solid foundation for accurate and meaningful analysis.

After data cleaning, our data set contained the following posts per platform (Table 2): for Twitter/X, *n* = 1,159,296; for Facebook Pages, *n* = 34,658; for Facebook Groups, *n* = 10,016; and for Reddit, *n* = 2,557 threads of 18,403 posts (*N* = 1,222,373 posts). Overall, the Facebook and Reddit platforms were laggards, with hashtags only occurring from 2013 onward. Posts on Reddit, in contrast to the other two platforms, have not decreased over the years, but have remained stable.

[3] Further details on the procedure and data can be viewed on the Open Science Framework at the following link: https://osf.io/5u6yd/.

https://doi.org/10.65476/xgpt3649

Table 2. Posts per Year and Platform.

Year	Twitter	Facebook pages	Facebook groups	Reddit	Sum
2011	23,630				23,630
2012	95,669				95,669
2013	125,455	1,723	297	13	127,488
2014	170,716	2,548	764	14	174,042
2015	191,356	3,220	1,024	227	195,827
2016	150,355	5,111	957	316	156,739
2017	169,338	5,869	1,403	426	177,036
2018	83,064	6,485	1,833	513	91,895
2019	102,402	5,599	1,800	495	110,296
2020	47,311	4,103	1,938	553	53,905
Sum	**1,159,296**	**34,658**	**10,016**	**2,557 threads = 18,403 posts**	**1,222,373**

Three points were particularly noticeable during the data-cleaning process: First, almost one million posts from Twitter dropped out during data cleaning, with numerous retweets and bot messages, including tweets that only contained a link, usually to a data business startup. Second, the difference between Facebook Pages and Facebook Groups is striking. Both serve different audiences, which is reflected in the way users interact and the keywords they like. Facebook Groups are generally intended for personal exchange and interaction between members. In contrast, Facebook Pages are often used to provide information about companies or brands (B2C marketing). In the data set, the focus is on posts from Facebook Pages and, therefore, a higher proportion of marketing messages can be assumed. This may result in an imbalance in the data set between messages from companies and interaction between users.

Last, our collected Reddit data often featured texts exceeding processing limits, with threads commonly surpassing 160–200 characters because of long posts and numerous comments. To address this and prevent bias toward early text segments during (semi) automated analysis, we chunked all longer entries into shorter processable sequences of 128 tokens from the original text, disregarding original sentence structures. This resulted in a data set of 18,403 text segments for Reddit, whereby these entries are nonhierarchical (i.e., it was no longer possible to distinguish between post and reply).

Classifying and Topic-Modeling Text Data With Transformer-Based Language Model Techniques

To analyze the large data sets, we combined approaches of constructionist analysis and topic modeling. Following Van Gorp (2010), cultural motifs are semantic patterns of knowledge materializing in observable linguistic usage. These patterns can be reconstructed using methods like text classification and topic models (Barberá, Boydstun, Linn, McMahon, & Nagler, 2021). Our approach employed text classification with active learning to assign cultural motif labels to every text sample. Active learning iteratively improves the classification model by selecting the most informative samples for annotation. Once they were classified, we applied topic modeling to identify topics associated with a cultural motif, capturing the thematic structures and nuances.

https://doi.org/10.65476/xgpt3649

We leveraged models from the Bidirectional Encoder Representations from Transformers (BERT) family for sentence classification and topic modeling because of their superior performance across these tasks (Devlin, Chang, Lee, & Toutanova, 2019). The embeddings generated by these models provide dynamic, contextual representations of text, effectively encoding the semantics of words based on their surrounding context. This approach mitigates issues such as polysemy and homonymy and enables the transfer of extensive linguistic knowledge, including grammar, syntax, semantics, and factual information, from large pretrained models.

Classification

For the classification process, we used a robust active-learning framework (Schröder, Müller, Niekler, & Potthast, 2023) to detect and identify cultural motifs in extensive data sets. We utilized the Small-Text library to facilitate our active-learning experiments, benefiting from its user-friendly and consistent interface. For the classification task, we selected the SetFit method (Tunstall et al., 2022) using the model paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych, 2019) with the default configuration provided through Hugging Face (2023) because of its efficiency and effectiveness in our experiments. Our approach incorporated a combination of weak supervision and active-learning techniques (Schröder et al., 2023) for data set creation and model training. Active learning utilized query strategies to select the most informative samples from an unlabeled pool of data, guided by a classifier trained on an existing labeled data set.

Model performance was progressively enhanced through an iterative process of active learning, integrating automated and human coding. During each of the roughly 20 training rounds, the classifier strategically selected approximately 100 examples about which it was most uncertain. These high-uncertainty examples were then manually labeled by humans, and the expanded data set was used to retrain the classifier. The predefined stopping criterion for this refinement process was the completion of 20 training rounds each expanding the training data with new examples. This culminated in the creation of a new, comprehensive training data set covering all platforms and including both English and German texts. To evaluate our active-learning model, we created a balanced labeled set from the annotated data. This was crucial because a dedicated validation set was missing, and our initial data had a class imbalance, especially for the minority class, which we ensured was well-represented by down sampling other classes. We then ran a tenfold cross-validation on this set to get a more robust and less biased estimate of overall quality. Achieving an F1 score of .70 was deemed mediocre, but acceptable, primarily because of the strong variability and indexicality of UGC, which inherently presents significant challenges for automated text classification. A comprehensive breakdown of evaluation metrics for all classes is presented in Table 3.

https://doi.org/10.65476/xgpt3649

Table 3. Classifier Evaluation Based in Tenfold Cross-Validation of Balanced Data Set.

