

**Ensuring the Validity of Computational Text-Based
Measures of Social Constructs:
A Systematic Review, Validation Framework, and Applications**

Von der Philosophischen Fakultät der Rheinisch-Westfälischen Technischen Hochschule
Aachen zur Erlangung des akademischen Grades eines Doktors der Philosophie
genehmigte Dissertation

vorgelegt von

Lukas Roman Birkenmaier

Berichter: Universitätsprofessorin Dr. Claudia Wagner
 Universitätsprofessor Dr. Sebastian Stier

Tag der mündlichen Prüfung: 23.04.2026

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Table of Contents

Acknowledgements	6
List of Figures	8
List of Tables.....	10
1. Introduction	11
1.1 Problem Statement.....	11
1.1.1 Growing Application of Computational Text Analysis in Research	11
1.1.2 Empirical Measures of Social Constructs need to be Validated.....	13
1.1.3 CTAM Measures of Social Constructs are Challenging to Validate.....	15
1.2 The present dissertation	26
1.2.1 Research objectives	26
1.2.2 Contributions and Outline	26
1.2.3 Database.....	28
1.2.4 Publications	29
2. Understanding Contemporary Practices in CTAM Validation.....	31
2.1 Introduction	31
2.2 Theoretical Background.....	33
2.2.1 Promises and Pitfalls of Text as Data in Communication Research	33
2.2.2 Contemporary Approaches to CTAM Validation	35
2.2.3 On the Need for a more Comprehensive Understanding of Validation.....	37
2.2.4 Our Contribution	38
2.3 Research Design.....	39
2.3.1 Systematic Review	39
2.3.2 Coding Procedure	41
2.3.3 Expert Interviews	43
2.4 Results.....	44
2.4.1 Systematic Review	44
2.4.2 Basic Study Characteristics	44
2.4.3 When Did Researchers Validate?	46
2.4.4 How Did Researchers Validate?	48
Next, we turned to the evaluation of how researchers validate CTAM.....	48
2.5 Expert Interviews	53
2.5.1 What Is The Purpose Of CTAM Validation	53
2.5.2 Which Validation Steps Are Seldomly Reported.....	54
2.6 Discussion.....	57
2.6.1 Critical Reflection	57
2.6.2 Practical Recommendations to Improve CTAM Validation	58
2.6.3 An Outlook on more Unified Validation Frameworks	62
2.7 Limitations and Directions for Future Work.....	64
2.8 Conclusion	65
2.9 Appendix.....	67
2.9.1 Appendix: Selection of Journals.....	67
2.9.2 Appendix: Benchmark Literature.....	68
2.9.3 Appendix: Search Terms	68
2.9.4 Appendix: List of Studies Included in the Systematic Review	68
2.9.5 Appendix: Overview of Studies per Journal	69

2.9.6	Appendix: Interview Guideline.....	69
3.	Developing Conceptual and Practical Guidance for Validation	70
3.1	Introduction	70
3.2	Bias and Error in Computational Text Analysis.....	71
3.3	A Critical Perspective on Human Annotations as a Gold-Standard.....	72
3.4	Deriving the Validation Framework	74
3.4.1	Types of Validation Evidence.....	74
3.4.2	Selecting Validation Steps	76
3.5	Overview Framework.....	77
3.5.1	Substantive Evidence	77
3.5.2	Structural Evidence	80
3.5.3	External Evidence.....	82
3.5.4	Robustness Checks	83
3.5.5	Checklist.....	84
3.6	Critical Reflection	85
3.7	Conclusion	86
3.8	Appendix.....	88
3.8.1	Appendix: Glossary of Validity Terms	88
3.8.2	Appendix: Context-dependent Types of Validation Evidence	89
3.8.3	Appendix: Illustrative Example: Measuring Sexism in Social Media Comments.....	92
4.	Framework Application 1: Personality Traits.....	100
4.1	Introduction	100
4.2	Literature	102
4.2.1	Politicians Personality Traits Matter	102
4.2.2	Politicians Strategically Display Personality Cues	103
4.2.3	Personality Cues are Inherently Ambiguous	105
4.2.4	How Can We Measure Personality Cues?	106
4.2.5	The Need for Computational Approaches.....	107
4.3	Method	108
4.3.1	Data.....	108
4.3.2	Measurement	109
4.3.3	Validation.....	113
4.4	Results.....	113
4.4.1	Substantive Validation.....	113
4.4.2	Comparison with Human-annotated Test Set (External Validation) ..	115
4.4.3	Functional Tests (Structural Validation)	116
4.4.4	Exploring Sensitivity to Partisan Differences (External Validation) ..	118
4.5	Discussion.....	120
4.5.1	Summary and Discussion	120
4.6	Appendix.....	122
4.6.1	Appendix: Overview of politicians in the German Raw Corpus.....	122
4.6.2	Appendix: General Workflow	124
4.7.1	Appendix: Filter Terms.....	126
4.7.2	Appendix: Codebook	127
4.7.3	Appendix: Hyperparameters.....	130
4.7.4	Appendix: Computational Resources	131
4.7.5	Appendix: Prompts.....	132
4.7.6	Appendix: Validation Strategy.....	136

4.7.7	Appendix I: Filled-Out Checklist.....	137
4.8.1	Appendix: Test Units.....	143
4.9.1	Appendix: Personality and Ideology.....	144
4.9.2	Appendix: Sensitivity Analysis for Agency and CHES Dimensions	146
4.9.3	Appendix: Distribution of Agency and Communion across Data.....	147
4.9.4	Appendix: Personality Scores and Gender Across Parties.....	147
5.	Framework Application 2: Geographic Representation	149
5.1	Introduction	149
5.2	Spatial Representation.....	151
5.2.1	Electoral Institutions and the Logic of Representation	151
5.2.2	Representing the Local	152
5.2.3	Representing the Deprived.....	152
5.3	Research Design.....	153
5.3.1	Case and Data Selection.....	153
5.3.2	Place Identification	154
5.3.3	Analysis	156
5.4	Results.....	157
5.4.1	Measurement Validation	157
5.4.2	Descriptive Statistics.....	158
5.4.3	Hypotheses	160
5.5	Discussion and Conclusion	161
5.6	Appendix.....	163
5.6.1	Appendix: Exemplary SPARQL Query	163
5.6.2	Appendix: Geographical Matching	164
5.6.3	Appendix: Constituencies	164
5.6.4	Appendix: Validation	165
5.6.5	Appendix: Descriptive Overview Places	166
5.6.6	Appendix: Descriptive Overview Accounts.....	166
5.6.7	Appendix: Descriptive Statistics.....	167
6.	Conclusion.....	168
6.1.1	Main Findings and Contributions	168
6.2	Research Objectives	168
6.3	Contributions to the Research Community	170
6.4	Limitations	171
6.5	Further Directions for Research.....	173
6.6	Concluding Remarks.....	174
7.	Bibliography.....	176

Acknowledgements

Writing a dissertation is, at least in my case, anything but a lonely adventure. It has been a journey marked by great exchanges, collaborations, and support from many individuals who have shaped my thinking, research, and motivation to persevere. I am deeply grateful to all of them.

First and foremost, I would like to thank my two supervisors Claudia Wagner and Sebastian Stier. Claudia has guided me from the very beginning with exceptional feedback, critical reflection, and continuous encouragement. Her unique perspective on social science research has profoundly influenced the way I approach research questions, and I have learned a great deal from her clarity of thought. Sebastian has been a steady mentor and colleague even before he joined as the supervisor of this dissertation. His shared interest in my research topics and his openness to get involved in this project have made a lasting difference. Having someone so approachable and supportive in my environment was truly invaluable.

I furthermore owe special thanks to Clemens Lechner. As a mentor and primary collaborator, he has guided and accompanied me through every step of the dissertation process. His scientific expertise, precision, and encouragement have shaped my work more than any other influence. I have learned immensely from his way of thinking and from his commitment to high-quality research. His impact on this dissertation can't be stressed enough.

My sincere appreciation also goes to my coauthors Marius Sältzer, Constantin Wurthmann, and Anne-Katrin Stroppe. Working with them was an ongoing source of inspiration and intellectual growth. They helped me broaden my perspective beyond my own projects and showed me what collaborative research at its best can look like. I am also thankful to my SDD team, especially Lydia Repke, Matthias Roth, Piotr Koc, Lena Römer and Matthias Blümke for their constant support, teamwork, and humour. The team atmosphere has been a key source of motivation throughout this process. Further thanks go to the current and former members of the DBD Working Group, whose discussions, feedback, and ideas have also contributed to this dissertation: Jessica Daikeler, Vanessa Lux, Leon Fröhling, Indira Sen, Jan Schwalbach, Yannik Peters, Karim Farah, Henning Silber, and Katrin Weller. Their collaboration has shaped my work in many subtle but meaningful ways.

Finally, I want to thank all my friends and family who have supported me along the

way. Among those, some stand out. My parents, Thomas and Monika Birkenmaier, have been a great source of stability and encouragement. My girlfriend, Anna Donauer, has accompanied me through the highs and lows of this process with endless patience, understanding, and love. And Peter, Holger, Vici, Lale, Jan, Lea, Schwabini, Ivan, and all persons around the P11, who have witnessed many hours of me working on this dissertation while drinking coffee at the bar counter.

This dissertation is the result of all these influences, and I am sincerely thankful to each and every one of them.

List of Figures

Figure 1: Measurement of Social Constructs	13
Figure 2: Overview of Empirical Contributions.....	27
Figure 3: Text as Data Publications within the Web of Science Database	34
Figure 4: Coding Procedure	41
Figure 5: Coding Scheme.....	42
Figure 6: Study Characteristics.....	44
Figure 7: Number of Validation Steps (Uni- and Bivariate).....	47
Figure 8. Categories of Validation Steps	48
Figure 9: Overview of Specific Validation Steps.....	50
Figure 10: Overview of Studies per Journal.....	69
Figure 11: Three Fundamental Types of Validation Evidence	75
Figure 12: ValiText Framework.....	77
Figure 13: Conceptual Model (Adapted from Brunswik, 1955.....	104
Figure 14: Workflow	109
Figure 15: Classifier Performance for Balanced and Unbalanced Datasets.....	115
Figure 16. Confusion Matrices for the Test Units (n=18).....	117
Figure 17: Correspondence of Aggregated Politicians Shares for each Party	119
Figure 18: Ratio of Agency and Communion	120
Figure 19: Screenshot of the Respective Codebook Section for the Rile-Score	144
Figure 20: Correspondence of Aggregated Agency Politicians' Shares	146
Figure 21: Mean Levels of Agency and Communion across Communication Formats	147
Figure 22: Distribution of Scores by Gender	148
Figure 23: Computational Workflow	154

Figure 24: Predicted Values for Mandate Mode and District Match.	158
Figure 25: Distributions of District Distance and Service Distance.....	159
Figure 26: Regression Coefficients	160
Figure 27: Robustness Checks	161
Figure 28: Matching Graph	164

List of Tables

Table 1: Comparison of Textual and Survey Data	17
Table 2: Outcomes from Limitations and Risk Factors of Computational Methods ...	19
Table 3: Rarely Reported Validation Steps	55
Table 4: Validation Steps for Substantive Evidence (Context-Independent)	78
Table 5: Validation Steps for Structural Evidence (Context-Independent)	80
Table 6: Validation Steps for External Evidence (Context-Independent)	82
Table 7: Robustness Checks (Context-Independent)	83
Table 8: Use Cases	84
Table 9: Validation Steps for research interest characteristics (context-specific)	89
Table 10: Validation Steps for use case characteristics (context-specific)	91
Table 11: Interrater Agreement for Human Coding (n=1000)	114
Table 12: Overall Best-Performing Models on the Test Data (3 Random Seeds)	116
Table 13: Overview Politicians	122
Table 14: Validation - Comparison with Human Annotations	157
Table 15: Top 10 Accounts with no. Tweets	166
Table 16: Top 10 Accounts with no. Facebook Posts	166
Table 17: Descriptive Statistics	167

1. Introduction

1.1 Problem Statement

1.1.1 Growing Application of Computational Text Analysis in Research

Textual data, such as letters, social media posts, or newspaper articles, stand out as one of the most valuable sources of information for studying and explaining human behavior (Grimmer et al., 2022). Hence, it is not surprising that the systematic analysis of textual content has a long-standing tradition in social science research, dating back to the 17th century in the field of theology. Dovring (1954), for instance, describes one of the first recorded cases of how the Swedish state church employed systematic text analysis in 1743 to examine a collection of 90 hymns titled *Songs of Zion*, created by unsanctioned sources, to determine whether they were blasphemous or aligned with the Church religious standards. Using the identification of key symbols, the division of content into favorable and non-favorable categories and the coding of values into these categories, their approach contained many strategies at the core of modern understanding of text analysis.

Naturally, research in the past has relied on manual approaches for text analysis, commonly referred to under the term “content analysis”. Even though the term “content analysis” itself was only first used in 1941, its core principles have been an established research method since the beginning of the 20th century (Berelson & Lazarsfeld, 1948).¹ The popularity of text analysis was further fueled by the rise of mass media, which facilitated large-scale production and dissemination of textual content (Chaffee & Metzger, 2001). Likewise, public opinion and communication research emerged as a research field throughout the 20th century, examining phenomena such as public discourse and rhetorical elements in textual content (Krippendorff, 2018). One of the earliest examples of such work was conducted by Speed (1893), who documented how New York newspapers between 1881 and 1893 shifted their focus away from religious, scientific, and literary topics toward gossip, sports, and scandals.

Traditionally, content analysis is a labor-intensive process in which human coders manually examine textual sources and assign categories or labels based on predefined criteria. However, the exponential growth in the availability of textual data,

¹ In 1910, for instance, Max Weber at the first meeting of the German Sociological Society in 1910, proposed a large-scale content analysis of the press

coupled with advances in information technology, spurred the development of computational approaches to content analysis (Hays, 1960; Stone et al., 1966). The 1950s marked the early adoption of such methods. For instance, Sebeok and Zeps (1958) utilized computer programs to analyze Cheremic folksongs by testing for co-occurrences of words within thematically similar paragraphs. Their pioneering work demonstrated the potential of computational methods that could scale up manual content analysis to explore completely new research questions and methodological approaches.

The usage of computational methods then gained massive popularity during the 2000s, driven by technological advancements, better access to information technologies, and a renewed emphasis on leveraging computational methods for quantitative text analysis (Krippendorff, 2018; Macanovic, 2022). Grimmer and Stewart’s (2013) influential article offers a valuable snapshot of computational text analysis methods as they stood in the early 2010s, particularly within political science research. Under the label *text as data*, Grimmer and Stewart offer a comprehensive review of various types of computational text analysis methods (CTAM), including dictionary-based approaches as well as supervised and unsupervised methods, such as Random Forrest or Support Vector Machine (SVM) models, and specifically discuss the methodological challenges of applying these tools in social science research.

One limitation of these early CTAM, however, was their lack of semantic and contextual understanding, as they primarily relied on *bag-of-words* approaches. Such approaches treat text as unordered collections of words, ignoring grammar, word order, and nuanced meaning. This makes them effective for simple frequency or exploratory analysis but insufficient for capturing deeper semantic or contextual relationships. The introduction of Word2Vec and similar *embedding* approaches, however, marked a methodological breakthrough (Mikolov et al., 2013). By mapping words into high-dimensional vector spaces, text *embeddings* place semantically related words closer together, thereby preserving contextual and semantic relationships that term-frequency matrices cannot capture.

The usefulness of embeddings for analysing textual content was significantly enhanced by the introduction of the transformer architecture by Vaswani et al. (2017). Using their attention mechanism (i.e., the ability to weight the importance of different words in a sentence relative to each other), advanced neural network architectures like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) marked a significant leap in computational text analysis by enabling contextual

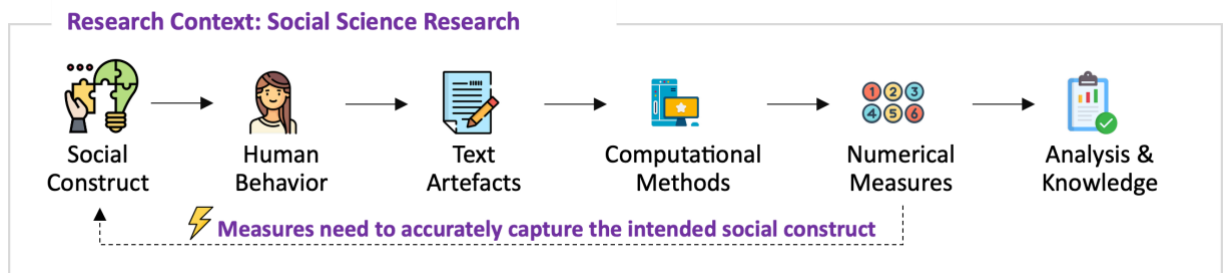
embeddings, where the representation of a word depends on its surrounding text. This is because attention-based model architectures process entire texts in parallel rather than sequentially, allowing for a deeper and more context-aware understanding. The transformer’s architecture further laid the groundwork for the development of state-of-the-art large language models (LLMs), such as GPT-4o (Brown et al., 2020) or Llama 3.3, (Touvron, Lavril, et al., 2023) which can capture complex dependencies and generate contextually coherent text across diverse inputs.

