Dissertation Paper I | Draft

Working title: **Mapping and Evaluating Measurement Validity Approaches for Text-as-Data Methods**

Lukas Birkenmaier  
MA Politik- und Verwaltungswissenschaften (University of Konstanz) &   
MSc Big Data and Digital Futures (University of Warwick)

***Coauthors:***

1. Dr. Clemens Lechner

2. Prof. Dr. Claudia Wagner

Abstract

Validating a measurement instrument in social science research is not a trivial task. This is especially the case if the instrument relies on text as data (TADA), because the constructs under study are inherently unobservable,

To describe and cluster current the validity of this emerging field of research, we conduct a systematic review of validation practices in social-scientific studies that rely on computer-assisted text methods (CATM). Our results show that XXX. Ultimately, we provide hints for further research as well as XX:

**Contents**

[1. Introduction 3](#_Toc108187412)

[The rise of digital behavioral data in political communication 3](#_Toc108187413)

[The crucial Task of Validating TaDa methods 3](#_Toc108187414)

[Aims 4](#_Toc108187415)

[2. Methods 5](#_Toc108187416)

[Systematic Review 5](#_Toc108187417)

[Literature Search 5](#_Toc108187418)

[Expert Interviews 7](#_Toc108187419)

[3. Analysis & Results 8](#_Toc108187420)

[4. Conclusion 8](#_Toc108187421)

[5. Appendix 10](#_Toc108187422)

[Appendix 1: Keywords 10](#_Toc108187423)

[Appendix 2: Benchmark Literature 12](#_Toc108187424)

[Appendix 4: Questionnaire 13](#_Toc108187425)

[6. TOPIC 0 | Summary 14](#_Toc108187426)

# Introduction

## Text as Data methods in political communication

In the past decade, social science research has experienced a remarkable rise in publications relying on digital behavioural data and computational methods (Brady, 2019; Edelmann et al., 2020). Thereby, the analysis of texts using computational text analysis methods (CATM) holds particular promise, as textual data enable researchers to facilitate substantively important inferences of human behaviour on an unprecedented scale (Grimmer & Stewart, 2013; Lazer et al., 2009). Because words and language constitute the most common way for people to express their internal thoughts and emotions, they lie at the core of psychology and communication and language research (Tausczik & Pennebaker, 2010). Those premises have been especially well received in the field of political communication research, where the number of studies relying on CATM has exploded in recent years. However, the the applications of these methods is currently debated. In particular, there is an ongoing discussion about the measurement validity (Baden et al., 2021). Measurement validity is

“with significant post hoc validation costs, as the researcher “must combine experimental, substantive, and statistical evidence to demonstrate that the measures are as conceptually valid as measures from an equivalent supervised model” (Grimmer and Stewart 2013, 5).”

## Validity and Text as Data Methods

However, a growing body of research highlights several methodological challenges in the automated analysis of textual data (Baden et al., 2021; Howison et al., 2011; van Atteveldt et al., 2021). One of the key challenges relates to the validity of CATM-based measures. Several scholars have expressed doubts about whether CATM can operationalize and validly measure the constructs under study, thus questioning these methods’ ability to answer substantive theory-driven research questions (Adcock & Collier, 2001).

Each method for analyzing textual content imposes its own particular set of assumptions and, as a result, has particular advantages and weaknesses for any given question or set of texts.(Quinn)

This is especially relevant because the high dimensionality and complex structure of texts renders it almost impossible to *truly* understand the underlying dependencies of texts (Yeomans, 2021).[[1]](#footnote-2)[[2]](#footnote-3) This culminates in the problem that even hand-coded estimates (“gold standard”) for the same texts are often [not consistent](https://www.linguee.de/englisch-deutsch/uebersetzung/not+consistent.html) so that reliability measures can vary significantly between different coders (cf. Song et al., 2020). Therefore, authors such as Baden et al. (2021) have called for more unifying validation efforts to show “exactly how and how convincingly [CATM] operationalize relevant conceptual properties” (Baden et al., 2021, p. 14). When reviewing the literature, one can observe a great heterogeneity of validation approaches in studies using CATM developed over time. Among others, these approaches include the comparison of estimates with other measurements and external data (Bach et al., 2021; Kmetty & Németh, 2022; Röttger et al., 2021), or the development of multistage validation frameworks, which cover different dimensions of validity at the same time (Goet, 2019; Lowe & Benoit, 2013). Even more worrying, scholars who want to validate their measures can easily lose track of what options they have, and which validation approaches are necessary to conduct high-quality research projects.