	Precision	**Recall**	**F1 score**
Low-profile surveillance	.697	.729	.709
Shift in datafying society	.820	.817	.817
Civic agency	.688	.700	.693
Innovations for societal progress	.553	.541	.545
Negative consequences	.761	.711	.732
Preventing wrongs	.736	.736	.734
Profits and prediction	.713	.710	.711
Macro avg	***.710***	***.706***	***.706***

In each training iteration, we assessed how well the human coder's labels aligned with the classifier's suggestions for the uncertain examples. This agreement, measured by the F1 score, improved from .47 to .61 across the active-learning iterations. This increase signifies that the classifier's uncertain suggestions became progressively more aligned with the coder's decisions, indicating a genuine improvement in the model decision-making capability. Finally, the fully trained classifier was applied to the entire data set to assign a label to each text fragment.

Topic Modeling

As demonstrated, it is possible to replicate manually detected cultural motifs based on precoded paragraphs using topic models (Knorr et al., 2023). Likewise, we were able to map the possible thematic spectrum of each cultural motif from the classified data set. For this purpose, we used BERTopic (Grootendorst, 2022), a topic-modeling technique that leverages transformer-based language model embeddings to create dense clusters of semantically similar documents. By using the pretrained model paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych, 2019), contextual nuances in the text are captured, allowing for more accurate and meaningful topic extraction compared with traditional methods like latent Dirichlet allocation, especially for short text utterances in UGC (Egger & Yu, 2022).

The process involved embedding documents using a transformer-based language model, reducing dimensionality of semantic representations with techniques like uniform manifold approximation and projection, and clustering the reduced embeddings using hierarchical density-based spatial clustering of applications with noise. This resulted in the identification of coherent linguistic usage contexts (i.e., topics that reflect the underlying structure or central aspects of each cultural motif in the data set, where a topic represents a cluster of thematically related terms; Hofmann, 1999).

Each topic consists of a statistically specific combination of keywords that occur frequently in the texts. Because we focused on the most prominent text contexts (i.e., topics) that characterize a cultural motif, we concentrated on the three most prominent topics grouped below a cultural motif. We employed the language model ChatGPT to create label descriptions for the topics generated by BERTopic. A structured prompt, provided alongside each topic's word list, guided ChatGPT to produce a distinct semantic and narrative label, along with a comprehensive description, for each topic. This methodology is acknowledged for its effectiveness in deriving meaningful topic descriptions (Piper & Wu, 2025). Care was taken to verify

https://doi.org/10.65476/xgpt3649

and validate all AI-generated content for accuracy and relevance. In the following, we present the most important topics given by the BERTopic analysis for all cultural motifs in the context of big data in user-generated communication.

Results

Cultural Motifs of Big Data (RQ1)

In response to RQ1, it can be stated that the three overarching cultural motifs in the data set were "profits and prediction," "shift in datafying society," and "negative consequences." In the UGC, the predominant cultural motif was "profits and prediction," found in approximately 20K posts every year. It remained the most prominent motif over the period studied. In the first part of the time frame in focus, the cultural motif "shift in datafying society" was the second-most frequent motif, appearing in about 10K posts per year. Toward the end of 2017, it was surpassed in volume by "negative consequences."

Thus, in the discourse around big data, the leitmotif was economic in nature and very much structured around revenue and business opportunities stemming from predictive analytics. However, a reckoning with possible negative consequences of data analytics came to the fore from 2017 to 2018 onward. Until then, big data was, next to commercial exploitation, mainly associated with a profound sea change in the way we generate knowledge in and about society, for good and bad. It implies a transformative potential that would upend—so the general thrust of the three motifs—all kinds of sectors and occupations (Figure 1).

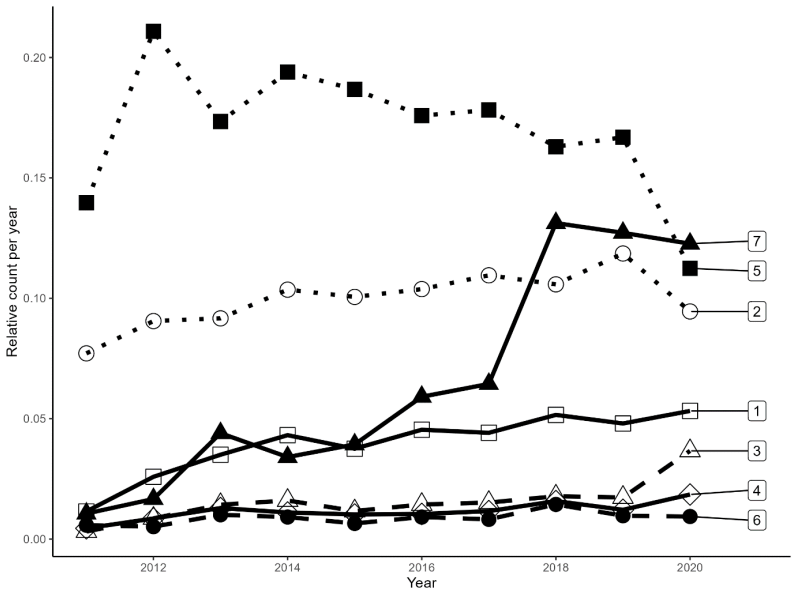

Figure 1. Development of cultural motifs, 2011–2020.