Today, computational methods are more widely used than ever in social science research, with computational content analysis emerging as a prominent approach for addressing substantive research questions across a broad range of disciplines (Bestvater & Monroe, 2023; Licht, 2023; Widmann & Wich, 2022). For instance, analysing large corpora of politicians’ speeches can uncover what drives the adoption of populist or anti-elite rhetoric in political communication (Licht et al., 2024), while studying online discourse on platforms like Reddit can help identify patterns of hate speech or misinformation (Sen et al., 2022).

1.1.2 Empirical Measures of Social Constructs need to be Validated

A central application of computational text analysis in the social sciences is the operationalization and measurement of social constructs within textual data. ² 1 provides an abstract visualization of this workflow.

Figure 1: Measurement of Social Constructs



As illustrated in Figure 1, this process begins with a *social construct* of interest, such as attitudes, personality traits, or ideology. These constructs manifest in *human behavior*, which leaves observable traces in the form of *textual artefacts*. Depending on the research context, these texts can take various forms, from single-word responses

² Throughout the text analysis literature, terms such as *detect*, *categorize*, *assign*, and *measure* are often used interchangeably to describe the process of assigning numerical values to textual content as outlined in Figure 1. In this dissertation, I adopt the term *measure* to describe the process of measuring social constructs in textual content, aligning with the broader aim of operationalizing abstract concepts for empirical analysis.

(e.g., the word “pathetic” in response to a social media comment) to large corpora or even complete social media feeds (e.g., all social media posts on a specific platform).³ Through *computational methods*, these artefacts are systematically transformed into *numerical measures*. This step enables researchers to apply statistical models and generate new *analysis and knowledge* about the relationships between social constructs. In this respect, computational text analysis aligns with the broader logic of measuring latent constructs in the social sciences: empirical data are translated into structured numerical representations according to explicit coding rules (Stevens, 1946; see B. D. Wright, 1997). Unlike traditional content analysis, where human coders manually assign numerical measures to specific text units, however, CTAMs depend on algorithms to implement decision rules. Nonetheless, measures derived from computational text analysis must adhere to the same quality criteria as other types of social science measurements.

While three main criteria determine the quality of scientific measurement instruments-i.e., objectivity, reliability, and validity-validity is the quality criterion that is generally considered the most fundamental one. In essence, validity is the extent to which a scientific measure captures what it is supposed to measure (Kelley, 1927; see also Repke et al., 2024 for a thorough discussion). Validation is necessary because multiple factors can lead to a situation where the numerical output of a measurement instrument does not accurately capture the underlying construct.

To demonstrate validity⁴, researchers must provide convincing validation evidence outlining that the measurement instrument accurately captures the intended social construct and distinguishes it from unrelated phenomena.

Take the example of a survey attempting to measure political trust. A common item might ask: “How much do you trust the government?” If respondents interpret “government” to mean the current ruling party, while the researcher intended it to mean public institutions in general, the resulting data will not validly reflect political trust as theoretically defined. In this case, the issue lies in *ambiguity* in

³ Throughout this dissertation, I define *textual data* as any piece of textual information that has the potential to offer sufficient insight into the construct being studied. In some cases - like measuring negative remarks about certain groups - a single word may be enough. In others - such as measuring persuasion style - longer text units, at least at the sentence level, are necessary.

⁴ Throughout this dissertation, I use the terms “validity” and “measurement validity” simultaneously since I focus on validity in the context of measurement. In the social science literature, validity can also describe the features of a research design, including aspects like internal validity (statement about causal relationships) and external validity (statement about the generalizability of results) (Campbell & Stanley, 2015; Cook et al., 1979). In this dissertation, however, the emphasis is specifically on measurement validity-how accurately a measurement captures the intended construct.

operationalization, one among many issues that can arise and distort the validity of empirical measurements (Frenkel-Brunswik, 1949; Kajs & McCollum, 2009). In this case, validation could include issues such as cognitive interviewing or empirical pretesting, which can reveal how respondents interpret terms like “government” and whether these interpretations align with the construct the researcher intends to measure (Lenzner et al., 2023). Post-hoc validation, in contrast, can help detect persistent issues in larger samples through techniques such as factor analysis or by inspecting data for systematic response patterns.

Similarly, consider an example of a computational text analysis designed to measure “empathy” in political speeches. This could involve training a machine learning classifier on a dataset where texts labeled as “empathetic” contain words like “care”, “help”, or “understand”. Hence, the model would learn to associate these words with empathy and achieve high convergence with human-labeled data, generally regarded as a crucial requirement for computational text classification. However, when applied to real-world political speeches, the method could begin to label any text with those words as empathetic-regardless of the broader context. For instance, the sentence “We must help corporations thrive so they can care for their shareholders” might be scored as highly empathetic, even though it arguably misses the intended human-centered meaning of empathy.

Taken together, both examples point to several challenges for the measurement - and the validation of - social constructs using empirical research methods.

1.1.3 CTAM Measures of Social Constructs are Challenging to Validate

Among social science measurement instruments, the validation of CTAM measures of social constructs proves especially complex, centering on three key challenges.

1. First, the inherent **ambiguity and multidimensionality of textual content** poses a fundamental challenge for any text analysis task. Because social constructs are inherently abstract and not directly observable, they can manifest in language in diverse and often ambiguous ways. The lack of genuine ground-truth labels, combined with the multidimensional nature of textual content, leaves room for multiple interpretations of meaning (Song et al., 2020; Yeomans, 2021), complicating efforts to validate discrete measures of these constructs inherent in textual content.

2. Second, the **limitations and biases inherent to computational methods** raise additional concerns, as they can introduce systematic errors into the measurement process. This issue is exacerbated by the rapid advancement of novel methods and models⁵ - particularly caused by the recent transition from traditional machine learning models to large language models (LLMs) - which has outpaced the establishment of validation standards within the research community. As a result, researchers frequently deploy novel methods and models without clear guidance on how method-related constraints and peculiarities influence measurement validity.
3. Finally, the interdisciplinary nature of computational text analysis results in a **blending of disciplinary traditions and terminology**. Researchers that apply computational text analysis often integrate perspectives from disciplines such as psychology, political science, and computational linguistics - each with its own terminology and implicit best practices. As a result, researchers often operate without a common understanding for assessing validity of CTAM measures of social constructs, instead relying on fragmented practices and vocabulary to structure validation.

In the next subchapters, I will dive deeper into each of the major challenges outlined above.

1.1.3.1 Ambiguity and Multidimensionality of Textual Content

The literature on content analysis offers extensive background on the challenges of measuring social constructs from textual content for scientific research (Krippendorff, 2018; Neuendorf, 2017). These challenges often stem from the inherent ambiguity and multidimensionality of textual data, which differ significantly from more structured forms of empirical data. In the following, I use survey data as a reference point, given their central role in social science measurement and their established standards for validity and reliability, to highlight key challenges in working with textual data. This comparison does not exclude other data types commonly used in the social sciences, such as experimental or medical data, but serves to illustrate most clearly how

⁵ Throughout this dissertation, I will refer to *strategy* as overarching measurement strategies (such as training/fine-tuning or prompting an LLM), and *model* as specific computational models (e.g., XLM-RoBERTa or Deepseek-V3)

methodological approaches differ in the measurement of social constructs. An overview of this comparison is summarized in Table 1.

Table 1: Comparison of Textual and Survey Data

Dimension	Textual Data	Survey Data
Data Collection⁶	(Mostly) Passively collected (found data)	(Mostly) Actively collected (designed data)
Construct Conceptualization	After the data collection	Before the data collection
Data Dimensions	Multidimensional (various text features)	Unidimensional (numerical values)
Operationalization Construct	Judgement according to some rules on which text features relate to the construct	Respondents' feedback on designed questionnaire items
Context-specific interpretation of data features possible?	Yes	No

At the beginning of the research process, text analysis usually starts with a sample of texts that provides the corpus to be analyzed by the researcher. Unlike survey data, which is actively collected with predefined constructs in mind, textual data, however, is often generated organically, without any intent of being used for systematic research (Howison et al., 2011). For example, Wikipedia has become a popular source of study for scholars investigating issues such as gender bias (Pradel, 2021; Schmahl et al., 2020) or online discussions (Xiao & Xiao, 2020). However, its content was never designed to facilitate such inquiries-it was created to disseminate knowledge collaboratively. Similarly, politicians' speeches, social media posts, or party manifestos are produced for practical or outreach purposes, not to answer specific research questions. This data generation process fundamentally contrasts with survey data, where researchers design questions, items, and sampling strategies to capture precisely the constructs they seek to measure.

Additionally, textual data differ significantly from survey data in its dimensionality and how constructs manifest themselves in the data. In survey research, constructs are typically captured through well-defined and standardized scale items, which produce data in a structured, unidimensional numerical format.

⁶ In this dissertation, I primarily work with textual data classified as "digital behavioral data," which are digital observations of human and algorithmic behavior recorded without the subject's active participation. However, most of the characteristics and challenges of "digital behavioral data" are also applicable for working with "designed textual data", i.e., open-ended responses in a survey field that require active participation by the respondent.

This format allows for a straightforward interpretation, where higher scores on a scale directly correspond to higher levels of the construct being measured. Textual data, on the other hand, is multidimensional and conveys meaning through a rich combination of word choices, syntax, context, and cultural references. As Krippendorff (2018) states, “texts do not have single meanings that could be found, identified, and described for what they are” (Krippendorff, 2018, p. 23). Take, for instance, the construct of “populist rhetoric” in political communication. In survey research, participants might be asked directly whether they perceive a statement as populist, using survey scales that measure dimensions of populism and producing numerical scores summarizing participants' attitudes (Castanho Silva et al., 2020; Pauwels, 2017; Wuttke et al., 2020). However, in textual data, populist rhetoric can manifest in varied ways-through the usage of specific slogans like “drain the swamp”, statements about certain groups like “the elite,” but also the usage of simpler language and vocabulary (Decadri & Boussalis, 2020). This diversity means that there are multiple valid ways to operationalize constructs, complicating the decision of which features to consider.

Similarly, it becomes evident that seemingly objective text features, such as specific sentences, can be interpreted differently depending on the reader's linguistic background, cultural and political context, and personal perspective (Artstein & Poesio, 2008; Plank, 2022). As a result, textual features are inherently subject to varied interpretations influenced by the context in which they are perceived and the linguistic cues emphasized (Drisko & Maschi, 2016). A rich literature in the content analysis domain demonstrates that human annotators often disagree when analysing the same texts, even when provided with identical instructions (Krippendorff, 1980). Accordingly, Krippendorff (2018) regards alpha agreement values between 0.667 and 0.8 as sufficient for tentative conclusions in scientific research, despite these constituting notable disagreements among raters working under uniform conditions and reading the same texts.⁷ Unlike text data that allow for varied interpretations, survey data provide a direct numerical measure of the social construct being studied (i.e., the response value for a specific item on a Likert scale), thus offering a rather objective data point that supports linear interpretation.⁸

⁷ As outlined throughout this dissertation, such context-independent thresholds can be questioned, as the acceptable level of agreement may vary considerably depending on the complexity of the task, the ambiguity of the material, and the intended use of the data.

⁸ This objectivity of the data does not relate to the traditional quality criteria of scientific measurement, i.e. validity, reliability and objectivity. Instead, the focus here is on the features of the output data, i.e., the numerical values that can only be interpreted in a linear way with higher values corresponding higher levels of the targeted construct

1.1.3.2 Limitations and Biases of Computational Methods

In addition to the challenges inherent to text analysis outlined in the previous chapter, relying on computational text analysis introduces its own set of limitations and biases (Baden et al., 2021; Howison et al., 2011; van Atteveldt et al., 2021).

At its core, as Grimmer and Stewart (2013, p. 269) famously state in their first principle of automated text analysis that “all quantitative models of language are wrong (but some are useful).” Intuitively, Grimmer and Stewart point to the presence of various sources of measurement mistakes that are introduced by quantitative text analysis methods. This is important as random mistakes (*errors*), being uncontrolled and incidental, can still be considered as part of a valid measurement instrument, as it is expected that random errors cancel out across multiple measurements. On the other hand, systematic mistakes (*biases*)⁹ are more concerning, as they represent non-random distortions that lead measurements to diverge from the true target construct (Bound et al., 2001; Kimberlin & Winterstein, 2008). In the validation process, researchers, therefore, need to assess whether the limitations of computational models in a specific research design lead to systematic biases that could negatively impact the validity of the measurement.

Below, I outline potential limitations and risk factors of computational approaches to text analysis, together with their respective negative outcomes for the measurement process in Table 2. However, given the heterogeneity of methods and models that may vary across different computational workflows, I focus on two common exemplary use cases in computational text analysis: (1) *fine-tuning* a language model such as BERT, which involves relying on a pretrained model and adjusting its parameters on a task-specific dataset, and (2) *few-shot prompting*, that is, crafting specific input prompts to guide a language model such as Llama 4 toward desired outputs without modifying the model itself (Touvron, Lavril, et al., 2023).

Table 2: Outcomes from Limitations and Risk Factors of Computational Methods

Outcomes

⁹ The term *bias* is multifaceted and context-dependent, yet it consistently signals a form of deviation or distortion. In NLP, it often refers to systematic patterns in data that lead models to reproduce social inequalities. In statistics, *bias* is usually used as a technical term for systematic error - how far an estimator's expected value is from the true value.

Potential Limitations and Risk Factors	Use Case (1) Fine-Tuning (e.g., a BERT model)	Use Case (2) Few-Shot Prompting (e.g., a Llama 4 model)
Poor and Shallow Proxies	- Model is trained on text features that falsely correlate with the construct to be measured	- Prompted outputs reflect shallow associations rather than a true understanding of the data
Technical Innovations and Variability of Methods & Models	- The heterogeneity of novel methods and models results in erroneous or superseded model selections	
Researchers' degrees of freedom	- Measures are sensitive to data (pre-)processing decisions - Model is sensitive to hyperparameters selection - (No additional limitations and risk factors beyond the general ones)	- Model is sensitive to prompt wording
Algorithmic Bias	- Initial pre-training data contains inherent societal bias - Initial pre-training data contains cognitive bias - Labeled data are influenced by cultural perspectives and subjective judgments of annotators	- Prompt design encourages biased answers
Interpretability	- Intermediate representations are hard to interpret	
Generalization	- Model overfits to training data and fails on out-of-domain data	- (No additional limitations and risk factors)

1.1.3.2.1 Poor and Shallow Proxies

A major concern in both fine-tuning and prompting approaches is that machine learning models often pick up on patterns only loosely associated with the intended construct. When fine-tuning language models (1), models may latch onto superficial text features that correlate with the target variable but fail to capture its conceptual meaning. For example, methodological research by Hirst et al. (2014) vividly demonstrated that a classifier trained to classify political ideology instead relied on textual features related to incumbency status, which merely correlated with ideological markers in the training data. Likewise, when relying on few-shot prompting (2), poor proxies might manifest through surface-level associations embedded in the pretraining data rather than by a deep understanding of the intended concept. A sentiment analysis prompt, for instance, might yield overly positive interpretations of text containing polite or formal language in a diplomatic setting, regardless of actual sentiment.

1.1.3.2.2 Technical Innovations and Variability of Methods & Models

The rapid pace of technical innovation and the wide methodological variability in computational text analysis increase the risk of using models and workflows that are

erroneous or superseded. One key aspect that sets computational text analysis apart from more traditional measurement instruments - such as surveys or interviews - is the rapid pace of technical innovation and the high variability in available methods and models.¹⁰ Traditional measurement approaches in the social sciences, like expert interviews, benefit from well-established methodological frameworks and guidelines that have evolved over decades (Bogner et al., 2009). In contrast, computational methods are developing at an unprecedented speed, with new models and techniques emerging continuously, both from proprietary and open-source providers.