## Aims

To provide guidance in CATM research, the aim of the current study is to collect, structure, and describe the various validation approaches in CATM research. Thus, it is asked: How are issues of measurement validity addressed in CATM studies? What are the main validation strategies in CATM research? And how can the current approaches be structured into overarching categories? To answer these research questions, this study provides a two-fold approach. First, a systematic review of validation practices from publications in the field of political communication. And second, focus group interviews with CATM experts. Subsequently, both approaches will be presented and discussed in more detail.

# Methods

## Systematic Review

To conduct the review of validation practices for CATM methods, we apply a systematic framework to identify a representative and unbiased sample of relevant studies within the field of political communication (Durlak & Lipsey, 1991; Liberati et al., 2009). To search for relevant studies, we manually searched the major journals in the field of political and communication science using the scimago journal ranking and screening relevant publications. These are *Communication Methods and Measures,* [*Digital Journalism*](https://www.scimagojr.com/journalsearch.php?q=21100430187&tip=sid&clean=0)*,* [*Political Communication*](https://www.scimagojr.com/journalsearch.php?q=22838&tip=sid&clean=0)*,* [*Applied Linguistics*](https://www.scimagojr.com/journalsearch.php?q=18651&tip=sid&clean=0)*,* [*Research on Language and Social Interaction*](https://www.scimagojr.com/journalsearch.php?q=22340&tip=sid&clean=0)*, Journal of Communication, Public Opinion Quarterly, Administrative Science Quarterly, American Sociological Review, American Political Science Review, Annual Review of Political Science, Annual Review of Sociology, American Journal of Political Science, Leadership Quarterly, Political Analysis* and *Party Politics*. We then searched these journals for publication title, abstract and keywords using multiple groups of keywords in the field of text analysis and political communication. After an initial search round using a minimal search string ((part\* or politic\* or \*govern\*) and (text\* or corpus)), the preliminary literature list was optimized using the LitsearchR package to facilitate quasi-automatic search strategy development. The optimized search string ultimately contained several keywords relating to both CATM (e.g., "*autom\* text\* analysi*\* or *corpus* or *topic model*) and the field of political communication (e.g., “*plenar\* debat*\* or *news cover*\* or *elect*\*). Because we want to review only the most recent validation practices, we limited our search results to studies between January 2015 to May 2022. This limited our literature list to 917 studies.

In a second step, title, abstract and keywords of all 917 publications were screened to evaluate their eligibility using the tool a*bstrackr* (Wallace et al., 2012). To be included into the final sample, studies must apply a computer-assisted text method using data from the domain of political communication. Thus, the method needs to needs to operationalize a typical latent construct, such as rhetoric or word tone, in the content of an empirical and theory-driven application.

**Methods:** The review will include publications which rely on computational methods using text as data. In particular, these might entail rule-based (e.g., dictionaries), unsupervised (text scaling algorithms, topic modelling), and supervised (machine learning) methods to operationalize and measure variables of interest for political and communicational research (Grimmer & Stewart, 2013). Generally, previous studies have pointed to the fact that validation exercises vary greatly between groups of methods.[[3]](#footnote-4) For instance, an unsupervised text scaling model like *Wordfish* (Slapin & Proksch, 2008) requires specific assumptions about the ideological composition of textual data (but no labelled training data), whereas a deep-learning ML model like BERT (Devlin et al., 2019) primarily requires sufficient labelled training data while being characterized as a “black box”. That being said, a systematic description and comparison of validation exercises across groups of methods might be an important first insight from this systematic review.