Note. (1) innovations for societal progress, (2) shift in datafying society, (3) preventing wrongs, (4) low-profile surveillance, (5) profits and prediction, (6) civic agency, (7) negative consequences.

https://doi.org/10.65476/xgpt3649

The other four cultural motifs whose existence and development we analyzed were less often evoked from the outset and did not change much throughout the entire decade. Fourth came the cultural motif "innovations for societal progress," which stood in close connection to the motif "shift in datafying society." Fifth ranked was "preventing wrongs," which was closely linked to "low-profile surveillance." Last was the cultural motif "civic agency." Thus, in essence, big data is primarily featured in an aspirational discourse of prospective innovations, new opportunities, and potentials. Issues of security and surveillance came second and were tied to concerns about crime prevention, privacy, and the right to one's own data. Economic and administrative considerations dominated the discourse; discussions of civic engagement and participation in social datafication came second.

Cultural Motifs in Development (RQ2)

Considering the development of cultural motifs in response to RQ2, it is evident there are larger diachronic trends instead of temporary shifts. Overall, this development can be understood as an ongoing public reckoning with big data. In its course, no new motifs were established. Rather, the discourse reiterated and adapted the existing repertoire of cultural sensemaking around new technologies (Feenberg, 2002). It encompassed the common antipodes of chance and risk coupled to positive and negative positions. In this structure, a technology's specific design, usage, and implications are spelled out and made the element of approval or rejection with some more balanced and contextualizing views in between. Thus, what is more striking in the big data discourse is not the rise and fall of motifs, but their robustness and existence over the entire period.

Looking at the prevalence of the three primary cultural motifs "profits and prediction," "shift in datafying society," and "negative consequences" makes it evident that big data was initially mainly tied to forward-looking and affirmative expectations, reflecting how data could drive economic growth and improve societal functions. The commercial aspect was particularly prominent, with a view on businesses leveraging data analytics for profit maximization and strategic forecasting. With a similar enthusiasm, the motif "shift in datafying society" culminated around 2017, emphasizing the revolutionary integration of data into all walks of life, personal and professional. The motif declined in prominence around the same time as the main motif "profits and prediction," indicating a possible end of the hyperbolic excitement and a normalization of datafication in society around 2018.

By contrast, the motif "negative consequences" began to rise in frequency from 2015 onward and remained prominent between 2018 and 2020, when it reached its peak. This motif underscored critical perspectives on big data, focusing on issues such as privacy concerns, data misuse, and ethical implications. While the motif "negative consequences" stressed more personal initiative to address these challenges, the motif "civic agency" highlighted issues of transparency, data policies, and ethical governance on a political level. Here, the possible negative impacts were also treated with ambivalence. On the one hand, discussions surrounding the potential dangers of mass surveillance were prevalent; on the other, there was an emphasis on the opportunities for preventive measures against such threats. However, both motifs—preventing wrongs (rank 5) and low-profile surveillance (rank 6)—were among the less well-established, exhibiting a comparably low peak in 2019.

https://doi.org/10.65476/xgpt3649

In a way, the motif "innovations for societal progress" seemed to bridge the discourse pivoting on the three prevalent motifs (profits and predication, shift in datafying society, negative consequences) and a less articulated discourse on the political and legal governance of big data anchored in the three motifs of civic agency, low-profile surveillance, and preventing wrongs. All of them were driven by an interest in big data's leverage. In one direction, the sensemaking pointed to chiefly commercial uses, whereas the other directed toward societal uses, reflecting a more nuanced understanding that included its potential drawbacks. These critical views have risen in prominence since 2017, underscoring possible negative consequences too.

Cultural Motifs and Topics Interlinkages (RQ3)

The thematic interpretation of the cultural motifs rests on an analysis that yielded several topics encapsulated by a cultural motif. With respect to RQ3, we found each cultural motif captured a mix of topics. Put differently, each topic provided an aspect of the semantic setup of a cultural motif. With the statistical information about the typical clusters of keywords, we imputed the thematic focus for each topic. Accordingly, we explicate the cultural motifs along the three most prominent topics (Table 4). For clarifying the semantic relationship between cultural motifs and topics, we selected representative example texts from one of the three platforms. These examples are solely presented to illuminate the conceptual linkage. In the following, we group the seven cultural motifs along their shared orientation toward either seeing big data as a catalyst for transformation or as contextualizing big data sociopolitically.[4]

[4] We acknowledge the assistance of the ChatGPT language model by OpenAI in generating and refining the cultural motifs. The final content and interpretation remain the responsibility of the authors.

https://doi.org/10.65476/xgpt3649

Table 4. Cultural Motifs on Twitter/X, Facebook, and Reddit (UGC) and Their Three Dominating Topics Demonstrated With an Example From January 1, 2011, to December 31, 2020.

