Extended Abstract: Adding web data of local governance authorities to the toolbox of regional studies

Due to the increasing availability of web data, as well as the development of new methods for extracting information from this data, the last decade saw a rise of web mining approaches for empirical research in the social sciences, including regional science and economic geography. This has provided novel ways of analyzing regional economic activities with applications ranging from research on R&D activities of firms (Gök et al., 2015), firm growth (Li et al., 2018), innovation ecosystems (Kinne & Axenbeck, 2020; Kinne & Lenz, 2021), the *digital layer* of firm activities (Abbasiharofteh et al., 2023a, 2023b; Kriesch, 2023; Krüger et al., 2020), innovation diffusion (Lenz & Winker, 2020) as well as sentiment and content of regional news data (Ozgun & Broekel, 2021, 2022). While most of these previous studies have focused on mining web data at the firm-level, the aim of our study is the conceptualization of a comprehensive approach for using web content and web structure data of local government institutions. By creating a novel database, showcasing the versatile potential of web mining and machine learning techniques, we are the first study to propose an approach for the creation of indicators that depict local governance institutions based on web content data. With this, we contribute to progress in empirical research by providing a new way of quantifying and analyzing the role of local governance institutions as facilitators for regional (economic) development, adding to the data- and toolbox of spatial research. Our study is structured as follows. Firstly, we describe our methodology for creating the database. Secondly, we use the well-established topic modelling method 'GoogleBERT' to identify policy priorities on German municipal websites. Thirdly we introduce the method of Few-Shot Learning to classify texts on the topic of sustainability on the websites of German counties and municipalities. Fourthly, we discuss potential applications of the results as well as implications for further research.

To generate a comprehensive web-database of local governance institutions we collect website data at both the municipal and county level in Germany via the web crawling- and scraping-framework 'Scrapy'. We employ exhaustive crawls, to maximize the amount of data generated, creating an extensive html-database. Subsequently, we extract the core text from the html which is then cleaned and filtered, leaning on criteria from Rae et al. (2022), to exclude irrelevant as well as poor data. These text blocks are then split into paragraphs. This leaves us with a text-database of circa half a million paragraphs for German counties and circa 4 million paragraphs for German municipalities.

While the use-cases for our data are manyfold, the primary issue of this study is to map regional policy priorities, providing a firm foundation for discussing regional policy narratives. Government websites are used to improve external stakeholder perceptions (Youngblood & Mackiewicz, 2012) and reflect government priorities and strategies (Feeney & Brown, 2017; Sandoval-Almazan & Gil-Garcia, 2012). Considering that global or national narratives can be observed to trickle down into regional policies (Lund & Vildåsen, 2022), depicting these priorities and strategies thereby allows us to map policy narratives and discourse in and between regions. To achieve this, we firstly apply the well-established neural network-based approach in form of the transformer model 'GoogleBERT' (Rogers et al. 2020) to identify the policy priorities of German municipalities. Secondly, we use Few-Shot Learning (Tunstall et al. 2022) to classify texts concerning the topic of sustainability on the websites of German counties and municipalities.

Preliminary topic modelling results using 'GoogleBERT' on the municipal level show that web data in Germany can be a valuable source of information about regional policy priorities. Using the text data of approximately 500.000 subpages we identified approximately 150 topics. For the 30 most frequent topics we differentiated general topics, which contain essential information such as municipal administration in contrast to policy topics, which indicate local policy priorities. Key policy priorities that are discussed on these websites include, amongst others, traffic and public transport, climate

change mitigation, digital infrastructure, urban land-use planning, tourism and urban as well as business development. These results suggest that web mining can be a valuable tool for depicting regional policy priorities.

As mentioned before, to showcase the potential Few-Shot Learning provides for topic modeling we classified the policy narratives and discourses concerning sustainability. To our knowledge there have been no studies that used web mining and topic modelling to systematically analyze policy narratives and discourses about any one topic on a large scale. Studies in this context have so far focused on the analysis of smaller sample sizes. While Meschede (2019) for example used content analysis to research the information dissemination related to sustainability on local government websites of the 15 largest German cities, Meub et al. (2023) used web mining to depict smart city potentials for three German regions. As outlined before, with our study we try to achieve a connection of the innovative methods of web mining and topic modelling with the analysis of information dissemination. The focus was on a large-scale analysis of sustainability policy narratives and discourses on local and regional government websites. Preliminary results using Few-Shot Learning on municipal- as well as county-level data show that sustainability is more prevalent on the county-level than on the municipal-level. The topic can be found on 80 percent of the counties websites resulting in circa 19.000 paragraphs, as well as on 35 percent of the municipalities websites resulting in approximately 82.000 paragraphs. Aggregated on the state-level we can observe that the relative frequency of the topic declines from West- to East-Germany. Focusing on the content of the text we find that it is possible to accurately identify all paragraphs concerning differing sustainability narratives such as bioeconomy or circular economy. Furthermore, it is not only possible to identify these differing sustainability narratives, but also to identify policy strategies that are implemented to tackle the topic of sustainability.