**Publication Types**. The review will include peer-reviewed publications from the most relevant journals in the field of communication and political science. The journals are identified using the Scimago Journal Ranking.[[4]](#footnote-5) Additionally, the choice of journals will be evaluated by relying on expert in the field. The lists of journals includes the following journals:

**Time and Language.** Some additional criteria are applied:

* Only publications published after 2015
* Only publications in the English language

Once the final sample of publications were collected, publications are coded based on a coder manual (Durlak & Lipsey, 1991). Optimally, at least two coders will be involved, so that metrics of interrater agreement for a sample of publications can be reported (Gisev et al., 2013). However, the coding of validation efforts might be too difficult for a research assistant with no previous knowledge on text analysis. Therefore, a solution could be to code a subset of studies (e.g., 20-30 studies) and proceed to independent coding after sufficient intercoder reliability is achieved (see Landis & Koch, (1977))  
*Coding*

* usually above 80 percent agreement was termed acceptable (Belur et al., 2021)

## Expert Interviews

To get an even more profound understanding of validation practices, we will also conduct interviews with experts in the field of CATM research. The interviews will be carried out at the [name] conference using open ended questions related to the practices and challenges of CATM validation. Due to the exploratory nature of this study, experts will be selected based on their reputation and publication record in the field of political communication and CATM application. In total, we aim to conduct around eight interviews, with participation being voluntarily and without compensation.

Below, the main questions of the questionnaire are reported. The draft for the full questionnaire can be found in Appendix 4:

* What is your personal and professional background?
* What is validity to you? Have you come across different validity conceptualizations within different research fields?
* What approaches do you know to validate a measurement method?
* How would you describe the current state of knowledge on computer-assisted text method validation?
* What approaches do you know of to validate computer-assisted text methods?
* What do you see as barriers to sufficiently validating CATM?
* How would you describe validation practices of CATM, which are usually not reported in papers?

Ultimately, the interviews will be transcripted and analysed using content analysis (Krippendorff, 2018).

# Analysis & Results

# Conclusion

* No standards for Cutoff values for correlations (convergent/criterion validity)
* No universal guidelines for feature importance (Jankowski & Huber, 2022)

**Mapping and Evaluating Measurement Validity Approaches for Text-as-Data Methods**

**Problem Statement.** In the past decade, social science research has experienced a remarkable rise in publications relying on digital behavioral data and computational methods (Brady, 2019; Edelmann et al., 2020). Thereby, the analysis of texts using computational text analysis methods (CTAM) holds particular promise, as textual data enable researchers to facilitate substantively important inferences of human behavior on an unprecedented scale (Grimmer & Stewart, 2013; Lazer et al., 2009). However, a growing body of research highlights several methodological challenges in the automated analysis of textual data (Baden et al., 2021; Howison et al., 2011; van Atteveldt et al., 2021). One of the key challenges relates to the validity of CTAM-based measures. Several scholars have expressed doubts whether CTAM methods can operationalize and validly measure the constructs under study, thus questioning these methods’ ability to answer substantive theory-driven research questions. This is especially relevant because the high dimensionality and complex structure of texts renders it almost impossible to *truly* understand the underlying dependencies of texts (Yeomans, 2021).[[5]](#footnote-6)[[6]](#footnote-7) This culminates in the problem that even hand-coded estimates (“gold standard”) for the same texts are often [not consistent](https://www.linguee.de/englisch-deutsch/uebersetzung/not+consistent.html) so that reliability measures can vary significantly between different coders (cf. Song et al., 2020). Therefore, authors such as Baden et al. (2021) have called for more unifying validation efforts to show “exactly how and how convincingly [CTAM] operationalize relevant conceptual properties” (Baden et al., 2021, p. 14). When reviewing the literature, one can observe a great heterogeneity of validation approaches in studies using CTAM developed over time. Among others, these approaches include the comparison of estimates with other measurements and external data (Bach et al., 2021; Kmetty & Németh, 2022; Röttger et al., 2021), or the development of multistage validation frameworks, which cover different dimensions of validity at the same time (Goet, 2019; Lowe & Benoit, 2013). Even more worrying, scholars who want to validate their measures can easily lose track of what options they have, and which validation approaches are necessary to conduct high-quality research projects.