Cultural motif	Prominent BERTopic and UGC posts example per each topic
1. **Innovations for societal progress:** advancements that drive positive change in society	Health care, cancer, medicine, analytics, doctors, improve, medical, patient "Doctors Use Big Data to Improve Cancer Treatments."
	Learning, intelligence, machine, artificial, innovation, blockchain, future, next "Healthcare will be radically transformed by big data, constant connectivity and machine learning."
	Ebola, coronavirus, fight, spread, disease, outbreak, diseases, malaria, Africa "How Big Data and real-time analytics could help fight the spread of Ebola."
2. **Shift in datafying society:** data and technology are transforming various aspects of society	Analytics, cloud, learning, new, machine, digital, intelligence, smart, internet, technology "How to Use Cloud #BI #Analytics to Drive Innovation."
	Social, revolution, media, world, change, future, ways, changing, good, revolutionizing "[Podcast] Using Data to Create Social Change."
	Thanx, wish, following, discussion, good, potential, cases, use, using, impact "Bill Fox thanx for following and I wish a good discussion on #digitaltransformation #Bigdata #ehealth and #management."
3. **Preventing wrongs:** disaster management mitigating harm	Security, cyber, hackers, analytics, fight, intelligence, cloud, artificial "Using 'Big Data' to Fight Hackers."
	Pandemics, disaster, fight, disasters, lives, response, crisis, terrorism, conflict "Using Big Data to Fight Pandemics #bigdata."
	Risk, food, risks, safety, insurance, management, oil, insurers, alert, supply "Using big data could alert us to risks in the food supply chain."
4. **Low-profile surveillance:** cybersecurity and data privacy	Security, privacy, cloud, IBM, cyber, need, IoT, protect, insight, challenges "IBM Addresses Security Challenges of Big Data, Mobile and Cloud Computing."

https://doi.org/10.65476/xgpt3649

	Fraud, detection, insurance, tax, prevention, theft, fight, investigations "Big Data: The Future of Insurance Fraud Prevention."
	Democracy, legal, election, elections, voters, government, regulatory, law, public "Big data is not a game played by different rules." "Big data: Managing the legal and regulatory risks."
5. **Profits and prediction:** improving financial and political operations and decision making	Business, analytics, marketing, companies, customer, use, new, intelligence, market, ways "Enterprises advanced their big data initiatives by converting plans into working projects and even implementations of big data that have transformed the business."
	Ways, revolutionizing, HR, Wissen, HLEN, Wen, humanizing, via, profoundly, Pentland "10 Ways #BigData Is Revolutionizing #SupplyChainManagement."
	Election, Obama, Trump, elections, campaign, win, presidential, president, Cambridge, Analytica "Elections and the Internet. Big Data Research from the Oxford Internet Institute."
6. **Civic agency:** individuals and institutions can navigate and exercise power	Privacy, Facebook, consent, world, debate, new, concerns, social, renewed, instead "A long but fascinating read about current legislation around our data, how it is impacting our perception of privacy, and the need (or lack of need) for legislation around ethical principles to protect our future."
	Security, secure, securing, internet, hadoop, IoT, things, protect, datenschutz, protecting "8 ways you can help secure the Internet of Things."
	Legal, law, side, lawyers, issues, industry, profession, firms, moneyball, contracts "The future legal and security system cannot be separated from the internet and big data."
7. **Negative consequences:** vulnerabilities and risks associated with big data and AI	Facebook, privacy, security, breaches, major, e-mail, services, found, Google, media "Exclusive—Big data breaches found at major e-mail services: Expert."

https://doi.org/10.65476/xgpt3649

	Artificial, analytics, intelligence, learning, algorithms, machine, biases, bias, Oracle, trust "Big data and machine learning algorithms could increase risk of collusion: ACCC #News phone."
	Election, Trump, democracy, government, Russian, voters, Russia, elections, voting, House "Russian hacking? No. This is how Trump won the US election: The Data That Turned the World Upside Down."

Note. The posts cited in Table 4 and in this article were collected and analyzed without user details or contact information. They are derived from the time-dependent platform peaks described above but not from the author information.

Big Data as Catalyst of Transformation

Four of the seven motifs revolved around the transformative power and potential of big data, primarily for making profit or revamping society as such. There was an overarching thematic line here that conceived of big data as the catalyst for the most significant social revolution of the decade, with those able to use the data deemed in a position of power.

With that general orientation, the main motif in the user-generated discourse "profits and prediction" was economically oriented. Here, all three identified topics represented different facets of how big data and predictive analytics could be used to improve financial and political operations and decision making. By integrating big data into business processes, companies could predict trends and optimize marketing strategies to drive growth. As such, the first two of the most prominent topics contained keywords that link analytics with marketing, decision-making processes with data exploitation, and companies with their consumers. Promises like "10 Ways #BigData Is Revolutionizing #SupplyChainManagement" are emblematic of that way of thinking. Especially the second topic characterizing the motif illustrated the broader impact of big data across different industries. It foregrounded big data's ability to revolutionize sectors like supply chain management and marketing. This also included ideas about predictive analytics transforming traditional practices, thus leading to more efficient operations and, ultimately, higher profitability. In the context of political campaigns, a third topic dealt with how predictive analytics may be used to influence elections. It therefore took up a distinct kind of profit that was political in nature, not economical. Here, keywords such as "election," "Obama," "Trump," and "Cambridge Analytica" were prominent. Overall, the topics referring to profits and prediction specified how data analytics could be successfully used for decision-making processes across different areas (business, industry, politics) and to maximize outcomes, whether those outcomes were financial profits or electoral victories.

