BOOK REVIEW

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press

https://doi.org/10.17356/ieejsp.v10i4.1410

Over the past decade, methods for analysing text as data have gained significant prominence in several fields, such as health, business, industry, but also in the social sciences and humanities. The increasing availability of large datasets, the development of advanced analytical techniques, and the decreasing costs of computational resources have collectively enhanced the capacity to extract valuable insights from textual data.

Authored by Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart, *Text as Data* is a valuable resource for researchers looking to leverage text as data in the social sciences, digital humanities and other fields where language is key to understanding human behaviour. Justin Grimmer is associate professor of political science at Stanford University. In the field of political science, his research has primarily focused on legislative communication and its influence on political representation. Margaret (Molly) Roberts's research focuses on the intersection of political methodology and information politics, with particular emphasis on automated content analysis and censorship in China. Her current work encompasses a range of projects, including investigations into censorship, propaganda, topic modelling, and advanced methods of text analysis. Brandon Stewart is an Associate Professor in the Department of Sociology at Princeton University. His research focuses on the development of novel quantitative statistical methods for application across the social sciences.

Each author possesses extensive expertise in data science, with a substantial background in the field. They have produced leading work in the field of data science that spans a diverse range of subjects within both American and international politics (Grimmer, 2010; Grimmer & Stewart, 2013; Roberts et al., 2014). Through the application of computational methods, they have produced and validated numerous novel insights into the nature, causes and consequences of political communication. Their work exemplifies the potential of treating text as data to substantially advance various domains of social science research.

In recent years, the cost of analysing vast collections of texts has undergone a dramatic reduction leading to remarkable advancements across various disciplines. Social scientists, digital humanities scholars and industry professionals now regularly leverage

large-scale document corpora. This shift is attributable, in part, to technological developments, but additionally, methodological innovations have played a crucial role. A growing body of literature – originating in computer science and computational linguistics, and subsequently expanding into social sciences and digital humanities – has introduced tools, models and software that facilitate large-scale text analysis and organization. The origins of this development trace back to the late 1980s and early 1990s, when first statistical techniques, such as latent semantic indexing, emerged, enabling the complex analysis of larger text corpora (Miner et. al., 2012). However, the widespread adoption of these methods occurred primarily after 2000, as newer techniques for processing the rapidly expanding volume of digital content were introduced (Liu, 2015).

According to the 'Text as Data' approach, textual data is treated as an organized, structured file, formatted within a numerical database to serve as input for computational algorithms (Gentzkow et al., 2019). These algorithms either first developed in computer science or built closely on those developments. For instance, within political science, scholars have employed topic models (Blei et al., 2003; Grimmer, 2010; Roberts et al., 2013) as well as supervised learning algorithms for document classification (Stewart & Zhukov, 2009; Pan & Chen, 2018). However, according to the authors, the knowledge transfer from computer science and related fields has created confusion in how text as data models are applied, how they are validated, and how their output is interpreted. This confusion emerges because tasks in academic computer science are different than the tasks in social science, the digital humanities, and even parts of industry. While computer scientists are often (but not exclusively!) 'interested in information retrieval, recommendation systems, and benchmark linguistic tasks, a different community is interested in using "text as data" to learn about previously studied phenomena such as in social science, literature, and history' (p. 23).

A large dataset of texts alone is insufficient; it is essential to formulate relevant research questions and derive meaningful answers. Additionally, one must demonstrate the limitations of the data's validity and account for potential sample biases and distortions. This book aims to illustrate how to treat 'text as data' for *social science tasks* and *social science problems*. It adopts a six-part structure, combined with several chapters and subchapters. Each part is structured around five fundamental concepts: representation, discovery, measurement, prediction and causal inference. Part I (Chapters 1–2) presents a comprehensive overview of how computational methods, particularly text analysis, are transforming social science research. It emphasizes the shift from traditional deductive approaches to more iterative and data-driven processes, where researchers engage in cycles of discovery, measurement, and inference. The part outlines the stages of the research process – discovery, measurement, and inference – explaining how text analysis can contribute at each stage. It also emphasizes that while computational methods are powerful, they cannot replace human judgment. Instead, these methods augment human capacity by helping researchers uncover insights that can be interpreted through theoretical frameworks.