**Approach.** To provide guidance in CTAM research, this first paper aims to collect, structure, and describe the various validation approaches in CTAM research. Thus, it is asked: How are issues of measurement validity addressed in CTAM studies? What are the main validation approaches in CTAM research? And how can the current approaches be structured into overarching categories? To address the research questions presented, I intend to conduct a systematic review on validation practices across studies in the field of CTAM research. Thereby, I want to review studies that apply lexicon-based methods (dictionaries) in the field of political science research as a target population. Dictionaries are one of the most popular CTAM method in social science research, which can be attributed to their transparency and simplicity (Chan et al., 2021; van Atteveldt et al., 2021). Moreover, dictionaries are widely used across disciplines, even though several studies show that the performance and trustworthiness of dictionaries can vary greatly (e.g., Boukes et al., 2020; González-Bailón & Paltoglou, 2015; Soroka et al., 2015). Therefore, dictionaries provide a well-suited field of research to explore and review relevant approaches to ensure measurement validity.[[7]](#footnote-8) To conduct the review, I want to apply a systematic framework (Durlak & Lipsey, 1991; Liberati et al., 2009)(Durlak & Lipsey, 1991; Liberati et al., 2009) to objectively search for relevant studies. This entails the definition of multiple groups of keywords related to the method (dictionary-based methods for the analysis of textual data) and the domain (e.g., political science research)[[8]](#footnote-9), the screening and exclusion of studies, and the synthesis and qualitative assessment of the selected studies towards their validation strategies.

**Contribution to the Field of Research.** The results of this study will contribute greatly to a better understanding of validation practices in CATM research. By systematically screening and describing the most relevant validation approaches, this research will acquire and structure knowledge on validation approaches within the rapidly evolving field of text-as-data research. Taking a broader perspective, this paper will therefore contribute to the ongoing discussion of measurement validity in digital behavioural data research. Ultimately, the analysis will provide a starting point for the development of theoretical concepts and quality indicators to strengthen the trustworthiness of the academic field and the credibility of its inferences in applied research projects (see paper 2).

# Appendix

## Appendix 1: Keywords

* + Naïve search
    - (Politi\* OR Party OR Govern\*) AND text\*
  + Adjusted search (*LitsearchR*)
    - Topic
      * **Political Communication** 
        + (elect\* OR govern\* OR parti\* OR polici\* OR "social\* media\*" OR polit\*)
      * **Textual Analysis**
        + ("autom\* text\* analysi\*" OR "content\* analysi\*" OR sentiment\* OR discours\* OR languag\* OR "machin\* learn\*" OR text\* OR word\* OR "comput\* communic\* scienc\*" OR Corpus\* OR lexicon\* OR Automa\* Content\* Analy\*)
    - Journals
      * "Communication Methods and Measures" OR "Digital Journalism" OR "Political Communication" OR "Applied Linguistics" OR "Research on Language and Social Interaction" OR "Journal of Communication" OR "Public Opinion Quartely"
      * “[Administrative Science Quarterly](https://www.scimagojr.com/journalsearch.php?q=16036&tip=sid&clean=0)” OR “[American Sociological Review](https://www.scimagojr.com/journalsearch.php?q=16929&tip=sid&clean=0)” OR “[American Political Science Review](https://www.scimagojr.com/journalsearch.php?q=15557&tip=sid&clean=0)” OR “[Annual Review of Political Science](https://www.scimagojr.com/journalsearch.php?q=15646&tip=sid&clean=0)” OR “[Annual Review of Sociology](https://www.scimagojr.com/journalsearch.php?q=16976&tip=sid&clean=0)” OR “[American Journal of Political Science](https://www.scimagojr.com/journalsearch.php?q=15555&tip=sid&clean=0)” OR “[Leadership Quarterly](https://www.scimagojr.com/journalsearch.php?q=21149&tip=sid&clean=0)” OR “[Political Analysis](https://www.scimagojr.com/journalsearch.php?q=145558&tip=sid&clean=0)”