In contrast, the second cultural motif, shift in datafying society, was broader in its outlook on data and technology transforming all sorts of aspects of society. The first topic contained keywords like "analytics," "cloud," "machine learning," "digital," and "intelligence." This suggests a focus on the infrastructure and tools that enable the collection, processing, and utilization of social data. The second topic captured how big data would revolutionize social life with words like "social," "revolution," "media," "world,"

https://doi.org/10.65476/xgpt3649

and "change." The third topic revolved around the engagement of individuals with data-driven technologies and their practical applications. Keywords like "use cases," and "impact" illustrate the practical engagements that define this shift.

The cultural motif "negative consequences" described the vulnerabilities and risks associated with big data. The first topic centered around risks associated with e-mail and social media services like Google and Facebook. The focus was on data breaches and compromised user privacy. A second topic took up the potential negative consequences of biases embedded in algorithms. The third topic was bound to a specific event during the US elections in 2017. It included keywords like "election," "Trump," "democracy," "Russian," and "voters." Posts communicate ideas such as "Russian hacking? No. This is how Trump won the US election: The Data That Turned the World Upside Down."[5] All these topics reflect specific concerns related to the possible misuse or negative impacts of big data.

Until 2017, the negotiation of possible negative consequences was intertwined with a fourth cultural motif of innovations for societal progress. Both cultural motifs—"negative consequences" and "innovations for societal progress"—share keywords related to artificial intelligence and analytics. However, eschewing negative connotations, innovations for societal progress related to advancements that would drive positive change in society. Its first topic focused on how innovations in big data and analytics are transforming health care, stating that big data could lead to cures for diseases like cancer. The second topic noted the role of emerging technologies like blockchain in driving the future of health, illustrated by statements like "Doctors Use Big Data to Improve Cancer Treatments" and "Healthcare will be radically transformed by big data, constant connectivity, and machine learning." A third topic revolved around the use of big data and real-time analytics to combat global health crises such as Ebola and COVID-19.

Sociopolitical Contexts of Big Data

Three cultural motifs shared a common orientation toward the social and political contexts of big data. While the motifs grouped under the rubric of big data as catalysts gravitated around its transformative force believed to override existing conditions and circumstances, the other three related big data technologies, analytics, and ambitions to existing practice and predominant usages.

"Preventing wrongs" related to disaster management that mitigated harm thanks to big data technologies; "low-profile surveillance" opened up the context of cybersecurity and data privacy. Therefore, the first topic in "preventing wrongs" focused on the use of advanced analytics to combat hackers and improve intelligence gathering. A second topic was about pandemic prevention as well as the prevention of larger crises, including terrorism, which situated the cultural motif "preventing wrongs" on a broader societal level. In the posts, big data was, for instance, said to alert us "to risks in the food supply chain," to "Fight Hackers," or to "Fight Pandemics."

[5] Note that the posts cited were collected and analyzed without user details or contact information. They are derived from the time-dependent platform peaks described above, but not from the author information.

https://doi.org/10.65476/xgpt3649

In close semantic connection, the first topic of low-profile surveillance focused on the challenges of protecting data in the age of cloud computing and the Internet of Things. It underscored the need for transparency and adherence to laws to protect democratic integrity and secure data privacy. Posts such as "big data is not a game played by different rules," "Big data: managing the legal and regulatory risks," or "IBM Addresses Security Challenges of Big Data, Mobile and Cloud Computing" expressed that line of thought.

The two cultural motifs were joined by the cultural motif of "civic agency." Its topics reflected on how both individuals and institutions may navigate and exercise power under the conditions of big data. The first topic centered on the growing concern for privacy in the digital world, with calls for legislation and ethical principles to protect people's data and personal freedom. A second topic zoomed in on cybersecurity. The third topic was how laws and regulations had to be shaped to protect citizens' rights. The broader theme in civic agency was that of empowering individuals and professionals to navigate the complex legal landscape to preserve and protect their privacy.

Discussion and Conclusion

Since its inception around 2010, big data has been related to different cultural motifs in parallel to tremendous technological progress and numerous debates. Some of these individual discourses have been studied, yet little is known about their evolution over a longer period of time.

In our analysis, we found seven cultural motifs that illustrate the extent to which big data, as a keyword, pertains to socially pressing issues that predicate on datafication. Especially three major cultural motifs—"profits and prediction," "shift in datafying society," and "negative consequences"—entail topics that refer to efficiency and profitability gains in the financial sector and healthcare system while also evoking possible detriments. Furthermore, the motifs and their topics suggest an ambivalent discursive construction of big data as a socio-material force that is a driver of innovation and a risk factor alike. Such ambivalent construction is not new, but reappears regularly to catch double-edged ramifications of a new technology. By this token, big data becomes yet another occasion for rehashing ideas reminding us of the sorcerer's apprentice who is hoping to unleash a technology's potential, but is overpowered by its magnitude (Moss, 2011).