Chapter 2 is built around six core principles of text analysis, which guide the application of computational methods to social science research. These principles underscore the critical role of theory, human expertise, and iterative processes in social science research. They highlight the complementary function of computational text analysis, which enhances rather than supplants human judgment. By integrating computational tools into

the research process, scholars are able to analyse large textual datasets more efficiently while still relying on theoretical frameworks and expert interpretation to guide their analysis and draw meaningful conclusions.

Part II (Chapters 3–9) delves into the fundamental phase of the research process: 'Selection and Representation.' Chapter 3 establishes key principles for selecting and representing texts as data, providing the theoretical grounding for subsequent discussions. Chapter 4 is particularly significant as it emphasizes the importance of corpus selection, framing it as a fundamental step that can shape the outcome of research. This chapter addresses four prevalent forms of sample selection bias – resource bias, incentive bias, medium bias, and retrieved bias – that can impact the representativeness of the textual corpus. These biases impact which voices or texts are included, often reflecting power structures or accessibility limitations. Chapters 5 through 7 explore various approaches to word representation, beginning with the traditional 'bag of words' model (Chapter 5), followed by the probabilistic 'multinomial language model' (Chapter 6), and the algorithmic 'vector space model' (Chapter 7). Since these models do not account for the contextual relationships and semantic similarities between words, Chapter 8 introduces the 'word embeddings' model, which provides distributed representations that capture such nuances. Finally, Chapter 9 reviews additional methods of text representation, focusing on language sequences that encode both the syntactic and semantic roles of words, illustrated through a range of examples.

Part III (10–14) delineates the initial core task within the social research process, termed 'Discovery,' a concept that is often overlooked in quantitative research methodologies. This section elucidates how models that treat 'text as data' can enhance both conceptual discovery and theoretical innovation in the social sciences. The discussion begins with an introduction to the concept of 'discovery' and the presentation of four foundational principles that should guide this process: context relevance, the absence of a definitive ground truth, the importance of evaluating the concept rather than the method, and the advantages of utilizing separate data sources (Chapter 10). These principles establish a framework for researchers to navigate the complexities inherent in textual analysis. Following this foundational chapter, Chapter 11 introduces keyword analysis methods, which focus on identifying discriminating words. These methods are characterized as relatively simple yet powerful techniques for uncovering the distinctive features of documents, facilitating deeper insights into their content and context. Subsequent chapters (12–14) expand upon the theme of organizational discovery within texts, presenting three distinct methodologies for text categorization. Chapter 12 discusses unsupervised clustering analysis, which groups texts based on inherent patterns without pre-defined labels. Chapter 13 introduces mixed membership topic models, which allow for a nuanced understanding of the various topics present within a single document, thereby acknowledging the complexity of textual data. Finally, Chapter 14 explores low-dimensional document embeddings, a technique that represents documents in a condensed vector space, enabling efficient comparisons and analyses across large corpora.

Part IV of the text (Chapters 15–21) focuses on the second core task in social science research: 'Measurement,' which involves quantifying conceptualizations that have been discovered. The section begins by outlining five foundational principles for effective

measurement: clear goals, identifiable source material, an explainable and reproducible coding process, validated measures, and documented limitations (Chapter 15). Following this, Chapter 16 introduces the simplest measurement method – 'word counting.' Chapter 17 provides an overview of supervised classification methods, while Chapters 18–20 delve into specific components of this approach. Chapter 18 addresses coding a training set, Chapter 19 discusses classifying documents with supervised learning, and Chapter 20 evaluates the performance of classification models. The final chapter (Chapter 21) emphasizes the application of discovery methods to measurement and the critical importance of extensive validation in both supervised and unsupervised models. This validation is essential for confirming the accuracy and reliability of measurement results, thereby enhancing the credibility of research findings. Overall, Part IV establishes rigorous standards for measurement in social science research, ensuring clarity, reproducibility, and validation in the quantification of concepts.