For instance, companies like OpenAI frequently update their models, and even major releases-such as Llama-2 (Touvron, Martin, et al., 2023) by Facebook - are quickly superseded by newer and more powerful versions. Moreover, even when researchers build on the same foundational models, such as BERT, the methodological implementations of these models can differ significantly. These variations include lightweight variants like ModernBERT (Warner et al., 2024), the use of adapter layers (Houlsby et al., 2019), or task-specific adaptations such as BERT-NLI (Yin et al., 2019). Additionally, the technical frameworks and libraries used to deploy these models can vary significantly¹¹, as well as the technical dependencies of different versions and packages that are used in the workflow, making it difficult to replicate findings and detect methodological errors.

1.1.3.2.3 Researchers' Degrees of Freedom

Another risk factor lies in the wide range of decisions required by researchers that can affect the outcome of the measurement process. For instance, researchers usually process the textual data before feeding it into computational models, such as conducting anonymization, removing meta-data, or filtering for subsets of the data (Denny & Spirling, 2018). These preprocessing decisions can substantially affect the measurement scores and, consequently, the outcomes of subsequent analyses, potentially compromising the validity of the scientific findings. Moreover, as with most computational models, the measures can also be influenced by the set of hyperparameters used to train the models, such as learning rate, batch size, or the number of epochs. Even slight modifications to these parameters can lead to measurable differences in model performance or interpretation (Arnold et al., 2023).

Additionally, use case-specific workflows introduce further variability. For instance, few-shot prompting (2) requires prompt design decisions, where subtle

¹⁰ For an excellent overview of state-of-the-art resources, see <https://huggingface.co/>

¹¹ One example for this is e.g., the “vLLM” framework (<https://github.com/vllm-project/vllm>)

changes in prompt phrasing or the sequence of few-shot examples can lead to significantly different outputs (Sclar et al., 2023; Zhao et al., 2021). For example, prompting an LLM with the question “Classify all statements that can be seen as populist?” might lead to a more inclusive response, accommodating a broader range of interpretations. In contrast, asking “Classify all statements that are populist?” could push the model toward applying a narrower, more definitive classification, potentially excluding statements that are ambiguous or context-dependent.

1.1.3.2.4 Algorithmic Bias

It is also crucial to acknowledge that the data used for training language models are not neutral; they often reflect and perpetuate societal biases embedded within them (Bender et al., 2021; Guo et al., 2024; Pagano et al., 2023; Ray, 2023)

While a discussion of all possible biases is beyond the scope of this dissertation, Guo et al. (2024) highlight two different sources of *societal bias*.

On the one hand, *representational bias* arises when certain demographic groups - such as those defined by gender, age, race, religion, ethnicity, culture, political orientation, or socioeconomic status – and their associated stereotypes are either overrepresented or underrepresented in the training data (Alvero et al., 2024).

On the other hand, *spatial and temporal bias* refers to the influence of geography and time on training data. Language models trained primarily on texts from certain regions, especially Western countries, may internalize and reflect those cultural norms while failing to adequately represent others (Santurkar et al., 2023; Talat et al., 2022). Likewise, models trained on older texts may inherit outdated or prejudiced societal values, such as racist or sexist language common in historical sources (Mehrabi et al., 2022).

Beyond societal bias, a growing body of research highlights that language models also replicate *cognitive biases*—systematic patterns of flawed human judgment under uncertainty (Tversky & Kahneman, 1974). Various cognitive biases observed in human behavior can be mirrored by LLMs (for an overview of relevant biases in the field of medical language models, see Schmidgall et al., 2024). For example, consistent findings show that LLMs exhibit availability bias, as evidenced by their tendency to misclassify rare cases in imbalanced datasets, echoing the human inclination to rely on more frequent and easily recalled patterns (Lin & Ng, 2023; Zhao et al., 2021).

Bias in language models can also emerge beyond the initial training phase. During fine-tuning (1), it is important to acknowledge that the labelling of training data is not value-neutral; it is shaped by the cultural perspectives and subjective

interpretations of human annotators (Plank, 2022). These can vary significantly across cultural and sociodemographic backgrounds, with the same expression interpreted as either polite or rude depending on the annotator’s cultural context (Uma et al., 2021). Such inconsistencies - especially when coupled with low-quality data - can further entrench or amplify bias in measurements (Song et al., 2020).

Additionally, in the case of few-shot prompting (2), the way prompts are designed—including the specific examples they contain—can influence model outputs in ways that reinforce biases. For example, if all examples in the prompt label only right-wing or left-wing slogans as “divisive,” the model may learn to associate divisiveness exclusively with right-wing or left-wing discourse, reinforcing ideological bias in classification.

1.1.3.2.5 Interpretability

All language models share a critical limitation: a lack of interpretability. This means that the decision-making processes are buried in high-dimensional vector representations and non-linear transformations, often requiring tools like LIME, SHAP, or integrated gradients to approximate explanations (Molnar, 2020; Zhao et al., 2024). Concerns about interpretability and the need for transparency in algorithmic decision-making emerged well before large language models. The “right to explanation” in the EU’s GDPR (Goodman & Flaxman, 2017), for instance, underscored growing legal and ethical demands for machine learning systems that make their reasoning accessible (Doshi-Velez & Kim, 2017). In the context of CTAM research, interpretability is likewise crucial to understanding which textual features drive model outputs to ensure these are caused by the features of interest, and not spurious relations (Hirst et al., 2014).

1.1.3.2.6 Generalisation

Finally, limitations in generalization persist, primarily for fine-tuning approaches (1). For instance, fine-tuned BERT models tend to perform well on in-domain data but often falter in out-of-domain settings, indicating overfitting. A classifier fine-tuned on parliamentary debates might not transfer well to social media posts or news articles. A prompted LLaMA 4 model may generalize more flexibly due to their broad pretraining, but performance can still degrade in domain-specific or long-context scenarios-especially when the prompt lacks sufficient context.

1.1.3.3 Blending of disciplinary traditions and terminology

While the definition of validity is relatively straightforward (“measuring what [the measurement instrument] is supposed to measure”), in practice, CTAM validation is further complicated by the blending of disciplinary traditions and terminology. While these challenges are especially prevalent in the context of CTAM, they are not unique to it and can also be found across various methodological disciplines and subfields of social science research (Borsboom et al., 2004; Hughes, 2018; Kane, 2001). As such, the term “validity” and the process of “validation” carry multiple interpretations and are applied with considerable variability within social science research.

“One notable example of varying understandings and terminology is the evolution of the concept of validity in psychometric research- the research field focusing on psychological and educational testing, predominantly using survey instruments-which has devoted substantial resources and engaged in extensive debates on the valid application of measurement instruments. In the early 20th century, validity in the psychometric domain was conceptualized as a property of tests, defined as the extent to which a test measures what it claims to measure (Kelley, 1927; Ruch, 1924). Thus, validity was predominantly assessed by matching the content of psychometric measures with theory (content validity) and by examining predictive capabilities (predictive validity) (Hughes, 2018). In the 1950s, Cronbach and Meehl (1955) introduced the concept of construct validity, which they viewed-unlike contemporary interpretations-as a distinct type of validity. Construct validity described a much broader concept of validity that is rather concerned with the examination of a construct’s nomological net and required careful interpretation of theoretical constructs within specific research contexts. With ongoing methodological research, the 1970s and 1980s saw the rise of a unified approach to validity, spearheaded by Messick redefined validity under an overarching concept of construct validity, advocating for the use of multiple sources of validity-supporting evidence-such as content-related and criterion-related evidence-to establish construct validity. Thus, construct validity, in a unified sense, also emphasizes the importance of contextual factors in assessing validity and encourages careful consideration of the consequences of scientific interpretations and decisions derived from test scores (Hughes, 2018). Today, the unified definition of construct validity remains the prevailing perspective in the psychometric domain, as outlined in the Standards for

Educational and Psychological Testing (American Educational Research Association, 2011).

As in psychometrics, political science has developed its own understandings of different types of validation evidence—though these are often less formalized, reflected more in evolving best practices than in codified standards. In political science, Adcock and Collier (2001) carefully document the challenges of reaching consensus on validation, particularly in efforts to bridge qualitative and quantitative traditions. Beyond methodological divides and debates over the meaning of core constructs, they point to a major source of confusion: the inconsistent and overlapping terminology used to describe validation procedures. With more than 37 different modifiers applied to the term “validity”—such as content, criterion, convergent, and construct—validation emerges as a fragmented and continuously evolving process. This evolution is often driven by the emergence of new research contexts that demand adaptations to existing validation standards.

For instance, one commonly applied approach for validation is the comparison of model outputs with human annotations. These human-generated labels are typically treated as the gold standard—assumed to provide objective and accurate representations of the underlying constructs. However, while they are essential for grounding computational results in human interpretation, there is increasing recognition that human labels alone cannot fully establish validity. In response to these limitations, researchers often apply additional validation steps (see Grimmer and Stewart, 2022). Yet these efforts are rarely coordinated or situated within a coherent framework. This is because existing validation frameworks originate from fields such as survey research and educational assessment, which do not fully address the unique challenges posed by computational text analysis.

Collectively, these challenges - the ambiguity and multidimensionality of textual content, the limitations and biases of computational methods, and the blending of disciplinary traditions - result in a central issue: the lack of guiding frameworks and guidelines for validating CTAM measures of social constructs (Baden et al., 2021; Grimmer et al., 2022). Usually, researchers in the social sciences can resort to methodological standards and frameworks - developed formally or informally within methodological disciplines - such as those used in psychometric (American Educational Research Association, 2011; Clark & Watson, 2019; Cronbach, 1951) or content analysis research (Krippendorff, 2018). In contrast, as Baden et al. (2021) argue, computational text analysis lacks comparable guidance, resulting in a

disconnect between theoretically grounded validation practices and the rapid, sometimes uncritical, deployment of large-scale computational text analysis methods. This gap, ultimately, creates ambiguity about how validation should be conducted, reported, and assessed in empirical research (Birkenmaier et al., 2023a).

1.2 The present dissertation

1.2.1 Research objectives

How can researchers ensure valid CTAM measures of social constructs? This guiding question is at the heart of this dissertation by addresses foundational issues of measurement validity and offering both conceptual and practical guidance for improving CTAM measures of social constructs.

As such, I define three distinct research objectives:

1. **Understanding contemporary practices in CTAM validation:** While validity is a core principle of empirical research, its meaning, scope, and application are often underdefined in computational text analysis. The first research objective of the dissertation is, therefore, to map current practices and to empirically identify challenges for CTAM validation.
2. **Developing conceptual and practical guidance for CTAM validation:** Building on the foundation of the first research objective, the dissertation then seeks to develop a clear and transferable conceptualization of validity that is grounded in the philosophy of measurement within the social sciences. At its core, this includes conceptual guidance for researchers, the introduction of shared terminology, but also practical guidance and tools.
3. **Demonstrating the usefulness of the framework for improving validation in empirical research:** The third research objective is to empirically demonstrate how the proposed validation framework can enhance practical validation of computational text analysis.

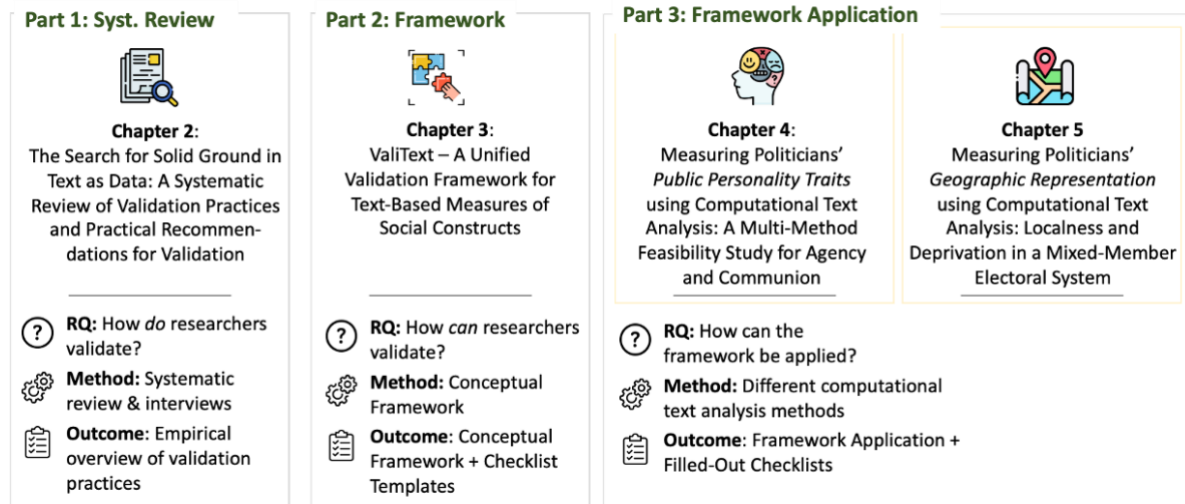
Taken together, these contributions offer both a theoretical foundation and an applied roadmap for more valid, transparent, and interdisciplinary computational text analysis.

1.2.2 Contributions and Outline

Recognizing the research objectives, this dissertation aims to advance both conceptual understanding and practical guidance for validation in computational text analysis

methods. Guided by the overarching question of how researchers can develop valid CTAM measures of social constructs, Figure 2 outlines the empirical contribution of this dissertation in greater detail:

Figure 2: Overview of Empirical Contributions



Part 1: Systematic Review (Chapter 2): The first part of the dissertation systematically investigates how researchers validate computational text analysis methods in applied settings. It addresses the guiding research question: How *do* researchers validate? By reviewing existing validation practices and conducting expert interviews, this part of the dissertation provides an empirical overview of common approaches and gaps in current validation efforts.

Part 2: Framework (Chapter 3): Building on insights from the systematic review, the second part of the dissertation develops a unified validation framework designed to guide the validation for CTAM measures of social constructs. It addresses the guiding research question: How *can* researchers *validate*? This part of the dissertation presents the theoretical underpinnings of the framework and offers practical templates and checklists to guide researchers in planning and documenting validation procedures.

Part 3: Framework Application (Chapters 4 & 5): The third part of the dissertation demonstrates the application of the framework in two distinct research contexts that involve the measurement of politicians communication. It addresses the central research question: How *can* the *framework be applied* to guide measurement validation? Focusing on the field of political communication, this part shows how the framework can be adapted to suit different computational text analysis tasks and how it can support validation.

On the one hand, Chapter 4 presents a multi-method validation study that validates measures of **politicians' public personalities** operationalized through the presence of personality cues in their public communication. In this study, I develop and validate fine-grained measures of public personality for the traits of *agency* and *communion*. This involves using various models - such as RoBERTa, Deepseek-V3, and GPT-4o - and applying multiple validation techniques, including comparisons with human annotations, the design of systematic tests, and sensitivity analyses of public personality in the statements of German politicians and in the 2024 U.S. presidential debate between Donald Trump and Kamala Harris.

On the other hand, Chapter 5 presents a study that highlights how the framework can be applied to validate measures of politicians' **geographic representation** through the frequency and context with which they reference locations within their constituencies. Rather than following a conventional construct classification pipeline, this study focuses on validating location references, primarily using named entity recognition and data linkage techniques.

Chapter 6 synthesizes the main empirical findings, discusses their theoretical and practical implications, critically reflects on the study's limitations, and concludes with concluding remarks that bring the dissertation to a close.

1.2.3 Database

This dissertation utilizes a range of data sources across its empirical chapters.

Chapter 2 draws from a systematic literature review of reported validation steps (i.e., specific tests to produce evidence for the validity of empirical measures). Beginning with an initial literature search that resulted in 920 publications, we further limited the publications using inclusion criteria and ultimately coded 96 peer-reviewed articles in the field of political communication for different characteristics, such as the method applied and construct targeted, as well as the validation steps that were reported. In addition to the literature review, we carried out eight semi-structured expert interviews to identify further activities researchers may engage in during CTAM validation.

Chapter 3 is conceptual in nature and does not draw on any original empirical data. Instead, it develops a theoretical framework based on the literature reviewed in Chapter 3.

Chapter 4 draws on a diverse set of public statements from 46 leading German politicians covering the period from 2017 to 2024. Three data sources are included: First, a novel collection of 99 public broadcast interview transcripts across five major German media formats (Birkenmaier, Sieber, et al., 2025). Second, a dataset of all public Twitter and Facebook posts for these 46 politicians.¹² Third, transcripts of parliamentary speeches held in the German Bundestag by these 46 politicians. In addition to the corpus of German politicians, in the validation process, we also examine a corpus of around 2,000 translated statements from the 2024 U.S. presidential debate between Donald Trump and Kamala Harris.