## Appendix 2: Benchmark Literature

* Goet, Niels D. ‘Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015’. *Political Analysis* 27, no. 4 (2019): 518–39.
* Gilardi, Fabrizio, Theresa Gessler, Mael Kubli, and Stefan Muller. ‘Social Media and Political Agenda Setting’. Political Communication 39, no. 1 (2 January 2022): 39–60.
* Di Cocco, J., & Monechi, B. (2022). How Populist are Parties? Measuring Degrees of Populism in Party Manifestos Using Supervised Machine Learning. Political Analysis, 30(3), 311-327. doi:10.1017/pan.2021.29
* Røed, Maiken. "When do political parties listen to interest groups?." Party Politics (2022): 13540688211062832.
* Welbers, K., van Atteveldt, W., Bajjalieh, J., Shalmon, D., Joshi, P. V., Althaus, S., Chan, C.-., Wessler, H. & Jungblut, M. (2022). [Linking event archives to news: A computational method for analyzing the gatekeeping process](https://madoc.bib.uni-mannheim.de/61633/1/Linking%20event%20archives%20to%20news%20a%20computational%20method%20for%20analyzing%20the%20gatekeeping%20process.pdf). *Communication Methods and Measures, 16*(1), 59–78. https://doi.org/10.1080/19312458.2021.1953455

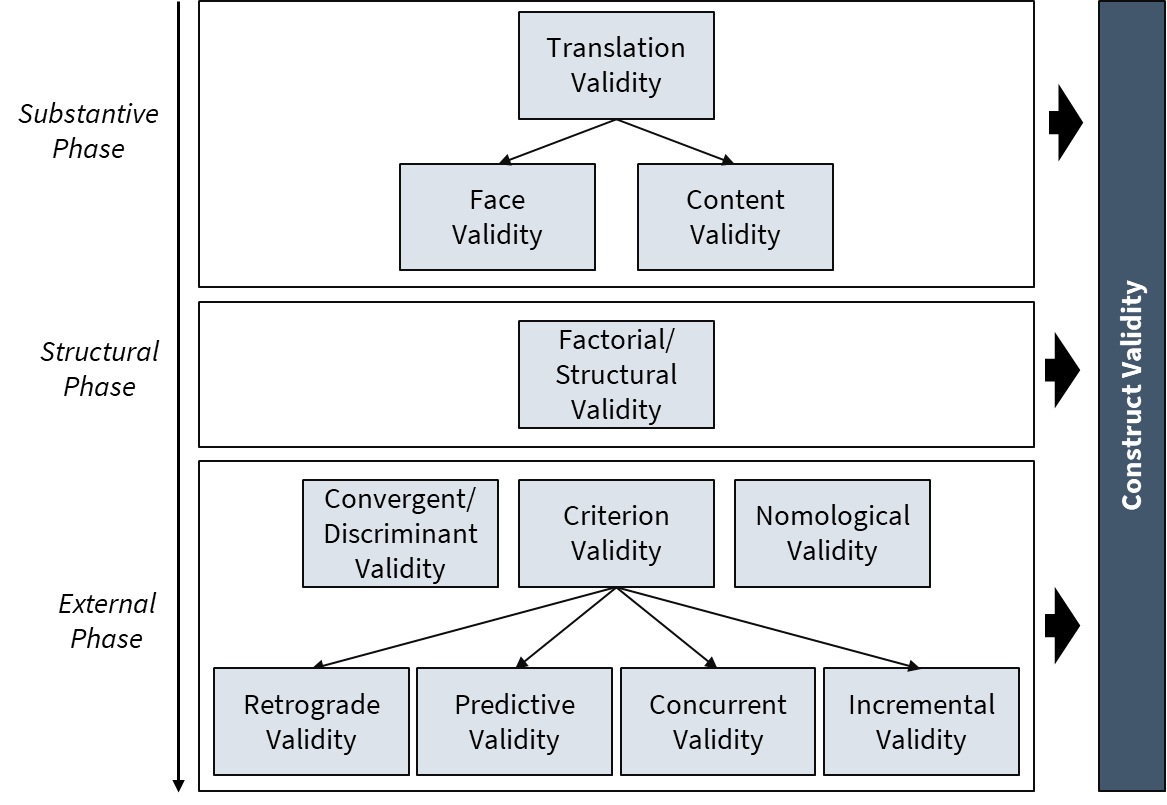
Appendix 3: Overview Validity Concepts

Figure 1: Own visual depiction of validity concepts

## Appendix 4: Questionnaire

|  |  |
| --- | --- |
| Background Paper |  |
|  |  |
| Consent | Furthermore, I would like to inform you that I will record the prepared questions and your answers for my bachelor thesis. This recording will be transcribed to be able to analyze the answers. The recorded data is purely for research purposes Your data will only be presented in aggregated form and anonymized. The research paper, including the appendices, can only be viewed by members of the Department of Geography at Heidelberg University, including myself.  Do you agree that I record this interview for research purposes? |