Part V of the text (Chapters 22–27) addresses the final stage of the social research process – 'Inference,' which encompasses prediction and causal inference. Chapter 22 introduces general principles of inference and differentiates the concept of 'prediction' from 'causal inference.' Chapter 23 outlines four major types of predictive tasks: source prediction, linguistic prediction, social prediction, and nowcasting. While prediction has become increasingly important in the social sciences, many research projects ultimately aim for causal inference. Chapter 24 provides an in-depth examination of causal inference, clarifying its relationship with prediction and measurement, and establishing key principles for applying causal inference in textual analysis. Chapters 25–27 further explore the application of text data in experimental settings, detailing its use as an outcome, a treatment, and a confounder. Together, these chapters underscore the importance of inference in social research, highlighting the methodologies that enable researchers to derive meaningful conclusions from textual data.

Part VI (Chapter 28) is the concluding part of this book. It reaffirms the principles articulated in earlier chapters, which collectively validate text as data methods as a robust tool for advancing research within the social sciences. However, the chapter emphasizes that, despite the apparent power and promise of these methodologies, they cannot wholly supplant human analytical capabilities or address the fundamental challenges inherent in social science research. Issues such as threats to inference – including confounding variables, reverse causality, and dependence – persist regardless of the analytical techniques employed. Consequently, the application of text as data methods must be approached with careful consideration and a sense of modesty, recognizing both their potential and limitations in the pursuit of meaningful social science inquiry.

In articulating my personal experience with the book, I find it most fitting to commence with the title of one of its subsections. In the introductory subsection titled 'What This Book Is Not,' the authors clarify the primary focus of the book, emphasizing that it is centred on research design rather than the technical intricacies of contemporary methods or software applications. The authors explicitly delineate their objective: to utilize textual analysis as a means of addressing questions pertaining to social data. This approach may leave readers seeking immediate, practical coding examples or a comprehensive examination of specific technical methodologies feeling dissatisfied. By establishing this frame-

work, the authors set a clear expectation for the reader, directing attention towards the conceptual underpinnings of research design in text analysis rather than providing a technical manual.

However, this position is, I believe, correct and justifiable. There are many manuals of a technical nature. But the integration of the 'text as data' approach in social research, the place of new methods among the existing ones are rarely discussed. Rather than technological innovations, the book discusses the challenges of using computational methods that are integrated into existing paradigms of social research: the critical role of theory and human expertise, the complementary function of computational text analysis, distinction of prediction and causal inference etc. That is, the book looks at the technological innovation in IT from the perspectives of social research. As the authors write: 'a central argument of this book is that the goal of text as data research differs from the goals of computer science work' (p. 5). In this respect, this work is unique and fills an important gap.

Overall, Text as Data emphasizes the treatment of 'text as data' within the context of social science tasks and problems, providing illustrative examples throughout. Additionally, it serves as a comprehensive guide for researchers, delineating the capabilities and limitations inherent in text data methodologies. The text facilitates readers' familiarity with the diverse range of tasks that text data methods can effectively address, thereby enhancing their understanding of the potential applications in social science research.

Tamás Varga
[vtamas7@student.elte.hu]
(Eötvös Loránd University)

References

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

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18*(1), 1–35. https://doi.org/10.1093/pan/mpp034

Grimmer, J. & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Gentzkow, M., Shapiro, J. M. & Taddy, M. (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica, 87*(4), 1307–1340. https://doi.org/10.3982/ecta16566

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D. & Fast, A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Elsevier.

Pan, J. & Chen. K. (2018). Concealing corruption: How Chinese officials distort upward reporting of online grievances. *The American Political Science Review, 112*(3), 602–620. https://doi.org/10.1017/s0003055418000205

Roberts, M. E., Stewart, B. M., Tingley, D. & Airoldi, E.M. (2013). *The Structural Topic Model and Applied Social Science.* Conference Paper (Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation). https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/stmnips2013.pdf

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Sciences, 58*(4), 1064–1082.

Stewart, B. M. & Zhukov, Y. M. (2009). Use of force and civil–military relations in Russia: an automated content analysis. *Small Wars & Insurgencies, 20*(2), 319–343. https://doi.org/10.1080/09592310902975455