Chapter 5 uses a dataset of all public Twitter and Facebook posts for 364 German politicians that were members of the German Bundestag between 2017 and 2024. In addition, to derive the measure of deprivation, it relies on data from the Point of Interest Bund (POI-Bund)

1.2.4 Publications

Apart from references to previous studies, this work only includes research results to which I contributed significantly. It is based on the following list of publications:

- **Birkenmaier, L.**, Lechner, C. M., & Wagner, C. (2023). The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. *Communication Methods and Measures*, 1–29. <https://doi.org/10.1080/19312458.2023.2285765>
- **Birkenmaier, L.**, Wagner, C., & Lechner, C. (2025). ValiText--a unified validation framework for CTAM measures of social constructs. arXiv preprint arXiv:2307.02863. <https://doi.org/10.48550/arXiv.2307.02863>
- **Birkenmaier, L.**, & Lechner, C. (2025). Measuring Politicians' Public Personality Traits using Computational Text Analysis: A Multi-Method Feasibility Study for Agency and Communion https://doi.org/10.31219/osf.io/7nhjz_v1
- **Birkenmaier, L.**, Sieber, L., & Bergstein, F. (2025). PolInterviews--A Dataset of German Politician Public Broadcast Interviews. arXiv preprint arXiv:2501.04484. <https://doi.org/10.48550/arXiv.2501.04484>
- **Birkenmaier, L.**, Stroppe, A.-K., Wurthmann, C. L., & Sältzer, M. (2025). Between Home Turf and Hinterland: Directly Elected MPs Focus More on Local and Deprived Places than List Candidates on Social Media. Working Paper. Currently under review in *Research & Politics*

¹² Since the Twitter data used in this dissertation was collected prior to the platform's rebranding to "X," I continue to refer to it as Twitter data throughout this dissertation.

- Repke, L., **Birkenmaier, L.**, & Lechner, C. (2024). Validity in Survey Research-from Research Design to Measurement Instruments. GESIS Survey Guides https://doi.org/10.15465/gesis-sg_en_048

Additionally, the following publications and preprints are related to this thesis.

- Daikeler, J., Fröhling, L., Sen, I., **Birkenmaier, L.**, Gummer, T., Schwalbach, J., Silber, H., Weiß, B., Weller, K., & Lechner, C. (2024). Assessing Data Quality in the Age of Digital Social Research: A Systematic Review. Social Science Computer Review, o(o). <https://doi.org/10.1177/08944393241245395>
- Lechner, C., Aichholzer, J., **Birkenmaier, L.**, Bluemke, M., & Lechner, C. (2025). Mapping the Political Personality (PoliPers): An Integrative Framework for Assessing Ideal and Actual Personality Profiles in Politicians. Preprint <https://doi.org/10.31234/osf.io/w42pj>
- **Birkenmaier, L.**, Daikeler, J., Fröhling, L., Gummer, T., Lechner, C., Lux, V., ... & Ziaja, S. (2024). Defining and Evaluating Data Quality for the Social Sciences: Position Paper. (GESIS Papers, 2024/06). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften. <https://doi.org/10.21241/ssoar.96764>

2. Understanding Contemporary Practices in CTAM Validation

The first part of the dissertation systematically examines how CTAM measure validation is currently practiced in social science research. It addresses the guiding research question: How do researchers validate their measurements? To answer this, I subsequently conduct a systematic review in the field of political communication research that maps existing approaches to validation and identifies recurring patterns, strengths, and shortcomings. In addition, I conduct expert interviews to provide deeper insights into how researchers understand and implement validation in practice. Taken together, this part of the dissertation offers a comprehensive overview of contemporary validation practices and highlights key challenges and gaps that need to be addressed to strengthen methodological rigor.

This chapter is based on the following publication: *Birkenmaier, L., Lechner, C. M., & Wagner, C. (2023). The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. Communication Methods and Measures, 1–29.*
<https://doi.org/10.1080/19312458.2023.2285765>

2.1 Introduction

Over the past two decades, using text as data has become an increasingly popular approach in communication research and continues to evolve rapidly. The term “text as data” refers to the application of computational text analysis methods (CTAM) for analyzing written or spoken language as a source of data in research. Whereas the term CTAM encompasses diverse methods and applications, it commonly entails the use of algorithms and automated software tools to analyze texts, varying in the extent of human supervision required (Grimmer & Stewart, 2013). In communication research, CTAM are usually applied to study and operationalize social science constructs, namely, abstract and theoretical concepts which are not directly observable, such as emotions, audience frames, or ideology (Krippendorff, 2018).

One of the abiding challenges for the application of CTAM, however, is to ensure the validity of CTAM measures, that is, the extent to which a measure accurately reflects the construct or phenomenon it is intended to measure (Adcock & Collier,

2001; Kelley, 1927; King et al., 1994). There are several threats to the validity of text-based measures. For example, the meaning of words and phrases can depend on the context in which they are used, and this context can be difficult to capture and interpret. In addition, text data can be noisy and ambiguous, with multiple meanings or interpretations possible for a given word or phrase (Boxman-Shabtai, 2020; Krippendorff, 2018). To address these challenges, researchers have developed a plethora of validation approaches to demonstrate that their empirical measures, nonetheless, validly operationalize the constructs of interest. So far, however, methodological guidance for researchers as to how they should validate their text-based measures is fragmented, and lacks conceptual clarity as well as a commonly shared terminology (Baden et al., 2021; Grimmer et al., 2022). Therefore, researchers are often faced with ambiguity on when and how validation should be conducted, as well as determining criteria that indicate the effectiveness of well-executed validation. We submit that without sufficient attention paid to measurement validity as the foundation for empirical research, communication research will not be able to reap the full potential offered by the increasingly sophisticated CTAM.

In this paper, we aim to lay the groundwork for a more principled and transparent approach to validation for text as data. Our paper comprises a descriptive and a normative part. In the descriptive part, we take stock of current validation practices to provide an empirical foundation for a more systematic engagement with validity and validation for CTAM. Toward that end, we carry out a systematic review of reported validation steps (i.e., specific tests to produce evidence for the validity of empirical measures) in the field of political communication. To enhance our understanding beyond the scope of information available in published research, we further carry out expert interviews to explore any additional activities that researchers undertake during CTAM validation. In the normative part, based on this empirical overview, we then derive practical recommendations and provide guidance on how researchers should approach validation conceptually. Additionally, we offer a glimpse into present and upcoming validation frameworks that may prove advantageous for researchers seeking guidance in CTAM validation,

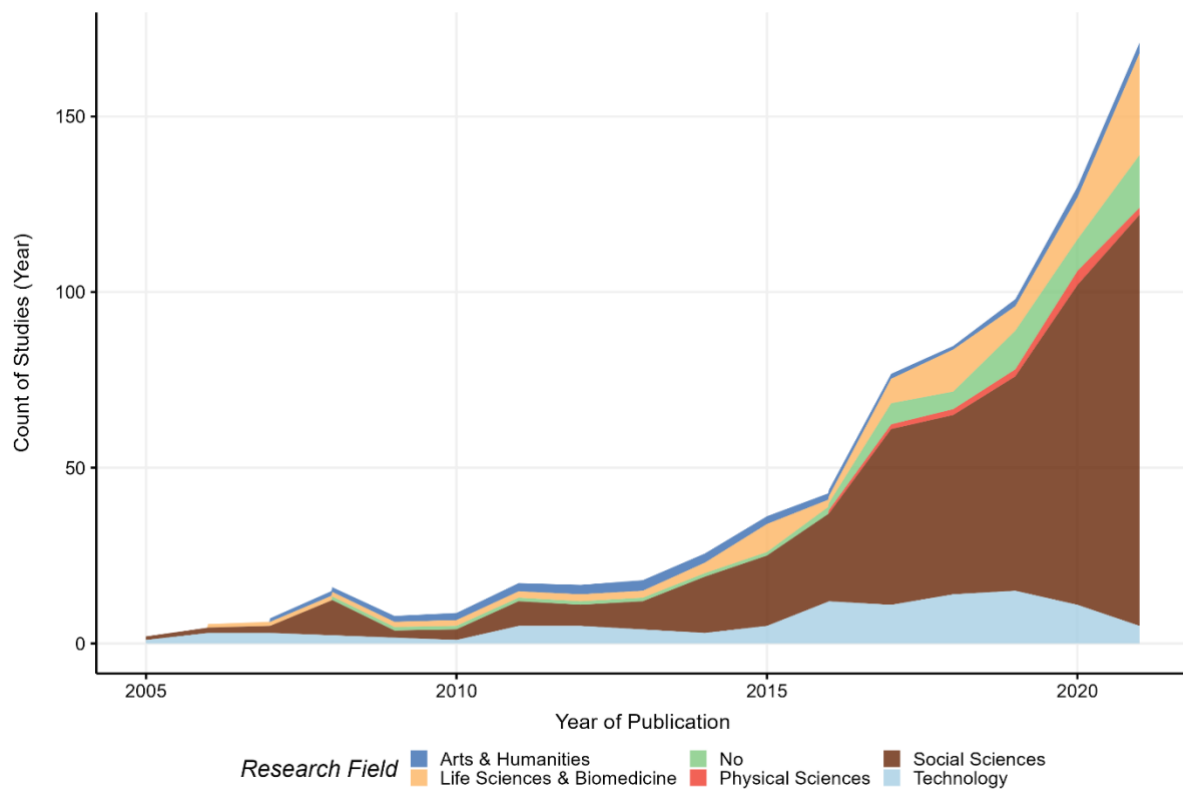
2.2 Theoretical Background

2.2.1 Promises and Pitfalls of Text as Data in Communication Research

With the advent of digital sources, such as social media, blogs, and news articles, researchers have been able to access and analyze unprecedentedly large and diverse datasets to study a wide range of phenomena in communication research (Brady, 2019; Edelmann et al., 2020). Concomitantly, advances in CTAM have made it easier for researchers to analyze and extract insights from textual data. A variety of CTAM have been developed which can be best grouped and defined according to their underlying principles. Whereas *rule-based* methods usually define a finite set of rules to process and classify texts into known categories, *supervised* methods allow a text model to autonomously develop ways to predict known output categories. *Unsupervised* methods, finally, are the least restrictive type of CTAM, allowing the model to derive its own output categories based on observed patterns in the data (Baden et al., 2021).

Towing to the variety of applications and research contexts, CTAM have become an increasingly important tool for studying communication (see Figure 3). For example, previous research utilized CTAM to study integrative complexity of online discussions (Dobbrick et al., 2021), media coverage and news frames (Baden & Tenenboim-Weinblatt, 2017; Eisele et al., 2023), or communication styles of political actors (Mueller & Saeltzer, 2020; Rudkowsky et al., 2018; Stier et al., 2018).

Figure 3: Text as Data Publications within the Web of Science Database



Notwithstanding the promising opportunities offered by computational methods, measurement validity and the operationalization of complex latent constructs raise significant concerns within political communication research (Baden et al., 2021). These concerns are a direct result of the complexities involved in quantifying social science constructs through textual data. Primarily, this includes the latent nature of social science constructs that may manifest in textual data in various forms. Because textual data is characterized by high dimensionality and complex grammatical as well as syntactic structures, it is impossible to make specific and definite statements about all predictable connections that exist between words in a text (Yeomans, 2021). Consequently, even human coders usually do not arrive at identical interpretations of the same text. Song, Tolochko, et al. (2020a), for instance, showed that “gold standard” human-annotated labels can suffer from profound inconsistencies, frequently resulting in low levels of inter-rater reliability between human coders. The methodological literature on qualitative content analysis provides extensive documentation on potential pitfalls, such as meaning multiplicity and polysemy (Boxman-Shabtai, 2020; Ceccarelli, 1998), or the importance of context (Mayring, 2004; Stemler, 2000).

The application of computational methods often comes with additional challenges for validation. These challenges hail from the design of computational methods that make implicit or explicit assumptions about how constructs manifest in text. For instance, dictionaries often build on drastically simplified assumptions about the structure of texts and ignore all sort of relevant textual information, such as sentence structure or word order (“*bag of words*”) (Lowe & Benoit, 2013). More complex models, in particular large language models (LLM) (see Devlin et al., 2019), promise to solve some of these limitations. However, “black box” LLMs also come with new methodological challenges for validation. Among others, prominent problems include data leakage (Gibney, 2022; Kapoor & Narayanan, 2022) or systematic biases in the training data, such as unrepresentative datasets, or model-inherent human stereotypes (van Giffen et al., 2022).

2.2.2 Contemporary Approaches to CTAM Validation

Thorough validation, involving both theoretical reasoning and empirical evidence, is an essential requirement when applying CTAM in empirical research. Validation ensures that what is meant to be measured in the texts corresponds accurately to the actual measurement. Likewise, validation helps researchers to minimize the uncertainty inherent in any empirical measurement (see for instance, Flake et al., 2017). Consequently, there should be no question regarding the overall necessity of CTAM validation. Thus, Grimmer and Stewart (2013) underscored the need for researchers to consistently “validate, validate, validate” (p. 271) in their work. When reviewing the literature, however, determining *how* to validate is often unclear. This unclarity stems from the fact that conceptual and practical guidance on how to validate CTAM measures is scarce and fragmented, prompting scholars to seek guidance from other research fields.

On the one hand, researchers can refer to the general methodological literature on validity in social and behavioral science research. Various general validation frameworks exist in subdisciplines such as psychology (American Educational Research Association, 2011; Flake et al., 2017), political science (Adcock & Collier, 2001; Goertz, 2008), or survey research (Lewis-Beck et al., 2003; Rammstedt et al., 2015). However, guidelines on validation between these subdisciplines vary significantly. Even more important, existing validation frameworks mostly emerged in the context of research using survey and assessment data and, as such, are typically not

tailored to the unique challenges associated with computational text analysis. Furthermore, the terminology used to describe particular steps of validation can exhibit substantial variations.¹³

For more specific guidance to text analysis, scholars can turn to the literature on qualitative content analysis, which has a long history of measuring social science constructs from text.¹⁴ However, the methodological literature on content analysis usually neglects the incorporation of computational methods, instead concentrating narrowly on human coding to conduct text analysis.

Therefore, in the absence of a comprehensive validation framework for CTAM, researchers often adopt approaches used by similar studies, or conduct validation using a mixture of different validation steps (Goet, 2019; Grimmer et al., 2022; Quinn et al., 2010). Subsequently, we outline some major conceptual distinction that are commonly used, albeit implicitly, to classify different types of validation steps. However, it is important to acknowledge that these distinctions are neither precisely defined nor exhaustive, primarily due to the lack of overall conceptual clarity in the field.

2.2.2.1 Internal and External Validation

One key conceptual distinction can be made between internal and external validation (Birkenmaier et al., 2023b; Grimmer & Stewart, 2013; Quinn et al., 2010). Broadly speaking, internal validation relies on common knowledge or domain-specific tests to demonstrate that the model and its measures appear plausible, while external validation seeks to compare the obtained measures with external data. Hence, for internal validation, validation steps typically involve various tests across the measurement process which rely heavily on the researcher's judgment, with subtypes being *face validity*, *statistical validity*, *content validity*, *semantic validity*, or *construct validity*¹⁵ (DiMaggio et al., 2013; Grimmer & Stewart, 2013). External

¹³ For example, in the psychometric tradition, validity is often associated with the overarching concept of “construct validity” which requires the collection of different types of evidence that support the validity of the construct (Cronbach & Meehl, 1955; Messick, 1995). In the field of political science, Adcock and Collier (2001) suggest testing for three types of measurement validation, where “construct validation” is just one type alongside “content validation” (assessing the adequacy of the measure and its content) and “convergent/discriminant validation” (determining its convergence with related or unrelated measures).

¹⁴ The term *semantic validity*, for instance, originates from the literature on content analysis and is usually used to refer to the “extent to which each category or document has a coherent meaning” (Quinn et al., 2010) or “[being] semantically coherent” (Greene & Cross, 2017).

¹⁵ Due to the lack of clearly defined terminology, some subcategories mentioned for internal validation can also represent external validation for different interpretations (see *construct validity* for Adcock and Collier (2001)).

validation steps, on the other hand, usually aim to compare the obtained measures with independent information or exogenous events, with subtypes being *convergent validity*, *discriminant validity*, *concurrent validity*, *predictive validity*, or *hypothesis validity* (for a more detailed explanation, see Quinn et al., 2010). Thus, the strength of external validation lies in its ability to showcase the performance of CTAM on “out-of-sample” data, that is, independent of the data used to adjust and train the CTAM (DiMaggio, 2015). Therefore, external validation is usually determined post-hoc, that is, by inspecting the CTAM scores and comparing it with some form of external information after the measurement has been conducted.