Future potential applications of this novel data source are expansive and varied. Mapping local and regional policy narratives opens the door to a wealth of research questions. Future research could for example analyze the regional differences between policy priorities and narratives and how the political party affiliations of local and regional government influence these. Another interesting focus of analysis could be the comparison of county and municipal level priorities to assess the cohesiveness of regional policy narratives and the spatial differences concerning this matter. Furthermore, through the classification of text concerning certain topics it will be possible to create large, novel databases for qualitative research, analyzing whether this sort of text can then be used for content or discourse analysis to examine how and in what context certain topics are discussed. We strongly encourage fellow researchers to exploit this novel data source.

Finally, it is important to underscore that this research is highly explorative. While web mining and topic modeling have been used in social science, including regional studies and economic geography before, the combination with regional policy data from German municipalities and counties offers a fresh perspective. We believe that the methods that we have proposed have the potential to be a valuable tool for understanding regional policy priorities and narratives and thereby providing a new way of analyzing the role these actors play in regional (economic) development. We hope that that the opportunities arising from this unique blend of data and methodology will encourage other researchers to use web mining to explore this area.

Literature

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D. & Resch, B. (2023a). The digital layer: alternative data for regional and innovation studies. *Spatial Economic Analysis*, 1–23. https://doi.org/10.1080/17421772.2023.2193222

Abbasiharofteh, M., Kinne, J., Krüger, M. (2023b). Leveraging the digital layer: the strength of weak and strong ties in bridging geographic and cognitive distances. *Journal of Economic Geography.* Advance online publication. https://doi.org/10.1093/jeg/lbad037

Feeney, M. K. & Brown, A. (2017). Are small cities online? Content, ranking, and variation of U.S. municipal websites. *Government Information Quarterly*, *34*(1), 62–74. https://doi.org/10.1016/j.giq.2016.10.005

Gök, A., Waterworth, A. & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. https://doi.org/10.1007/s11192-014-1434-0

Kinne, J. & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

Kinne, J. & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PloS one*, *16*(4), e0249071. https://doi.org/10.1371/journal.pone.0249071

Kriesch, L. J. (2023). *Web Mining und Natural Language Processing als methodisches Komplement in der Wirtschaftsgeographie* [, Universitätsbibliothek Gießen]. DataCite.

Krüger, M., Kinne, J., Lenz, D. & Resch, B. (2020). The digital layer: How innovative firms relate on the web. *ZEW Discussion Papers*, *20*(03).

Lenz, D. & Winker, P. (2020). Measuring the diffusion of innovations with paragraph vector topic models. *PloS one*, *15*(1), 1-18. https://doi.org/10.1371/journal.pone.0226685

Li, Y., Arora, S., Youtie, J. & Shapira, P. (2018). Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation*, *76-77*, 3–14. https://doi.org/10.1016/j.technovation.2016.01.002

Lund, H. B. & Vildåsen, S. S. (2022). The influence of Industry 4.0 narratives on regional path development. *Regional Studies, Regional Science*, *9*(1), 82–92. https://doi.org/10.1080/21681376.2022.2029552

Meschede, C. (2019). Information dissemination related to the Sustainable Development Goals on German local governmental websites. *Aslib Journal of Information Management*, *71*(3), 440–455. https://doi.org/10.1108/AJIM-08-2018-0195

Meub, Lukas, Proeger, Till, Fuhrich, Svenja, Ullrich, Matthias, Bizer & Kilian. (2023). *Zukunftsfelder für Smart City: Eine Webscraping-Analyse von Betrieben und Organisationen der Landkreise Hildesheim, Peine und der Region Hannover.* ifh Göttingen.

Ozgun, B. & Broekel, T. (2021). The geography of innovation and technology news - An empirical study of the German news media. *Technological Forecasting and Social Change*, *167*, 120692. https://doi.org/10.1016/j.techfore.2021.120692

Ozgun, B. & Broekel, T. (2022). Assessing press releases as a data source for spatial research. *REGION*, *9*(2), 25–44. https://doi.org/10.18335/region.v9i2.379

Rae at al. (2022). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. https://arxiv.org/abs/2112.11446

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics, 8, 842-866.

Sandoval-Almazan, R. & Gil-Garcia, J. R. (2012). Are government internet portals evolving towards more interaction, participation, and collaboration? Revisiting the rhetoric of e-government among municipalities. *Government Information Quarterly*, *29*, S72-S81. https://doi.org/10.1016/j.giq.2011.09.004

Stich, C., Tranos, E. & Nathan, M. (2023). Modeling clusters from the ground up: A web data approach. *Environment and Planning B: Urban Analytics and City Science*, *50*(1), 244–267. https://doi.org/10.1177/23998083221108185

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M. & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. ArXiv:2209.11055. https://doi.org/10.48550/arXiv.2209.11055.

Youngblood, N. E. & Mackiewicz, J. (2012). A usability analysis of municipal government website home pages in Alabama. *Government Information Quarterly*, *29*(4), 582–588. https://doi.org/10.1016/j.giq.2011.12.010

Yun, J. & Geum, Y. (2020). Automated classification of patents: A topic modeling approach. *Computers & Industrial Engineering*, *147*, 106636. https://doi.org/10.1016/j.cie.2020.106636