|  |  |
| --- | --- |
| Introduction | What is your professional background? |
| How is your work related to text analysis and political communication? |
| Measurement Validity | How would you **describe validity**? |
| Have you come across **different validity terms** within different **research fields**? |
| Measurement validity and computer-assisted text methods? | Have you come across **different validity terms** in the field of computer-assisted text methods? |
| How would you describe the process of measurement validation? Which steps are usually reported in the final paper? |
| **What approaches** do you know to validate a computer-assisted text method? |
| How would you **describe** the **current state of knowledge** on validating computer-assisted text method? |
| Challenges Measurement Validity and CATM | What are the barriers that **hamper** the validation of computer-assisted text methods? (*Costs, missing guidelines etc*.) |
| Do you have any ideas on **how to improve** the validation of computer-assisted text methods? |
| Conclusion | Is there anything else you would like to tell us? |

# TOPIC 0 | Summary

**Enhancing the Measurement Validity of Digital Behavioural Data**[[9]](#footnote-10)

**Introduction.** In the past 20 years, social science research has experienced a remarkable rise in publications relying on digital behavioral data and computational methods (Brady, 2019; Edelmann et al., 2020). The potential advantages of these developments are impressive. They allow researchers not only to get access to new kinds of previously unexplored data, but also to study what people think, feel, and do, in near real-time and large scale, by analyzing and combining digital traces of human behaviour (Azucar et al., 2018; Eichstaedt et al., 2021). However, a growing body of research also highlights profound challenges in the validation of DBD-based methods (Baden et al., 2021; Howison et al., 2011; van Atteveldt et al., 2021). One of the most crucial questions is whether methods relying on DBD can operationalize and validly measure the underlying constructs of interest in social science research. Keeping in mind that focal constructs in social-science research are often unobservable and multidimensional, operationalizations that lack validity could seriously hamper these methods’ ability to answer substantive theory-driven research questions because what exactly is being measured is unknown (Flake et al., 2017). This is especially relevant in the context of textual data because the high dimensionality and complex structure of texts render it almost impossible to *truly* understand the underlying dependencies in written language (Song et al., 2020; Yeomans, 2021).

**Dissertation Outline.** Emphasizing the challenges associated with validation practices in DBD research, this dissertation will describe, structure, and advance current validation practices in the field of DBD research. In the first paper, I will conduct a systematic review of validation practices in social-scientific studies that rely on computer assisted text methods (CATM). The aim is to provide an overview over how current studies approach the issue of validity and that, if any, evidence for the validity of their measures they present. , Toward that end, I will identify and structure the most relevant validation practices for a sample of studies that used dictionaries in the domain of political science research. This study will set the groundwork for approaching the issue of measurement validation.

In the second paper, I will build upon the validation practices identified in this systematic review and develop a validation framework, which will synthesize and combine current approaches in the field of DBD research. Additionally, the framework will benefit from a discussion of best practice validation approaches in other domains, such as psychometrics Adcock & Collier, 2001). The third and fourth paper will then conduct empirical research in the field of DBD that showcase measurement validation practices while also addressing substantive questions. In the third paper, I will explore determinants of political polarization and partisan language using textual data. Even though the exact research question is still open for discussion, the paper will include a thorough validation of the measurement using the validation framework developed in paper 1 and 2. In the fourth paper, I want to extent the scope of analysis by assessing the validity of methods using other forms of DBD. Among others, this might include sensor, image, or video data to answer substantive research questions.