2.2.2.2 Gold-Standard and Non-Gold-Standard Validation

Another key conceptual distinction can be made between validation steps that involve gold-standard data and those that do not (Grimmer et al., 2022). Gold-standard data is typically meticulously hand-coded and presumed to be completely accurate and objective, although this assumption may not always hold in practical settings. Consequently, gold-standard data is often seen as the ultimate benchmark, whereas validation steps that do not include gold-standard data are considered less meaningful and require more reasoning by the researcher (Song, Tolochko, et al., 2020a).

2.2.2.3 Generic or Method-Specific Validation

Furthermore, one can also distinguish between validation steps that are universally applicable, and validation steps that are only eligible for specific types of methods (Birkenmaier et al., 2023b; Grimmer & Stewart, 2013). The literature on unsupervised methods, for instance, proposes a great variety of metrics and validation steps to demonstrate the consistency of topics for specific variants of topic models (Chan & Sältzer, 2020; Chang et al., 2009; Ying et al., 2022). On the other hand, there are generic validation steps that can be universally applied to different types of CTAM. For example, these steps may include visualizing the model output or evaluating the correspondence of measures with human-annotated data.

2.2.3 On the Need for a more Comprehensive Understanding of Validation

Given the heterogeneity of research contexts and validation steps available, it appears natural that researchers need to adapt their validation approaches and cannot follow a one-size-fits-all solution. Although using different validation strategies for different

research designs is not problematic per se, it is, however, essential to arrive at conceptual clarity and follow a shared terminology when it comes to validation. Additionally, it's vital to have a clear understanding of the minimal requirements for validation. As outlined in the previous chapter, this is especially true for the question of *how* to validate CTAM.

In practice, though, the lack of clarity in validation often leads to confusion, to the extent that validation choices may seem arbitrary and hard to understand (Baden et al., 2021). This is further compromised by the absence of any clear distinction regarding the various forms of validation that are available to researchers. Likewise, one can find significant discrepancies between reporting practices and, if applicable, ways of providing reproducibility materials. Clearly, the ambiguity around validation poses a problem for researchers, who can easily lose track of how they should validate CTAM in their substantive research projects. Our core argument therefore is that we need to take a more systematic perspective on CTAM validation. This is in line with other researchers claiming that more unifying validation efforts are needed to show “exactly how and how convincingly [CTAM] operationalize relevant conceptual properties” (Baden et al., 2021, p. 14). To do so, however, we first need an encompassing understanding of current differences and practices in the field of CTAM validation.

2.2.4 Our Contribution

This paper contributes to closing this methodological research gap by tacking stock of current practices of CTAM validation. To do so, we pursue a twofold approach. First, we evaluate how studies that apply CTAM approach the issue of validation and identify and categorize the main validation steps they take. To do so, we conduct a systematic review of validation practices in peer-review publications from the field of political communication. Specifically, we aim to answer the question of *how* researchers validate CTAM. Second, we conduct expert interviews with scholars in the field of text analysis to gain insights on more subtle and unreported validation activities. Based on the assumption that researchers might conduct more extensive validation that, however, may at times go unreported, the interviews aim to complement and contextualize the information garnered from the systematic review. Together, we thus aim to obtain a comprehensive picture of the current state of CTAM validation that will enable us to 1) map the current field of CTAM validation and 2) derive normative

statements and practical recommendations for applied researchers that can lay the groundwork for a unified understanding on CTAM validation.

2.3 Research Design

2.3.1 Systematic Review

2.3.1.1 Literature Search

Any systematic review requires a clear focus. We deemed political communication a very apt case in point for a review of validation practices for two main reasons. First, this subfield has had a pioneering role in the application of CTAM for communication research (Theocharis & Jungherr, 2021). Second, political communication constitutes a highly multidisciplinary research field within communication research (Tenenboim-Weinblatt & Lee, 2020), drawing heavily from different domains of communication research (Song, Eberl, et al., 2020; Waisbord, 2019). Thus, although our systematic review will focus on political communication as a case in point, we expect that the validation practices identified in our review should align with similar practices in other communication research domains. This is because the lack of a unified approach to CTAM validation and even a shared terminology related to validation is a common challenge across all subfields, as highlighted by Grimmer et al. (2022). Therefore, we argue the conclusions and implications derived from our review will also be instructive for researchers from other subdisciplines of communication research, indeed to any research using text as data.

To start, we identified the five most cited peer-reviewed journals in both the fields of communication research and political science research based on the Scimago Journal Ranking (Scimago, 2022).¹⁶ Given the innovative character of many CTAM publications, we further identified six journals based on their relevance of publishing high-quality CTAM publications which were not listed in the initial journal list. The selection of these additional journals was further confirmed by thorough discussions with domain experts. Thus, we identified a total of 16 relevant journals (see Appendix 1 for a detailed justification and a complete list of journals assessed).

To search for relevant publications within these journals, we first relied on a naïve search strategy using a simple search string ((Politi* OR Party OR Govern*) AND

¹⁶ To conduct the review, we use a systematic strategy to transparently identify publications which rely on CTAM in social science research (Durlak & Lipsey, 1991). To do so, we rely on the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA), a systematic search and evaluation procedure for academic literature (Liberati et al., 2009).

text*). In a second step, we then added several keywords using the LitsearchR package, an R package to facilitate quasi-automatic search strategy development (Grames et al., 2019). To verify the effectiveness of our search string, we compared the results obtained from our literature search with a curated list of exemplary articles that we knew should fall in the scope of the review. These exemplary articles, which are listed in Appendix 2, were thoughtfully selected by the primary author from different journals and subdisciplines in political communication research. We successfully located all these articles in our preliminary literature list. As a result, the final search string contained several pruned terms related to text-based methods (e.g., “machi* learni*” OR “sentim* analys* OR text*”) and the research field of political communication (e.g., socia* medi* OR “parliament* record*”) (see Appendix 3 for the complete search string). As we wanted to describe only current validation, we further limited our search to studies published after 2014. Ultimately, this resulted in a preliminary literature list of 920 publications.

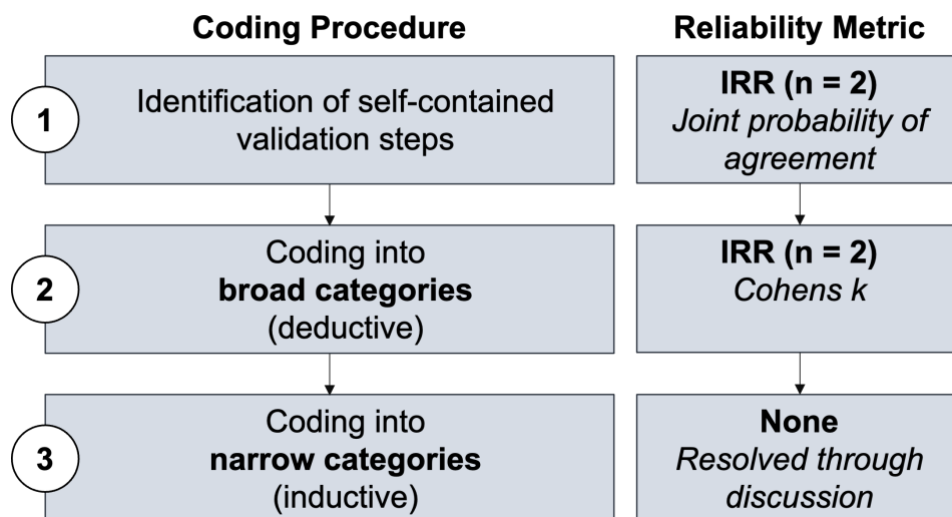
Next, we examined the title, abstract and keywords of all studies identified and limited our analysis based on a clear set of eligibility criteria. Specifically, we stipulated that the studies included in our review should apply a CTAM on a corpus of textual data to measure at least one latent social science construct. To capture the complete landscape of text analysis research, we deliberately adopted an inclusive definition on what constitutes a social-science construct, ranging from rather abstract (e.g., sentiment or communication style) to multidimensional (e.g., populism or ideology) constructs. As another eligibility criteria, we stipulated that only studies with a substantive research focus were eligible, meaning that the goal of a study as stated by its authors should be to answer an empirical and theory driven research question in the field of political communication by means of CTAM. These eligibility criteria narrowed down the number of relevant studies from 920 to 96. A cursory examination of the excluded studies showed that these studies primarily utilized qualitative approaches relying on hand-annotation, extensively discussed general characteristics of CTAM and their implication for the broader research field or focused solely on introducing or adapting a specific type of CTAM, or deviated from our political communication focus (e.g., by studying argumentation patterns in judicial documents).

2.3.2 Coding Procedure

2.3.2.1 Identification of Validation Steps

Figure 4 visualises the stepwise coding procedure for the 96 eligible studies we retrieved.¹⁷

Figure 4: Coding Procedure



For step 1, we identified, for each study, sections of the article which reported any kind of validation step. We defined a validation step as a complete and self-contained validation activity, such as comparing the output of a CTAM with human-annotated scores or using the CTAM scores to predict an external criterion. In order to faithfully represent how researchers described their validation, we coded only those validation steps that researchers themselves explicitly identified as a form of validation. To do so, we screened the entire article for sections containing evidence presented with the intent of validating the CTAM (stated explicitly, although perhaps not using the exact terminology such as “validity”, “validation”, or “validate”). Most of the times, sections presenting validity evidence were located before or after the presentation of empirical results and contained terms such as “validity” or “validation,” or synonyms such as “trustworthiness of results”, “evaluation of model performance”, or “reviewing the results”. Whenever the authors of the respective study explicitly referenced validation in supplementary materials, we also inspected and coded the validation steps reported therein. For each validation step, we documented the whole section in the manuscript and how its authors themselves reported the validation step. To guarantee the

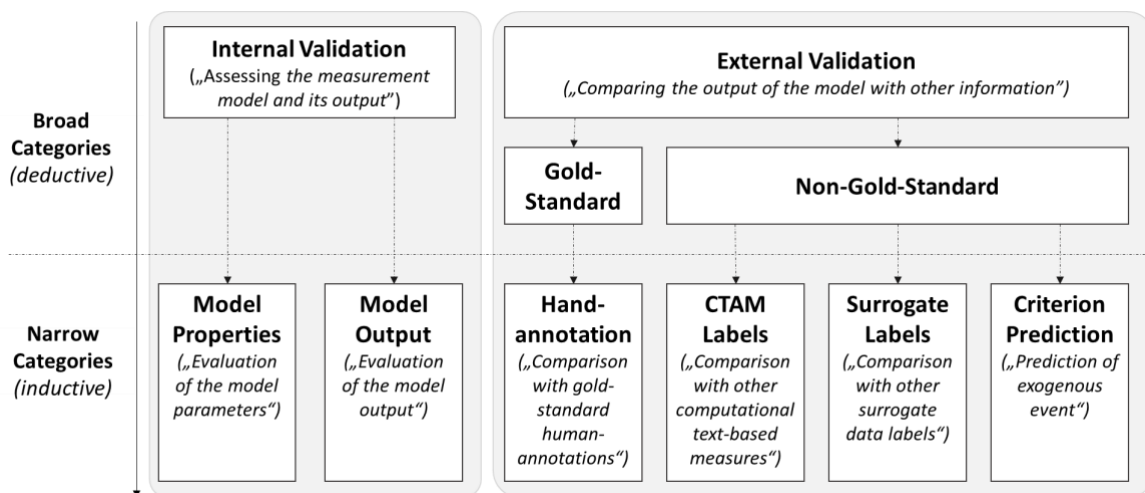
¹⁷ All data used for the review, including a list of all studies included in the review (n = 96) and a list of all coded validation steps, is available from <https://figshare.com/s/7418783cfa9f75c984f8?file=39635053> (see Appendix 4). Furthermore, an overview of the number of studies per journal is displayed in Appendix 5.

consistency in identifying sections containing self-contained validation steps, two coders independently coded a 20 percent subset of the studies, achieving a satisfactory level of concurrence in identifying the validation steps within the text (with a sufficient agreement rate of 82%).

2.3.2.2 Developing and Applying a Coding Scheme for Different Types of Validation Steps

After documenting the validation steps in each study (step 1), we coded each validation step according to its validation type (step 2 and 3). Because there is no widely shared taxonomy of different types of validation steps in the research community, we developed our own coding scheme based on the evaluation of previous validation approaches (see our chapter on “Deciding How to Validate” in the literature section). The complete coding scheme is visualised in Figure 5, and the process of refining it is outlined in greater detail below.

Figure 5: Coding Scheme



We started developing our coding scheme by defining broad categories for internal (i.e., validation steps that systematically evaluated the model and its output) and external (i.e., validation steps that focus on comparing the output of the model with other information) validation. We primarily distinguished between internal and external validation since both categories share unique core philosophies—namely, assessing the measurement model and its output (internal validation) and comparing the output with external information (external validation) (see e.g., D. Maier et al., 2021; Quinn et al., 2010). Based on this initial assessment, we further divided external validation into *comparison with gold-standard human-annotated data* and *non-gold-standard data*. As outlined in Grimmer et al. (2022), there is an in-depth

exploration of the importance of gold-standard data, and its inclusion or absence forms a critical component in the types of validation steps available. In the case that a validation step was not clearly assignable into one of these broad categories, we also provided an open category for further inspection.

After a first coding round, we then took a more inductive approach and, based on the literature, subdivided and structured the coded validation steps into more narrow categories. This was supported by an in-depth discussion between the authors where the goal was to map applied validation strategies and to group them into related categories of validation steps. As a result, we subdivided internal validation into two sub-categories: *model properties* (i.e., evaluation of model parameters and characteristics) and *model output* (i.e., evaluation of the output measures). Additionally, for external validation, we established more detailed categories for *non-gold standard data*, including *CTAM labels* (alternative text-based measures), *surrogate labels* (other labels associated with the textual data), and *external criteria* (external criteria unrelated to the textual data itself) (see Quinn et al., 2010).

2.3.3 Expert Interviews

In addition to the systematic review, we also conducted semi-structured interviews (Helfferich, 2022) with a subset of the researchers whose work was included in our review. To ensure their substantial familiarity with CTAM, we specifically targeted these researchers who had a minimum of two CTAM publications in our literature review, amounting to 45 individuals. Based on this initial list, we then applied a purposive sampling strategy and selected a total of eight interview partners (Robinson, 2014). Our goal was to select a diverse group of experts to encompass a range of perspectives and experiences on validation. Our selection criteria focused on occupational status (including both senior and junior scholars), gender (including both male and female scholars), and regional background (including scholars from different countries and universities). Ultimately, the first author of this manuscript conducted the eight interviews either face-to-face or via Microsoft Teams, lasting between 20 and 40 min (see Appendix 6 for the complete questionnaire). The recorded audio files were then transcribed using the transformer-based language model *Whisper* (Radford et al., n.d.) and subsequently manually corrected. Interviewees were ensured anonymity by declaring that they would remain anonymous, and their identity would not be revealed to any third parties or people outside the research project.

2.4 Results

2.4.1 Systematic Review

In the following, we present results from our systematic review. We begin with a summary of basic study characteristics, followed by a detailed description of when and how scholars reported validation.

2.4.2 Basic Study Characteristics

Figure 6 summarises the key study characteristics.