Therefore, in the discourse around big data, critical and affirmative motifs coexist without merging or giving way to new motifs. In a technologically highly innovative field and despite all societal transformations, there was discursive inertia. For sure, the discourse itself was volatile, yet in terms of reflecting big data's technology, analytics, and mythology, it rehearsed a well-known cultural repertoire. Looking at the basic discursive structure that undergirds the many trending topics, there is little movement. The inertia corresponds to what Lucia, Vetter, and Adubofour (2023) have named the "rhetoric of continuance and novelty" (p. 15) that emits from Facebook and other Silicon Valley enterprises. With this, the corporate players aim to balance their utopian promises of tech-enhanced futures with the majority of customers' reluctance to embrace all things new. As such, we find the same tropes applied to different innovations like that of global connectivity or global community, for instance, which were used to market different Meta services up to the Metaverse (Haupt, 2021). In another respect, the paucity of dominating motifs and the little development of the repertoire also demonstrate journalism's lack of critical imagination

https://doi.org/10.65476/xgpt3649

when faced with big data. Although it was instrumental in reporting the many different scandals that marked the period studied, it failed to provide genuinely new perspectives on the matter that would have helped users come to terms with datafication processes and their ramifications. This is also reflected in the quite monotonous and unimaginative illustrations used in US broadsheets to picture big data (Pentzold, Brantner, & Fölsche, 2018).

In its basic structure, the repertoire we found predates the big data hype, and some of its elements seem to appear once again in the discourse around AI (Mager & Katzenbach, 2021; Richter, Katzenbach, & Schäfer, 2023). Though it responds to an established cultural understanding of the chances and risks of technological innovations, at least in the Global North, it may not be able to appropriately capture nuances and genuine issues associated with big data. Somewhat surprisingly, in our material, the massive data scandal around Edward Snowden's revelations in 2013 did not do much to immediately change the basic setup of motifs. Only in the long run did the critical motifs gain in prominence. This shift, however, again did not coincide with another prominent data scandal around the Cambridge Analytica disclosures from March 2018, but preceded it (Knorr, Wolter, & Pentzold, 2024). The development was therefore more tectonic in character than imminent.

In these dynamics, where topics structure a motif's development, each topic is an aspect of a cultural motif reiterating common tropes found in the sociotechnical imaginaries surrounding technological innovation (Jasanoff, 2015). Topics may suggest new keywords, both platform-specific and cross-platform. This needs to be investigated further. It seems interesting to see to what extent they follow the logic of the well-known hype cycle of emerging technologies going through a period of exaggerated hopes and fears, a valley of disillusion, before becoming mundane (Mager & Katzenbach, 2021). Though the ideal trajectory of hype and gloom does rarely exist, the topics we found in big data discourse suggest some similar patterns.

In subsequent studies, limitations of our time series data utilized by BERT need to be addressed too. The time series generated by BERT was not continuous, but often more resembled a sequence of micro-events. These micro-events were short-term occurrences that garnered limited public attention measured in activities like sharing, retweeting, or commenting. As such, micro-events resulted in "scattershot" posts and discussions (Mahl, von Nordheim, & Guenther, 2023), often without a classic peak and rarely a spillover effect to other platforms or even news media. To contextualize micro-events across multiple platforms, a high level of data aggregation would be necessary to discern frequency patterns (Lehmkuhl & Promies, 2020). Moreover, a deeper level of analysis is required to capture both the broader context and impact of micro-events and their thematic entanglement of partly contradictory cultural motifs that characterize the discourse, such as "negative consequences" and "profits and prediction." These next steps will help to further unpack the cultural sensemaking around big data. Such investigation needs to include other languages and discursive arenas. The question is if these discourses evoke alternative motifs to the big data ideology of dataism (van Dijck, 2014). In these discourses, "big data" may feature as a keyword, yet not all cognate debates must be using the moniker. They can also employ either more loose rubrics of digital data and analytics or may run under specific notions like predictive policing or microtargeting (Flensburg & Lomborg, 2021; Kitchin, 2014).

https://doi.org/10.65476/xgpt3649

References

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis, 29*(1), 19–42. https://doi.org/10.1017/pan.2020.8

Barrie, C., & Ho, J. C. (2021). academictwitteR: An R package to access the Twitter Academic Research Product Track v2 API endpoint. *Journal of Open Source Software, 6*(62), 3272. https://doi.org/10.21105/joss.03272

Beer, D. (2018). *The data gaze: Capitalism, power and perception*. London, UK: SAGE Publications.

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). How frames can undermine support for scientific adaptations: Politicization and the status-quo bias. *Public Opinion Quarterly, 78*(1), 1–26. https://doi.org/10.1093/poq/nft044

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Burgess, J. E., Albury, K., McCosker, A., & Wilken, R. (2022). *Everyday data cultures*. Cambridge, UK: Polity Press.