**Expected findings.** Results from my dissertation will not only contribute to a better understanding of current validation practices for DBD methods but also help advance the standars of validity and validation practices in the field. Starting from a discussion of validation practices and their associated methodological and theoretical challenges (*descriptive part*), the dissertation will then provide a general validation framework and hands-on guidance on how to convincingly validate DBD based measures in applied research projects (*normative part)*.

Literature

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, *95*(3), 529–546.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/J.PAID.2017.12.018

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting voting behavior using digital trace data. *Social Science Computer Review*, *39*(5), 862–883.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2021). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 1–18. https://doi.org/10.1080/19312458.2021.2015574

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. *Sociological Methods & Research*, *50*(2), 837–865. https://doi.org/10.1177/0049124118799372

Brady, H. E. (2019). The Challenge of Big Data and Data Science. *Annual Review of Political Science*, *22*(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

Chan, C., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., Atteveldt, W. van, & Jungblut, M. (2021). Four best practices for measuring news sentiment using ‘off-the-shelf’ dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, *3*(1), 1–27.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Durlak, J. A., & Lipsey, M. W. (1991). A practitioner’s guide to meta-analysis. *American Journal of Community Psychology*, *19*(3), 291–332.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology*, *46*(1), 61–81. https://doi.org/10.1146/annurev-soc-121919-054621

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., & Buffone, A. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330–338.

Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, *27*(4), 518–539.

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

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, *12*(12), 2.

Jankowski, M., & Huber, R. A. (2022). *When Correlation is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models*. OSF Preprints. https://doi.org/10.31219/osf.io/uynhb

Kmetty, Z., & Németh, R. (2022). Which is your favorite music genre? A validity comparison of Facebook data and survey data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 07591063211061754. https://doi.org/10.1177/07591063211061754

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., & Gutmann, M. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine*, *6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*(3), 298–313.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15*(2), 121–140.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 819–824.

Yeomans, M. (2021). A concrete example of construct construction in natural language. *Organizational Behavior and Human Decision Processes*, *162*, 81–94. https://doi.org/10.1016/j.obhdp.2020.10.008

1. This has been described as the first principle of textual data analysis by Grimmer and Stewart (2013), namely “all quantitative models of language are wrong -but some are useful). Thus, “the complexity of language implies that all methods necessarily fail to provide an accurate account of the data-generating process used to produce texts” (p. 270) [↑](#footnote-ref-2)
2. “The quantities we seek to estimate from text [...] are fundamentally unobservable” (Lowe & Benoit, 2013, p. 299), [↑](#footnote-ref-3)
3. "Yet only the development of rule-based approaches is sometimes accompanied by similar efforts at concept-driven operationalization and the development of substantive validity criteria (e.g., in the construction of dictionaries, word banks and rule sets). For supervised applications, validation efforts remain largely limited to the manual creation of training data sets, and are rarely reported (Baden et al., 2021)" [↑](#footnote-ref-4)
4. [https://www.scimagojr.com/](https://www.scimagojr.com/journalrank.php?area=330) [↑](#footnote-ref-5)
5. This has been described as the first principle of textual data analysis by Grimmer and Stewart (2013), namely “all quantitative models of language are wrong -but some are useful). Thus, “the complexity of language implies that all methods necessarily fail to provide an accurate account of the data-generating process used to produce texts” (p. 270) [↑](#footnote-ref-6)
6. “The quantities we seek to estimate from text [...] are fundamentally unobservable” (Lowe & Benoit, 2013, p. 299), [↑](#footnote-ref-7)
7. Additionally, the restriction to dictionaries might help to focus on the different ideas and approaches to validate on the measures, instead of being distracted by a variety of different model specifications, such as for machine learning methods (e.g., SMV, BERT etc.) [↑](#footnote-ref-8)
8. Cf. Appendix 1 and 2 for an preliminary naïve and advanced search string. The R-package *litsearchr* is used to optimise the search strategy. The R-Code for the analysis is available under https://github.com/lukasbirki/Review\_Validity [↑](#footnote-ref-9)
9. Note: In a future version of the expose, I am planning to link the title to a specific area of (political) social science research, such as party contestation or polarisation in political discourse. However, I have not yet decided on a specific subfield yet. [↑](#footnote-ref-10)