Figure 6: Study Characteristics

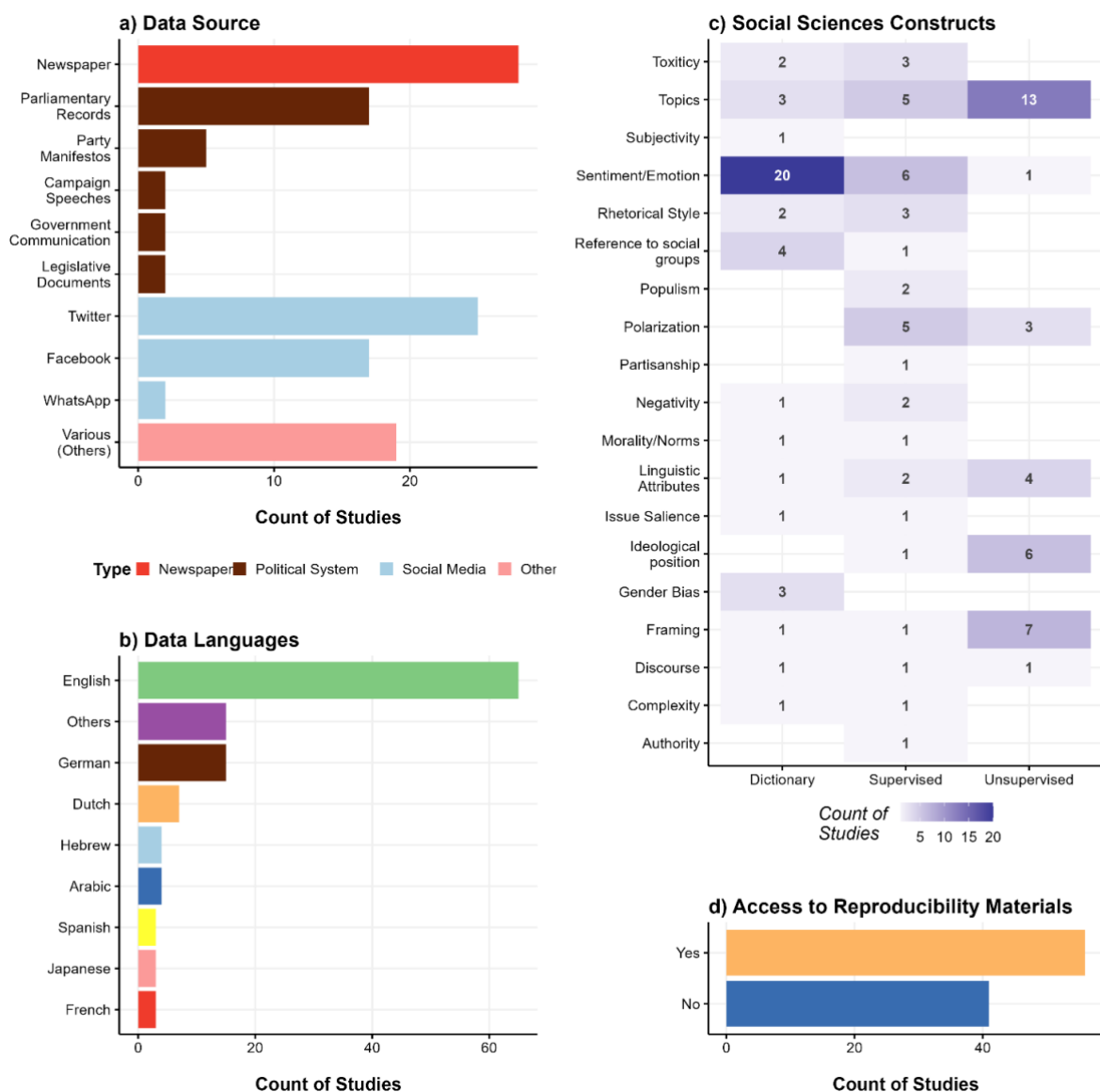


Figure 6a displays the origin of the most prominent data sources, clustered into major categories. Newspaper data constituted the most prominent data source analyzed ($n = 28$), followed by Twitter data ($n = 25$), parliamentary records ($n = 17$), and Facebook data ($n = 16$). Furthermore, researchers also worked with textual output of the political system, such as party manifestos, campaign speeches, governmental press releases or legislative documents.

Figure 6b shows the languages of the data sources used. The most dominant language present in the data was English ($n = 65$). This is in line with earlier findings on the dominance of English corpora in textual research (Pang & Lee, 2008). However, other languages were assessed as well, either as part of multilingual projects (e.g., de Leeuw et al., 2020; Maier et al., 2022) or in single-language studies usually focusing on a specific country or region (Bustikova et al., 2020; El-Masri et al., 2021; Yang & Fang, 2021).

Figure 6c provides information on the distribution of topics across the three main types of CTAM, namely dictionaries, supervised, and unsupervised models. Generally, all three types of CTAM were widely applied, with dictionaries being most commonly used (43,8%), followed by supervised (39,6%) and unsupervised methods (35,4%).¹⁸ Turning to the target constructs measured through CTAM in these studies, one could observe a great variety that reflects the diversity of research topics in the field of political communication. The most popular constructs were sentiments and emotional attitudes ($n = 27$), topics ($n = 21$), and political frames ($n = 9$). However, the distribution of research interests varied significantly between CTAM types. In the case of dictionaries, the great majority of studies were interested in sentiments and emotions, whereas unsupervised methods were commonly used to assess topics, political frames, or ideological positions. For supervised models, constructs of interest were more evenly distributed, ranging from issues such as polarization, sentiment, emotion, rhetorical style, populism, or negativity. This indicates that supervised methods are commonly used for the measurement of different constructs, which is not surprising since supervised methods can be easily trained in cases for which researchers have access to labelled data.

Lastly, Figure 6d provides information on whether studies provided access to reproducibility materials, either code or data. Overall, 41 out of 96 studies did not

¹⁸ Note: Percentages do not round up to one, because studies might apply more than one method type within the same project. Furthermore, 9,38 per cent of methods were not clearly assignable, such as using “black box” APIs ($n = 3$), word embeddings ($n = 3$) or “others” ($n = 2$).

report any reproducibility materials, which is worrying despite clear recommendations for good scientific practice within communication research (B. N. Bakker et al., 2021; Dienlin et al., 2021; Humphreys et al., 2013).

2.4.3 *When Did Researchers Validate?*

Next, we turn to the assessment of validation practices. To do so, we first evaluate *when* researchers reported to validate CTAM by plotting the total number of validation steps across studies.

Describing the number of validation steps per study is important for two reasons. First, a higher number of validation steps should serve as a proxy for the importance that the authors of a study place on validation, in particular for these instances where validation is omitted. Second, the quantity of validation steps likely corresponds to the quality of the validation procedure, as more validation evidence reduces uncertainty around the validity of the measures, thereby enhancing the trustworthiness of the measures. However, it's important to note that a high count of validation steps doesn't necessarily imply a comprehensive and rigorous validation process (see Jankowski & Huber, 2022). Nevertheless, achieving sufficient validation necessitates the presence of multiple complementary validation steps as a crucial requirement.

Figure 7: Number of Validation Steps (Uni- and Bivariate)

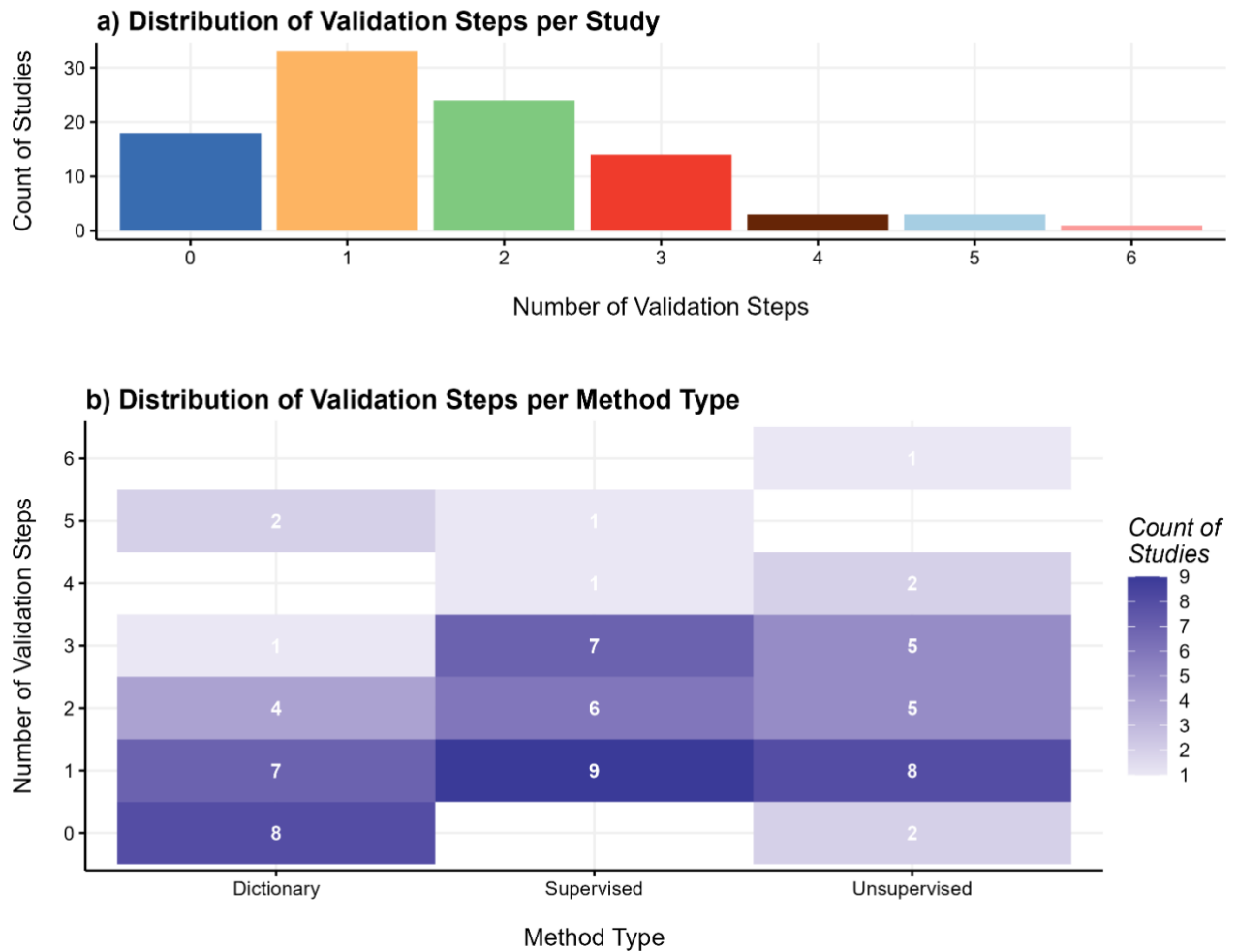


Figure 7a depicts the distribution of the number of validation steps per CTAM applied.¹⁹ The distribution is right-skewed, with most studies reporting one validation step per text-based method applied. Most studies (85,4%) reported at least one validation step (i.e., a complete and self-contained validation exercise), suggesting that researchers acknowledged the crucial role of validation in CTAM research.

Figure 6b breaks down the distribution of validation steps by CTAM type. Most studies that did not report any validation steps had applied dictionary-based methods. These findings support earlier work of Chan et al. (2021) and van Atteveldt et al. (2021), who argue that the convenience of using dictionaries is often accompanied by a lack of attention paid to validity, in that researchers often refer to previous efforts validating a dictionary, rather than presenting validation evidence themselves. For supervised

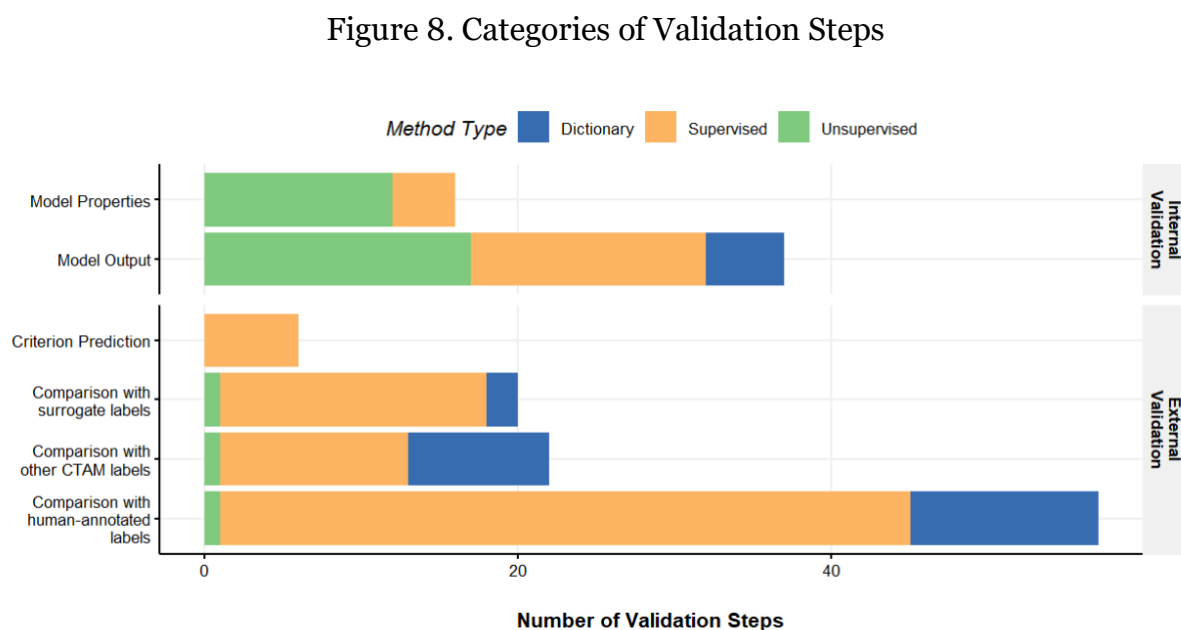
¹⁹ Because some papers applied more than one CTAM per project, Figure 5a depicts an adjusted distribution, as we divided the number of validation steps by the number of distinct CTAM applied.

methods, we counted at least one validation step for all publications included in our review. Keeping in mind that supervised methods require some form of labelled data to train the text model, these labels were also used by the researchers to evaluate the CTAM’s ability to accurately predict these labels. For unsupervised methods, the number of validation steps per publication varied the most, ranging from zero to six, indicating a great variety of validation approaches.

2.4.4 How Did Researchers Validate?

Next, we turned to the evaluation of *how* researchers validate CTAM. The first thing to stand out in the qualitative evaluation of validation steps was the absence of consistent terminology. Only 9 per cent of the validation steps in our sample of studies (15 out of 158) explicitly qualified the type of validation step they referred to, such as “semantic validity”, “face validity”, or “convergent validity”. This finding indicates that scholars may not have a shared terminology for discussing explicit validation steps. Frequently, researchers invoked validation practices to substantiate a vague notion of “validity” or “reliability of findings.”

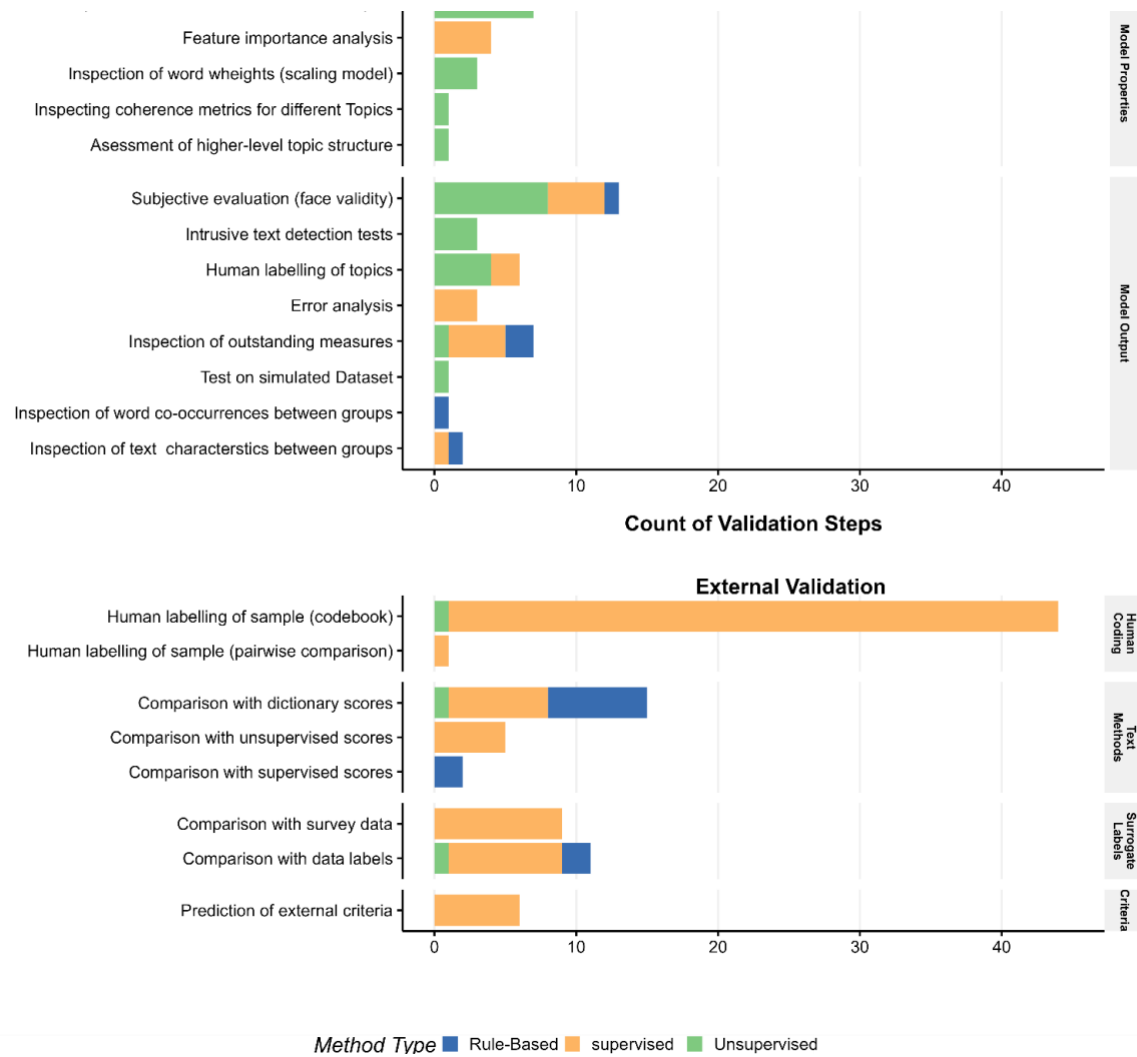
Figure 8 depicts the distribution of validation steps across different categories for each method type.