Cools, H., Van Gorp, B., & Opgenhaffen, M. (2024). Where exactly between utopia and dystopia? A framing analysis of AI and automation in US newspapers. *Journalism, 25*(1), 3–21. https://doi.org/10.1177/14648849221122647

Couldry, N., & Yu, J. (2018). Deconstructing datafication's brave new world. *New Media & Society, 20*(12), 4473–4491. https://doi.org/10.1177/1461444818775968

David, C. C., Ong, J. C., & Legara, E. F. T. (2016). Tweeting supertyphoon Haiyan: Evolving functions of Twitter during and after a disaster event. *PLoS One, 11*(3), e0150190. https://doi.org/10.1371/journal.pone.0150190

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Vol. 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, MN: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology, 7*, 1–16. https://doi.org/10.3389/fsoc.2022.886498

https://doi.org/10.65476/xgpt3649

Eisele, O., Heidenreich, T., Litvyak, O., & Boomgaarden, H. G. (2023). Capturing a news frame—Comparing machine-learning approaches to frame analysis with different degrees of supervision. *Communication Methods and Measures, 17*(3), 205–226. https://doi.org/10.1080/19312458.2023.2230560

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(1), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Feenberg, A. (2002). *Transforming technology: A critical theory revisited*. New York, NY: Oxford University Press.

Flensburg, S., & Lomborg, S. (2021). Datafication research: Mapping the field for a future agenda. *New Media & Society, 25*(6), 1451–1469. https://doi.org/10.1177/14614448211046616

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology, 95*(1), 1–37. https://doi.org/10.1086/229213

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (Version 0.9.4) [Computer software]. Retrieved from https://github.com/MaartenGr/BERTopic

Guenther, L., Bischoff, J., Löwe, A., Marzinkowski, H., & Voigt, M. (2019). Scientific evidence and science journalism. *Journalism Studies, 20*(1), 40–59. https://doi.org/10.1080/1461670X.2017.1353432

Haupt, J. (2021). Facebook futures: Mark Zuckerberg's discursive construction of a better world. *New Media & Society, 23*(2), 237–257. https://doi.org/10.1177/1461444820929315

Hofmann, T. (1999). Probabilistic latent semantic indexing. In F. Gey, M. Hearst, & R. Tong (Eds.), *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57). Berkeley, CA: ACM. https://doi.org/10.1145/312624.312649

Hugging Face. (2023). Transformers (Version 4.30.0) [Computer software]. Retrieved from https://github.com/huggingface/transformers

Jasanoff, S. (2015). Future imperfect. In S. Jasanoff & S. Kim (Eds.), *Dreamscapes of modernity* (pp. 1–33). Chicago, IL: University of Chicago Press.

Jünger, J., & Gärtner, C. (2021). Distilling issue cycles from large databases: A time-series analysis of terrorism and media in Africa. *Social Science Computer Review, 39*(6), 1272–1291. https://doi.org/10.1177/0894439320979675

https://doi.org/10.65476/xgpt3649

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London, UK: SAGE Publications.

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1), 1–10. https://doi.org/10.1177/2053951716631130

Knorr, C., Niekler, A., Behret, M., & Pentzold, C. (2023, July 10–14). *Cultural motifs on #bigdata—A semi-automated topic modeling from a socio-cultural constructionist perspective*. Paper presented at ADHO Conference "Collaboration as Opportunity", Graz, Austria. Abstract retrieved from https://dh-abstracts.library.virginia.edu/works/12511

Knorr, C., & Pentzold, C. (2025). Making sense of "big data": Ten years of discourse around datafication. *Big Data & Society, 12*(2), 1–15. https://doi.org/10.1177/20539517251330181

Knorr, C., Wolter, M., & Pentzold, C. (2024). Whistleblower memoirs: Deconstructing data consultants' insider stories. *Social Media + Society, 10*(1), 1–10. https://doi.org/10.1177/20563051231224730

Lehmkuhl, M., & Promies, N. (2020). Frequency distribution of journalistic attention for scientific studies and scientific sources: An input–output analysis. *PLoS One, 15*(11), e0241376. https://doi.org/10.1371/journal.pone.0241376

Lohr, S. (2016). *Data-ism. Inside the big data revolution*. London, UK: Oneworld Publications.

Lucia, B., Vetter, M., & Adubofour, I. (2023). Behold the metaverse: Facebook's Meta imaginary and the circulation of elite discourse. *New Media & Society, 27*(2), 790–807. https://doi.org/10.1177/14614448231184249

Mager, A., & Katzenbach, C. (2021). Future imaginaries in the making and governing of digital technology: Multiple, contested, commodified. *New Media & Society, 23*(2), 223–236. https://doi.org/10.1177/1461444820929321

Mahl, D., von Nordheim, G., & Guenther, L. (2023). Noise pollution: A multi-step approach to assessing the consequences of (not) validating search terms on automated content analyses. *Digital Journalism, 11*(2), 298–320. https://doi.org/10.1080/21670811.2022.2114920

Markham, A. N., & Tiidenberg, K. (Eds.). (2020). *Metaphors of internet: Ways of being in the age of ubiquity*. New York, NY: Peter Lang.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London, UK: John Murray.

https://doi.org/10.65476/xgpt3649

Moss, F. (2011). *The sorcerers and their apprentices: How the digital magicians of the MIT Media Lab are creating the innovative technologies that will transform our lives*. New York, NY: Crown.

