Cultural Motifs on #bigdata - A Semi-Automated Topic Modeling from a Socio-Cultural Constructionist Perspective

Knorr, Charlotte
charlotte.knorr@uni-leipzig.de
Leipzig University, Germany

Niekler, Andreas
aniekler@informatik.uni-leipzig.de
Leipzig University, Germany

Behret, Marius
m.behret@studserv.uni-leipzig.de
Leipzig University, Germany

Pentzold, Christian
christian.pentzold@uni-leipzig.de
Leipzig University, Germany

Abstract

In this study, we show a possible operationalization of the sociocultural framing concept by van Gorp (2010). Communication Research and Computational Humanities are exploring how the keyword #bigdata can be investigated via combined approaches of both framing analysis and topic modeling. A first study shows, how both methodological approaches complement each other profitably regarding the analysis of cultural motifs, which in turn allows to quantitatively capture such abstract and complex concepts in large data sets.

Aim

Part of the history of science is its self-constitution through discourse, also its conceptualization and implementation of both theoretical and methodological approaches. Hereby, scientific discourse can be facilitated by research collaborations in which already proven theoretical approaches, and methodological procedures can be tested collaboratively to adapt them, for example, to examine new media and data environments. This work reports on such a collaboration.

As part of the project Framing Big Data (Deutsche Forschungsgemeinschaft (DFG) 447465824), which started in April 2021, the departments Computational Humanities and Media and Communication Research (both Leipzig University) are conducting a joint research program to investigate how the media-communicative framing of Big Data can be collected, analyzed, and evaluated both on a cross-national and on a time-comparative basis. We examine the media-communicative framing of Big Data and globally datafied processes in both press- and user-generated aggregates (Facebook, Reddit, Twitter) over a period of the last 10 years and in three countries (U.S., Germany, and South Africa) using both frame-analytic and discursive approaches.

In a way, we are researching large quantities of text data about Big Data. In our data, thousands of press releases and social media data from Facebook, Reddit and Twitter, extracted on the keywords #bigdata and “Big Data”, we recognize discursive references to public events such as political debates on regulation, election campaigns, data scandals, and economic innovations in the tech business. Following Gamson and Modigliani, a discourse can be understood as “a set of discourses that interact in complex ways” (Gamson, & Modigliani, 1989, p. 2). However, the discourse on Big Data is apparently triggered by the keyword Big Data, which in turn is not associated with a single topic. Moreover, several themes at different times are merged, which – in addition – are hardly reflected by key events in our data. Furthermore, resonance effects between User Generated Content (UGC) and the press – as they are consensus in framing research (e.g., Benford, & Snow, 2000) – seem to be mostly absent in the discourse on Big Data.

This is where this report comes in: The Departments Media and Communication Research and Computational Humanities at Leipzig University are exploring how #bigdata can be investigated via combined approaches of both framing analysis and topic modeling. We ask: How can Topic Modeling be designed to reveal in Topics not only individual issues, but also their implicit and frame-analytic references?

Background

Starting from a socio-cultural framing approach (Van Gorp, 2010), we argument that each topic of #bigdata contains semantic domain-typical patterns. These domain-typical patterns are subject to a nature-given implicitness (Ryan, & Gamson, 2006). Furthermore, these patterns refer to the frames, which become visible over time in the discourse on Big Data. Hereby, cultural motifs provide meaning to these patterns. Following Van Gorp (2010), cultural motifs are “the implicit cultural phenomenon” (p. 97) that can be extracted from a statement. In this vein, cultural motifs become visible in annotating an event or an actor – be it via hashtag, tweet, post, thread, subreddit or even just a hyperlink. We argue, the cultural motifs serve as anchor for both framing analysis and Topic Modeling. How Big Data is referred to over time, which frames are initiated and addressed, is anchored in their topics with their specific cultural motifs.

Study

During the manual coding process, we preliminary read 20 books that dealt with Big Data as socio-technological phenomenon (boyd & Crawford, 2012). Applying an inductive-deductive based framing analysis (Van Gorp, 2010), we were able to identify eight cultural motifs (Fig. 2). Afterwards, the paragraphs from the books, the manual analysis was based on, were used as the basis for the Topic Model. We followed the approach of Maier et. al. (2018) to implement a methodologically valid processing chain for the modeling approach. Each paragraph represents a document, and typical NLP-procedures and preprocessing-steps (Fig.