In total, 34 per cent of validation steps related to internal validation, whereas 66 per cent of validation steps related to external validation. The most common type of external validation reported was comparison with human-annotated labels ($n = 57$), followed by other text-based measures ($n = 22$), surrogate labels ($n = 19$), and the prediction of external criteria ($n = 7$). Furthermore, Figure 8 demonstrates that the selection of validation steps may be influenced by the respective CTAM type. Whereas internal validation practices were primarily adopted for unsupervised methods, supervised methods were often accompanied by external validation.

To get a better understanding of the specific types of validation steps, Figure 9 provides a more detailed overview of the same data that, however, displays the concrete validation steps which will be subsequently discussed in greater detail.

Figure 9: Overview of Specific Validation Steps



2.4.4.1 Internal Validation

For internal validation, we broadly differentiated between assessments of the *model properties* and the *model output*, respectively.

2.4.4.1.1 Model Properties

To assess the *model properties*, researchers relied on different validation steps to evaluate the model parameters and its coherence with the model assumptions. For supervised methods, for example, several researchers reported that they assessed feature importance, that is, the evaluation of how much specific words or tokens contributed to the prediction of labels ($n = 4$). For unsupervised models, evaluation of model properties included the human inspection of dominant words for each topic ($n = 7$) or the examination of individual word weights for scaling methods ($n = 3$), thus

making sure that theoretical assumptions regarding the orientation of words were fulfilled. In order to assess the consistency of the methods applied, researchers furthermore reported sporadically on quantitative model-specific metrics. In their study on political agendas setting, Greene and Cross (2017), for example, inspected various coherence metrics for different topic models and examined the higher-order topic structure using a clustering approach.

2.4.4.1.2 Model Output

To assess the *model output*, researchers reported several validation steps, which notably differed in their level of subjectivity.

On the one hand, several validation steps were based on some form of subjective output inspection. Most of all, researchers applied *face validity* ($n = 13$), which was purely argument-based and required no formal analysis involved. Usually, face validity included the visual inspection of model outputs to evaluate, for example, the stability and general trends of measures across time or across groups. On a slightly different note, researchers also inspected texts that had extremely high or low measures ($n = 7$), compared far apart or opposite placed texts from different groups ($n = 7$), or evaluated general text characteristics ($n = 2$) and word-cooccurrences ($n = 1$) across known groups.

On the other hand, more systematic approaches were applied as well to inspect the model output. One example for this constituted intrusive texts detection ($n = 3$). This validation step was applied to make sure that texts within the same category can be distinguished from an unrelated and arbitrarily added text. Studies in our sample applied intrusive text detection for both human coders (Garbe et al., 2021; Rossiter, 2022) and CTAM themselves (L. Huang et al., 2020). Likewise, different forms of error analysis ($n = 3$) were applied as well (Dobbrick et al., 2021; Schub, 2022; van Atteveldt et al., 2021). Usually, this included the identification of patterns in the mistakes made by the CTAM with the goal of identifying systematic biases in the classification process (see Burlacu, 2021). Another approach exclusively applied for unsupervised methods was to rely on human judgement to assign meaningful labels to the most relevant words for each topic and to assess inter-rater reliability ($n = 6$).

2.4.4.2 External Validation

2.4.4.2.1 Comparison with Human-Annotated Labels

Comparison with human-annotated scores was the overall most often-reported validation step in our review. In nearly all cases ($n = 56$), this included the coding of texts via a previously defined codebook and the calculation of performance metrics such as accuracy, precision, recall, or F1 score. Usually, this also involved coding by at least two human coders and calculating the inter-rater reliability. In one case, a pairwise comparison approach (i.e., repeatedly comparing text entities in pairs) was used to label the data (J. W. Kim et al., 2021).

2.4.4.2.2 Surrogate Labels

Another form of external validation constituted the comparison of CTAM output with surrogate labels, that is, independently obtained but closely connected characteristics of the same or similar constructs (Adcock & Collier, 2001; Cronbach & Meehl, 1955). Thus, surrogate labels were usually derived from human assessed labels ($n = 9$) or contextual labels derived from the greater context of the data ($n = 11$). Starting with human-assessed labels, researchers used both respondent data and expert ratings. Whereas respondent data were typically survey responses from the same object (Parthasarathy et al., 2019; Temporão et al., 2018), expert ratings were usually derived from domain-expert knowledge, such as the left-right scores (*rile*) of the Comparative Manifesto Project for party manifestos (Lehmann et al., 2022) or the Freedom House Index for news coverage (House, 2022). On the contrary, contextual labels were usually derived from the greater context of the data. Examples of contextual labels ranged from committee (Fernandes et al., 2019; Greene & Cross, 2017; Rossiter, 2022) and party labels (Bruinsma & Gemenis, 2019; Lauderdale & Herzog, 2016), labels from legislative data or parliamentary transcripts, to mentions in newspapers (J. Gelman & Wilson, 2022).

2.4.4.2.3 Comparison with other CTAM labels

Next, comparing measures with other CTAM labels was another reported form of external validation. For most of the cases ($n = 15$), this included the application of dictionaries as a simple and easily applicable CTAM, followed by unsupervised ($n = 5$) and supervised ($n = 2$) methods. In many cases, scores from other CTAM functioned as a baseline model (e.g., Ballard et al., 2022), thereby providing evidence that the CTAM was able to produce better benchmark estimates. However, some scholars took a more interpretative approach by carefully inspecting and comparing the distributions

of different CTAM outputs to detect similarities, but also inconsistencies across measures (e.g., Rauh, 2018; Schub, 2022).

2.4.4.2.4 Prediction of external criteria

Finally, researchers also considered validation steps where the CTAM scores were used to accurately predict external criteria unrelated to the textual data itself ($n = 6$). This form of validation was especially popular for unsupervised methods, whereas no form of this validation could be found in any of the dictionary studies. Examples of external criteria predicted by the text-based measures ranged from specific events (e.g., parliament disputes based on the language used, see Gelman & Wilson (2022)) to other forms of observable behaviour, such as politicians roll-call votes (I. S. Kim et al., 2018; Lauderdale & Herzog, 2016; Rheault & Cochrane, 2020).

2.5 Expert Interviews

To further enrich and contextualize the information acquired from the systematic review, we carried out expert interviews with eight of the researchers whose work was included in our review. Given our assumption that researchers may undertake additional validation that may go unreported, the qualitative expert interviews yielded additional insights into researchers' validation practice. Subsequently, we first provide a short overview of how interviewee understand and define CTAM validation. Afterwards, we outline our findings on what additional validation steps interviewee undertake that may go unreported in published papers.

2.5.1 What Is The Purpose Of CTAM Validation

When asked to outline the purpose of CTAM validation, most interviewees emphasized that validation needs to establish a link between the constructs of interest and empirically observed measurement scores. For text analysis, this was reported to be an especially challenging task because texts were often passively collected and not designed by the researcher. One interviewee argued:

“[CTAM] are not going to give you anything beyond what's manifestly in the content of the text. And the methods are, well, they are not thinking for you, right? They're only going to give you whatever patterns they are going after”

Therefore, as one interviewee noted, all text-based measures, from simple dictionaries to LLMs, ultimately need to rely on proxies to operationalize the theoretical constructs of interest. Thus, “validation ensures a real correspondence between the meaning of

the proxy and the theoretical construct.” This was also described by two interviewees as the “fit” between the theoretical concept and the obtained measures.

Moreover, several interviewees also acknowledged that the aim of CTAM validation was to incorporate „human judgment” or „human insight” into the computational workflow. Thereby, there was a general agreement among interviewees that human judgment should not be restricted to the hand-coding of textual data but should include a more comprehensive evaluation at different stages of the measurement process. One interviewee thus described the goal of validation as ultimately being “confident enough to use [CTAM] rather than reading and coding the concepts by hand”.

2.5.2 Which Validation Steps Are Seldomly Reported

Generally, interviewees confirmed that they often carried out validation steps that, however, went unreported in the resulting papers. As a matter of fact, these unreported steps often related to internal validation across the research process, whereas reported validation steps tended to focus more on external validation. As one expert noted, the decision not to report validation steps was often driven by manuscript word limits:

“Because validating [a CTAM ca be] a bloody nightmare. There are so many small things that you can do. And writing them up always feels a little bit stupid. And you can fill an entire page, which the reviewers will tell you to take out anyway. But you definitely did that.”

On another note, interviewees bemoaned that it was often unclear to them what exactly constituted sufficient validation that they needed to report. This was connected to the general lack of concepts and guidelines for validation, a problem which was repeatably brought up in the interviews. Whereas there was an agreement among interviewees that guidelines and standards should be flexible enough to account for the heterogeneity of research contexts and text sources, the absence of a theoretical framework on validation was repeatably argued to come with profound practical problems for the interviewees who were unsure on which validation steps to report.

The major validation steps interviewees mentioned to have conducted but not reported in their work are displayed in Table 3, along with specific quotes from the interviews.

Table 3: Rarely Reported Validation Steps

Validation Step	Sample Quote from the interview
Competence Building	“What you're not going to see in the final paper about what we have done is we read some EU law textbook on how to write EU legislation [...] And then we knew that some directives [and structures] are known for being complex”
Human Identification of Concepts	“If it's completely impossible for humans to identify a concept of interest, then our text analysis approaches probably also fail. Therefore, I think it's important to start reading text first and trying to classify it manually and then you might realise that it's not possible at all or you might need broader categories, or you need to follow a different approach”
Justification of Preprocessing Decisions	“Justifying processing steps is part of validation. Like having a reason for why you remove these stop words is part of validation. And we need to think of this as part of validation, because it really affects the outcomes. [...] And it is almost the most difficult thing, because as long as all we have is a little bit of rule of thumb, seat-of-the-pants, experience values there and no real theory about why which processing steps would have what kind of implications, this is really hard to do”
Inspecting Descriptive Statistics	“So, my take is usually if there is no descriptive pattern [in the data], the project is dead basically. So, the first step should and usually is to look at differences in vocabulary between [categories], or conduct whatever exploratory analysis, [such as comparing document lengths between known groups]”
Qualitative (Error) Analysis	“So, all the measures that actually go beyond the quantitative [evaluation of model performance metrics], they normally go unreported. And if they are reported, they're only reported for the model shells. [...] because, you know, reviewers' length”
Rejection of Poorly Performing Models	“We developed also automated measures for topics but [they] seemed to not really work out so we just dropped that [...] thinking back now we could have reported that maybe in some appendix at least. But we didn't, so we just concentrated on what we thought works or what we could give some evidence that it works”

Several interviewees noted that they often *build up competencies* to critically engage with the constructs of interest prior to the conceptualization of measurement. As an example, one interviewee who studies EU legislation noted that they had engaged extensively with the technical literature on legislative writing, yet did not report doing so in the paper because he felt uncertain whether this could be seen as validation practice worthwhile reporting. Similarly, interviewees mentioned that they had extensively immersed themselves in the textual data during the stage of

conceptualization. For instance, one interviewee noted that before applying CTAM, they often asked colleagues to identify the concept of interest on a small subsample of texts.

Next, *justifications of preprocessing decisions* were also a seldomly reported validation steps. According to one interviewee, this is because preprocessing often involves “a little bit of rule of thumb, seat-of-the-pants experience values and no real theory”. Therefore, interviewees often reported that they abstained from going into detail by referring to some general note of standard preprocessing practices. Thus, reporting and justifying the multiple project specific preprocessing decisions was often disregarded.

Another unreported validation step brought up by the interviewees was the *inspection of descriptive statistics* for text corpora prior to the analysis. According to the interviewees, inspecting descriptive statistics usually helps to detect basic patterns in the data. As one interviewee noted, “the first step should and usually is to look at differences in vocabulary between [categories], or conduct whatever exploratory analysis, [such as comparing document lengths between known groups]”.

Moreover, interviewees emphasized that *qualitative error analysis* is a validation step that is often underreported. One interviewee underscored the significance of qualitative error analysis, which entails the meticulous examination of wrongly categorized texts, as a crucial step to better understand the biases and limitations of CTAM. However, the interviewee expressed frustration regarding the reporting of qualitative error analysis results, citing the absence of clear performance criteria as a major hindrance. As a result, the interviewee reports that he also frequently omits reporting these results from reports due to reviewers’ length.

Lastly, interviewees pointed out their frequent omission of details concerning the *rejection of poorly performing models*. This practice commonly involved initially experimenting with simple models like dictionaries before progressing to more advanced models or adjusting their methodologies. According to the interviewees, while the rejection of underperforming models might not be considered a direct validation step (in the sense of providing validation evidence for a specific CTAM), they believed it was still worth reporting because it provided valuable context for the validation process.

In sum, three major themes emerged from the expert interviews: (1) the interviewed researchers were aware of the importance of CTAM validation; (2) they noted a lack of universally (or at least widely) accepted concepts and guidelines for researchers’; and

(3) they appeared to engage in a variety of validation that they did not necessarily report in the resulting manuscripts. This suggests that the relatively few validation steps reported in most papers we covered in our systematic review may in fact constitute only a subset of the complete range of validation activities.

2.6 Discussion

2.6.1 Critical Reflection

It would be an understatement to say that CTAM holds great promise for communication research. CTAM offers many ways in which text analysis can be improved and scaled to handle large quantities of texts. Notwithstanding the benefits offered by CTAM, it is of crucial importance to devote sufficient attention to the validation of text-based measures. Without sufficient evidence for CTAM validity, it is unclear whether the inferences at which the researchers arrive in substantive research can be trusted, and it becomes challenging to build a cumulative body of evidence. Unfortunately, the question of how to best validate CTAM remains largely unresolved, as several critics have repeatedly pointed out (Baden et al., 2021; Ribeiro et al., 2016). This article aimed to lay the groundwork for a normative engagement with CTAM validation by mapping the current field of current validation practices, which will subsequently serve as a fundament for practical recommendations.

Our results show that, in terms of *when* validation was conducted, researchers carried out validation for most of the studies assessed. However, for 14 per cent of the studies, authors did not report any validation, and these cases were exclusively limited to dictionaries. To evaluate *how* validation was conducted, we proposed an initial classification of validation steps that (1) enabled us to review current validation practices in a principled fashion and (2) can serve as the basis for a systematic engagement with validity in future work. Our classification was based on a synthesis of the general literature on validity in the social and behavioral sciences but accounted for the specific use cases of CTAM. We primarily distinguished between internal and external validation, a distinction that is common in the literature on validity (Grimmer et al., 2022; Quinn et al., 2010). Our results show that authors leaned more towards external than on internal validation. For internal validation, authors applied a plethora of validation steps, both for the properties of the model and the interpretation of the model output. For external validation, annotations by human coders were the most

popular subject of comparison, followed by other CTAM labels, surrogate labels, and the prediction of external criteria unrelated to the textual data.

Accompanying findings of our qualitative interviews suggested that researchers furthermore carried out additional validation steps that they often did not report in the published manuscript. According to the interviewees, most often, these validation steps related to conceptual considerations, such as construct definitions and design decisions, the iterative process of inspecting the data using descriptive statistics, and the qualitative analysis of errors.

However, the predominant theme that emerged from this systematic engagement with validity was the lack of consistency in how researchers justified the selection and reporting of their validation steps. Thus, our empirical findings show that validation was rarely grounded in a comprehensive conceptual understanding of validity. This inconsistency extended to the terminology used, which varied significantly across studies. Therefore, we subsequently take a more normative perspective and derive practical recommendations based on our empirical findings.

2.6.2 Practical Recommendations to Improve CTAM Validation

To tackle the issue of conceptual ambiguity regarding validity, we present practical recommendations for communication researchers based on our empirical findings and the previous literature on validity. In a nutshell, our recommendations are that communication researchers should (1) provide an explicit definition and operationalization of the construct for their textual data, (2) always validate the measure they use, (3) combine both internal and external validation, (4) always compare their measures with human annotations, and (5) maximize transparency and replicability. These recommendations are meant to provide communication researchers with the guidance for validating CTAM that has so far been lacking in a comprehensive manner.