Nisbet, M. C., Scheufele, D. A., Shanahan, J., Moy, P., Brossard, D., & Lewenstein, B. V. (2002). Knowledge, reservations, or promise? *Communication Research, 29*(5), 584–608. https://doi.org/10.1177/009365002236196

Nolin, J. M. (2019). Data as oil, infrastructure or asset? Three metaphors of data as economic value. *Journal of Information, Communication and Ethics in Society, 18*(1), 28–43. https://doi.org/10.1108/JICES-04-2019-0044

Nye, D. E. (1996). *American technological sublime*. Cambridge, MA: MIT Press.

Paganoni, M. C. (2019). *Framing big data: A linguistic and discursive approach*. Basingstoke, UK: Palgrave Macmillan.

Pentzold, C., Brantner, C., & Fölsche, L. (2018). Imagining big data: Illustrations of "big data" in US news articles, 2010–2016. *New Media & Society, 21*(1), 139–167. https://doi.org/10.1177/1461444818791326

Pentzold, C., & Fischer, C. (2017). Framing big data: The discursive construction of a radio cell query in Germany. *Big Data & Society, 4*(2), 1–11. https://doi.org/10.1177/2053951717745897

Pentzold, C., & Fraas, C. (2023, November). Media frames as adaptive networks of meaning: A conceptual proposition. *Language & Communication, 93*, 95–106. https://doi.org/10.1016/j.langcom.2023.09.001

Pentzold, C., & Knorr, C. (2023). When data became big: Revisiting the rise of an obsolete keyword. *Information, Communication & Society, 27*(3), 600–617. https:/doi.org/10.1080/1369118X.2023.2227673

Piper, A., & Wu, S. (2025). Evaluating large language models for narrative topic labeling. In M. Hämäläinen, E. Öhman, Y. Bizzoni, S. Miyagawa, & K. Alnajjar (Eds.), *Proceedings of the 5th international conference on natural language processing for digital humanities* (pp. 281–291). Albuquerque, NM: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.nlp4dh-1.25

Portmess, L., & Tower, S. (2015). Data barns, ambient intelligence and cloud computing: The tacit epistemology and linguistic representation of big data. *Ethics and Information Technology, 17*(1), 1–9. https://doi.org/10.1007/s10676-014-9357-2

Puschmann, C., & Burgess, J. (2014). Metaphors of big data. *International Journal of Communication, 8*, 1690–1709.

https://doi.org/10.65476/xgpt3649

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1410

Richter, V., Katzenbach, C., & Schäfer, M. S. (2023). Imaginaries of artificial intelligence. In S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 209–223). Cheltenham, UK: Edward Elgar.

Rieder, G. (2018). Tracing big data imaginaries through public policy: The case of the European Commission. In A. R. Sætnan, I. Schneider, & N. Green (Eds.), *The politics and policies of big data: Big data, big brother?* (pp. 89–109). New York, NY: Routledge.

Ryan, C., & Gamson, W. A. (2006). The art of reframing political debates. *Contexts, 5*(1), 13–18. https://doi.org/10.1525/ctx.2006.5.1.13

Scannell, L., & Gifford, R. (2013). Personally relevant climate change. *Environment and Behavior, 45*(1), 60–85. https://doi.org/10.1177/0013916511421196

Scheufele, D. A., & Lewenstein, B. V. (2005). The public and nanotechnology: How citizens make sense of emerging technologies. *Journal of Nanoparticle Research, 7*(6), 659–667. https://doi.org/10.1007/s11051-005-7526-2

Schröder, C., Müller, L., Niekler, A., & Potthast, M. (2023). Small-text: Active learning for text classification in python. In *Proceedings of the 17th conference of the European chapter of the Association for Computational Linguistics: System demonstrations* (pp. 84–95). Dubrovnik, Croatia: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-demo.11

Streeter, T. (2011). *The net effect: Romanticism, capitalism, and the internet* (Vol. 32, Critical Cultural Communication). New York, NY: New York University Press.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv*. Retrieved from https://arxiv.org/abs/2209.11055

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society, 12*(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776

Van Gorp, B. (2007). The constructionist approach to framing: Bringing culture back in. *Communication Research, 57*(1), 60–78. https://doi.org/10.1111/j.0021-9916.2007.00329.x

https://doi.org/10.65476/xgpt3649

Van Gorp, B. (2010). Strategies to take subjectivity out of framing analysis. In P. D'Angelo & J. A. Kuypers (Eds.), *Doing news framing analysis: Empirical and theoretical perspectives* (pp. 84–109). New York, NY: Routledge.

Van Gorp, B., & Vercruysse, T. (2012). Frames and counter-frames giving meaning to dementia: A framing analysis of media content. *Social Science & Medicine, 74*(8), 1274–1281. https://doi.org/10.1016/j.socscimed.2011.12.045

Wyatt, S. (2021). Metaphors in critical internet and digital media studies. *New Media & Society, 23*(2), 406–416. https://doi.org/10.1177/1461444820929324

https://doi.org/10.65476/xgpt3649