1) were used to further process the textual data. We extracted significant collocations but not all bigrams to represent important multiword tokens. Then, we applied the implementation from the topicmodels package in R to the data and the hyperparameters of the Topic Model were deter-mined by using the Jensen-Shannon-Divergence from the ldatuning package. Afterwards, the generated Topic Model was validated and finally evaluated by comparing the manually coded cultural motives with the Top-20 terms from each topic from the Topic Model, which showed that the results from the Topic Model are similar to the results of the manual coding process.

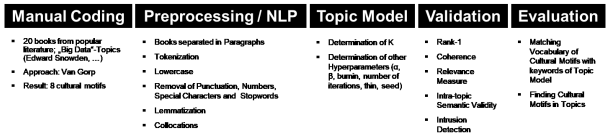

Figure 1: Representation of the process chain (left to right) from manual coding to the assignment of subjects to the manually determined motifs. For the steps, the methodological implementation was essentially based on (Maier et. al.).

Interestingly, quite comprehensible cultural motifs emerge such as data capitalism, Surveillance, crime prevention, increased efficiency or social empowerment (Figure 2). Where, i.e. within which digital spaces/platforms and with the help of which technical affordances #bigdata is thematized and thereby framed in terms of media communication, is not only a technical question, but also a socio-political one. A different openness to technological innovations can be stated during the last ten years. Also, the political governance for data protection and data storage have become more critical. In this vein, the results gained are also to be discussed collaboratively and against the background of the increasing datafication of society and science.

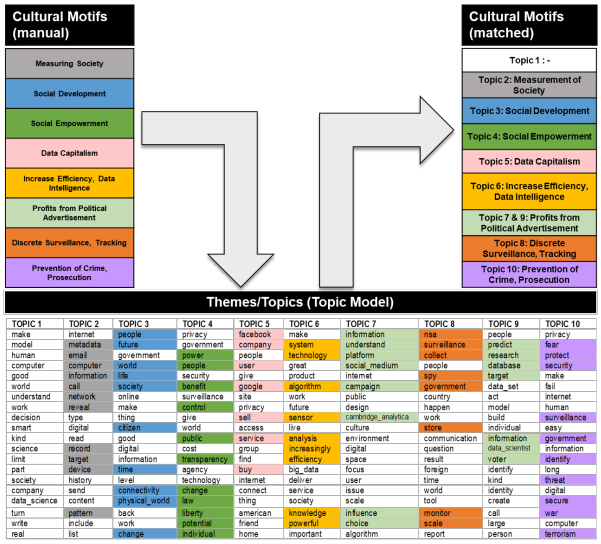

Figure 2: Illustration of the identified motives in the manually coded data and the assignment process to the result of the Topic Model. The assignment of specific colors was initially done by the researchers and the transfer of colors was then supported with a text search. First, the cultural themes were translated into English and then synonyms were identified. The synonyms were searched for in the topics table and colored when found. We added more words to the search based on our understanding of their meanings and cultural motifs and colored them accordingly.

Final Remarks

The collaboration between Computational Humanities and Media and Communication Research is fruitful for several reasons: First, the tangible benefits of a frame-based topic modeling can be highlighted. With the application of a topic model, it is possible to practically replicate the manually detected cultural motifs based on pre-coded paragraphs. Secondly, manual and automated approaches can be contrasted, which means in that case, to improve the manual frame analysis from the books with the help of automated procedures. Likewise, the results from the semi-manual and semi-automated methods can be used as seeds in a mixed-method scenario to analyze frames in large datasets. By creating a Topic Model based on an inductive-deductive framework analysis, we can search for cultural motifs in very large data sets and ultimately evaluate them quantitatively. A modeling process guided in this way enables the valid traceability of searched phenomena in very large data sources.

Bibliography

Benford, R. D., & Snow, D. A. (2000). *Framing Processes and Social Movements : An Over-view and Assessment* . Annual Review of Sociology , 26, 611–639. http://www.jstor.org/stable/223459

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

Gamson, W. A., & Modigliani, A. (1989). *Media Discourse and Public Opinion on Nuclear Pow-er: A Constructionist Approach* . American Journal of Sociology , 95(1), 1–37. https://www.jstor.org/stable/2780405

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

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). *Applying LDA topic modeling in communication research : Toward a valid and reliable methodology* . *Communication Methods and Measures* , 1–26. https://doi.org/10.1080/19312458.2018.1430754

Van Gorp, B. (2010). Strategies to Take Subjectivity Out of Framing Analysis. In P. D´Angelo & J. A. Kuypers (Eds.), Communication Series. *Doing News Framing Analysis: Empirical and Theoretical Perspectives* (pp. 84–109). Routledge.