Below we discuss each of these recommendations in turn. Afterwards, we will additionally offer a preview of new and upcoming validation frameworks that might be advantageous for researchers seeking guidance beyond the scope of our recommendations.

2.6.2.1 Justify your Construct Definition and Outline its Operationalization in Text

First and foremost, we strongly urge communication researchers to always justify their construct definition and outline how the definition guides the operationalization of the construct in the textual data. This is because the absence of a well-defined construct makes it unfeasible to present validation evidence for CTAM, as the basis for evaluating the measurement process remains ambiguous (Clark & Watson, 2019). Outlining the construct definition and ways of operationalization is already well-established practice for alternative research methods. The validation of survey scales, for example, typically involves some form of content validation that is concerned with conceptual reasoning on the domain and the operationalization of the construct (Adcock & Collier, 2001; Drost, n.d.; Flake et al., 2017; Rusticus, 2014). We argue that, for text analysis, conceptual considerations on how a construct manifests itself in text should be equally important. This is of particular significance in the field of communication research, as communication researchers delve into various social aspects where communication plays a pivotal role (Craig, 1999). Given the diverse contexts and data sources that are studied by communication researchers, it is imperative to provide precise definitions of the constructs under study to prevent confusion and ensure methodological soundness. Therefore, we explicitly recommend researchers to not only provide a definition for their construct grounded in literature (see Podsakoff et al., 2016), but also to discuss the implications of the definition for the measurement design, such as data and method selection, or preprocessing decisions (Denny & Spirling, 2018).

2.6.2.2 Always Validate CTAM that Measure Social Science Constructs

In light of the increasing use of large language models by communication researchers (Baden et al., 2021; Wankmüller, 2022), we want to reiterate in the strongest terms that validation is still an essential requirement for any measurement that involves CTAM. Admittedly, results of our review indicate that most researchers in our sample did validate their measurement, with the exception that dictionaries were sometimes applied without validation. Therefore, we noticed a broad consensus in the studies assessed that aligns with previous recommendations, for instance from Chan et al. (2021) and van Atteveldt et al. (2021), who emphasized the importance of validating due to the varying performance of CTAM across research contexts. However, there is reason to believe that the question on whether rigorous validation is necessary could be questioned by the promises of powerful pre-trained LLMs (F. Huang et al., 2023;

Kuzman et al., 2023; Reiss, 2023). Thus, the promises of LLMs carry the danger that communication researchers trust the output of these models blindly, without accounting for the of various types of biases and limitations inherent to LLMs (as discussed by Bender et al., 2021).

However, even the best model can produce unreliable measures when the method is not suitable for the specific research context. Therefore, to minimize the uncertainty inherent in any empirical measurement, it is essential to provide validation evidence to establish the validity of any scientific measurement instrument (Flake et al., 2017). While large language models can potentially aid researchers in the CTAM validation process by generating synthetic data or suggesting alternative labels, it is crucial to emphasize that human judgment and careful validation remain necessary to ensure valid measurements.²⁰

2.6.2.3 Combine Internal and External Validation

Considering the lack of a generally agreed terminology, we advise researchers not to become overly preoccupied with terminology for the time being. Instead, we suggest that researchers focus on the primary differentiation in our coding scheme, namely internal and external validation. This fundamental distinction between internal and external validation aligns with core principles in validation theory, despite potential variations in its terminology across literature (Flake et al., 2017; Loevinger, 1957).²¹

Based on this general distinction, researchers should then select and apply different validation steps for internal and external validation. Combining internal and external validation is important, as they provide two complementary perspectives on the quality of the validation process. For instance, relying solely on external validation carries the danger of model overfitting—an issue discussed in detail in Grimmer et al. (2022). On the other hand, relying solely on internal validation might provide little evidence that the model output converges with information outside of the textual model.

²⁰ Speaking more generally, our recommendation to always validate CTAM is also supported by the fact that methodological research in the social sciences generally abstains from the idea that a measurement instrument (i.e., a scale or a questionnaire) can be valid in itself and for all times, but can only be validated for specific contexts (Hedrih, 2019). Consequently, within the methodological literature, validation is typically conceived as an ongoing process that requires ongoing validation.

²¹ The Standards for Educational and Psychological Testing (2014), for instance, primarily distinguish between “Evidence based on Internal Structure” and “Evidence Based on Relations to Other Variables.”

Connected to that, it is important to apply different kinds of validation steps because each validation step comes with limitations. For instance, Jankowski and Huber (2022) have demonstrated that external validation using surrogate labels can cause problems when the labels are inadequate substitutes of the underlying construct. Similarly, relying solely on a single type of internal validation, such as post-hoc plausibility checks (Lipton, 2017) where one attempts to “make sense” of the model output or retrospectively assigns meaningful labels to topics in a topic model, can be problematic. This is due to the human tendency to perceive meaningful patterns in random or unrelated data (Cohen, 1960; Shermer, 2008), which can potentially lead to misleading conclusions.

Thus, we argue that validation is not about stringing together validation steps but rather aiming for a comprehensive and critical assessment of validity, including both internal and external validation, to limit the potential errors and pitfalls for different types of validation.

2.6.2.4 Always Conduct External Validation Using Human Annotations

Irrespective of the research context, we argue that human judgement should play a crucial role in all aspects of validation, including both internal and external validation. Nonetheless, we make a clear recommendation for communication researchers to incorporate, at the very least, external validation using human-annotated labels in any validation that includes CTAM. This is in line with guidance in the literature, which generally highlights the crucial role of human-annotations as a clear standard for evaluation in text analysis (Grimmer et al., 2022; Lacy et al., 2015; Lewis et al., 2013). However, as Song, Tolochko, et al. (2020) point out, human annotations are not always of high quality. Thus, researchers often put excessive trust in naïvely coded annotations, while they often do not consistently and clearly report the exact methodological details of the coding process. Therefore, we encourage researchers to adhere to the rigorous methodological standards found in the content analysis literature (Krippendorff, 2018). These principles encompass various steps, such as (1) providing a well-defined description of the annotation task, (2) creating a codebook through multiple feedback iterations, (3) adequately training coders on the codebook while ensuring sufficient performance, and (4) employing a minimum of two coders to calculate intercoder reliability.

2.6.2.5 Maximize Transparency and Reproducibility.

Lastly, we argue that scholars should maximize transparency and reproducibility of their validation. Transparency calls for the presentation of robustness checks to ensure that the empirical findings are robust to researchers' degree of freedom within the research process. Thus, by rerunning empirical analyses across different settings, consistent findings can provide important evidence that CTAM are robust against various issues. In our review, we observed several robustness checks, such as for how to aggregate scores from lower to higher level (e.g., from sentence to paragraph or document to corpus level) (Boukes et al., 2019), transforming numeric to categorical measures (Baden et al., 2020; Mueller & Saeltzer, 2020), selecting different number of topics in a topic model (Van Der Velden et al., 2018; Yarchi et al., 2021), or choosing a different text-based method (Mueller & Saeltzer, 2020).²²

To foster reproducibility, we advocate for adhering to open science standards, such as publishing all materials including data, code, and non-restrictive computational environments (e.g. a dockerfile), preregistering studies and submitting registered reports, as well as conducting replication studies to verify computational reproducibility (Dienlin et al., 2021, Schoch et al., 2023). In our review, we observed that 42 per cent of studies did not report access to reproducibility materials.

Furthermore, we recommend that researchers should generally avoid utilizing CTAM that are exclusively accessible through closed source blackbox APIs (such as Perspective API, ChatGPT, among others) due to the lack control over these models. Even worse, the models can undergo significant changes at any point without anyone noticing because the algorithms and models behind them are constantly changing (Rauchfleisch & Kaiser, 2020). This, in our view, poses an extremely problematic scenario for computational reproducibility, which is the reason that we generally abstain from using such APIs for research using text as data to measure social science constructs.

2.6.3 An Outlook on more Unified Validation Frameworks

We are confident that the five recommendations outlined thus far will assist researchers in improving their validation practices when applying CTAM. What lends

²² It is important to note that our list is not exhaustive as we did not explicitly document robustness checks in the coding process.

weight to our recommendations is that we based them directly on our empirical engagement with validation in our review. However, it is important to recognize that our recommendations, even though they provide strong conceptual guidelines on how to approach validation, cannot equip researchers with all-encompassing practical guidance for all scenarios involving CTAM validation. Therefore, we conclude this chapter by providing a glimpse into recent and ongoing work concerning more unified practical validation frameworks, designed to aid scholars in CTAM validation for different research contexts.

Building upon the findings of this review, we propose a unified validation framework, called *ValiTex*, specifically designed for the validation of CTAM measures of social science constructs (Birkenmaier et al., 2023b). The contribution of our framework lies in its practical character, as it provides a consistent terminology and comprehensive guidance on various validation steps available to researchers. At its core, *ValiTex* calls on researchers to demonstrate three overarching types of validity evidence: *substantive* evidence (outlining the theoretical underpinning of the measure), *structural* evidence (examining the properties of the text model and its output), and *external* evidence (testing for how the measure relates to independent information).²³ For each overarching type of validation evidence, *ValiTex* defines a detailed taxonomy of empirical validation steps that enable researchers to demonstrate validity. To ensure practical guidance, all validation steps, together with detailed description, are accessible via a checklist that is adaptable for distinct CTAM and use cases.²⁴

Similarly, researchers might also make use of method-specific frameworks tailored towards specific workflows or methods that require validation. Often, these method-specific contributions propose valuable workflows, targeted at the unique challenges of validating specific types of CTAM. Recent examples of such method-specific frameworks and guidelines target in particular unsupervised methods (Bernhard et al., 2023; Chan & Sältzer, 2020; D. Maier et al., 2022; Terragni et al., 2021; Ying et al., 2022) or supervised methods (see Chapter 20 on validation in Grimmer et al., 2022; Park & Montgomery, 2023). While they tend to be geared towards specific method-specific workflows, these approaches still furnish researchers with practical and valuable guidance.

²³ In the *ValiTex* framework, *substantive* and *structural* evidence constitute more fine-grained subcategories of internal validation.

²⁴ The framework and checklist are available at <https://www.valitex.info/>.

Moreover, there is ongoing research in a promising domain that is advancing the validation of multi-lingual text analysis (for an introduction, see Licht & Lind, 2023). In particular, Baden et al. (2022) proposed a framework designed for the validation of computational multilingual text analysis. In their framework, they explicitly highlight the various challenges of working with multilingual data, such as how comparable meanings are expressed differently across languages and provide a workflow on how to best validate multilingual text analysis. They propose to follow a process logic, including data validation, input validation, process validation, and output validation. Similarly, Ho and Chan (2023) introduce a model-agnostic workflow for validating the *transferability* of multilingual text analysis, that is, the extent to which CTAM performance can be maintained when switching from one language to another. For validation, they propose a workflow combining both quantitative and qualitative tests to ensure transferability. Combined, these two frameworks serve as an excellent illustration of how validation practices can be enhanced within a specific field that was previously lacking coherence.

Despite notable variations in the terminology and the focus of the validation frameworks presented so far, they constitute major steps toward addressing the challenges associated with CTAM validation. Therefore, we believe that these frameworks can serve as sources of valuable practical guidance for CTAM validation. It is our hope that the empirical insight and practical recommendations gleaned from our review can help to harmonize current methodologies and contribute to forming a generally agreed understanding for CTAM validation.

2.7 Limitations and Directions for Future Work

We would like to note three limitations of our current work. Firstly, in the absence of widely shared terminology and frameworks, we developed our own classification of validation steps. Although this classification was firmly rooted in the validity literature and discussions with domain experts, there is an inherent subjectivity in our categorizations, and other researchers might have chosen different categories. Still, from a pragmatic standpoint, we believe our heuristic classification has value in itself and might be further adapted and discussed by other researchers (see Birkenmaier et al. (2023b) for an extended classification of validation steps).

Secondly, in our systematic review, we only included publications within a limited number of high-ranking journals. Because the research field of text as data is, however, quickly developing, it cannot be ruled out that we missed relevant

publications. Connected to the comparatively low number of studies assessed, we were furthermore unable to make statements about how perspectives on validation changed over time. Future research may have the potential to map the research field more comprehensively. As results from our qualitative interviews point to the crucial role of peer-review processes and journal requirement, further research could systematically examine the role of publication outlets in shaping how scholars validate CTAM.

Lastly, our systematic review focused on political communication. We chose this focus because political communication is a field that heavily utilizes CTAM and has been pioneering many of the methods used in communication research. Given this focus, the results of our systematic review cannot be directly generalized to other subfields. However, in our view, there is little reason to believe that validation practices in political communication are markedly different from those encountered in adjacent (and often overlapping) fields of communication research. This is because the goal of validation is usually to demonstrate that a measurement instrument measures what it purports to measure for a specific research context. Whereas the precise research contexts might change according to the factors such the research question and constructs assessed (e.g., studying the effects of offensive language or different communication frames), or the choice and characteristics of data (e.g., analyzing Tweets with a strict 140-character limit or newspaper articles with varying length), the underlying validation steps remain largely consistent across research fields. Even more important, the absence of guiding frameworks and standardized vocabulary is a common challenge across all social science research domains. Therefore, we argue that the empirical findings of our review and especially our normative recommendations for more rigorous validation practice hold relevance beyond political communication for all subfields of communication research, and essentially for all social science disciplines that employ CTAM to measure social science constructs. At the same time, we invite researchers to expand our inquiry into validation practices – and how to improve them – to other domains of communication research.

2.8 Conclusion

In conclusion, whereas CTAM validation remains a challenging task, we hope that our review can contribute to a more systematic engagement with CTAM validation for communication research. Based on a comprehensive review of current field of CTAM validation, we derived practical recommendations that could be the starting point to structure discussion around validation and avoid common pitfalls. Looking

ahead, we encourage researchers to build on our work towards a solid methodological foundation for CTAM validation that is guided by strong conceptual claims on what constitutes sufficient validation evidence.

2.9 Appendix

2.9.1 Appendix: Selection of Journals

Communication Research Journals (“Communication”)

<https://www.scimagojr.com/journalrank.php?category=3315>

1. Communication Methods and Measures
2. Political Communication
3. Communication Research
4. Digital Journalism
5. Journal of Communication

Political Science Journals (“Sociology and Political Science”)

<https://www.scimagojr.com/journalrank.php?category=3312>

1. Administrative Science Quarterly (**no eligible publication found**)
2. American Sociological Review
3. American Political Science Review
4. Annual Review of Political Science (**no eligible publication found**)
5. Journal of Service Research (**no eligible publication found**)

Other journals considered

1. American Journal of Political Science
 - Reasoning: Well established journal with proven record of publishing CTAM studies for a wide range of topics associated with political communication, such as party politics (e.g., Jung, 2020).
2. Political Analysis
 - Reasoning: Well established methods-journal that discussed questions of validation from an early stage on (e.g., Goet, 2019)
3. Legislative Studies Quarterly
 - Reasoning: Well established journal with proven record of publishing CTAM studies in the domain of parliamentary communication (e.g., Proksch et al., 2019)
4. Information, Communication & Society
 - Reasoning: Journal with a focus on the impact of communication technologies for society and political communication. Therefore, we promised to encounter promising research using CTAM.
5. JOURNAL OF INFORMATION TECHNOLOGY & POLITICS
 - Reasoning: Journal with a focus on the impact of communication technologies for society and political communication. Therefore, we promised to encounter promising research using CTAM.
6. Applied Linguistics
 - Reasoning: Direct Reference to textual content

2.9.2 Appendix: 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
- Schürmann, L., & Stier, S. (2023). Who Represents the Constituency? Online Political Communication by Members of Parliament in the German Mixed-Member Electoral System. *Legislative Studies Quarterly*, 48(1), 219-234.
- Rossini, P., Sturm-Wikerson, H., & Johnson, T. J. (2021). A wall of incivility? Public discourse and immigration in the 2016 US Primaries. *Journal of Information Technology & Politics*, 18(3), 243-257.
- Eisele, O., Litvyak, O., Brändle, V. K., Balluff, P., Fischeneder, A., Sotirakou, C., ... & Boomgaarden, H. G. (2022). An emotional rally: exploring commenters' responses to online news coverage of the COVID-19 crisis in Austria. *Digital Journalism*, 10(6), 952-975.

2.9.3 Appendix: Search Terms

- Naïve search
 - (Politi* OR Party OR Govern*) AND text*
- Adjusted search
 - 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*)

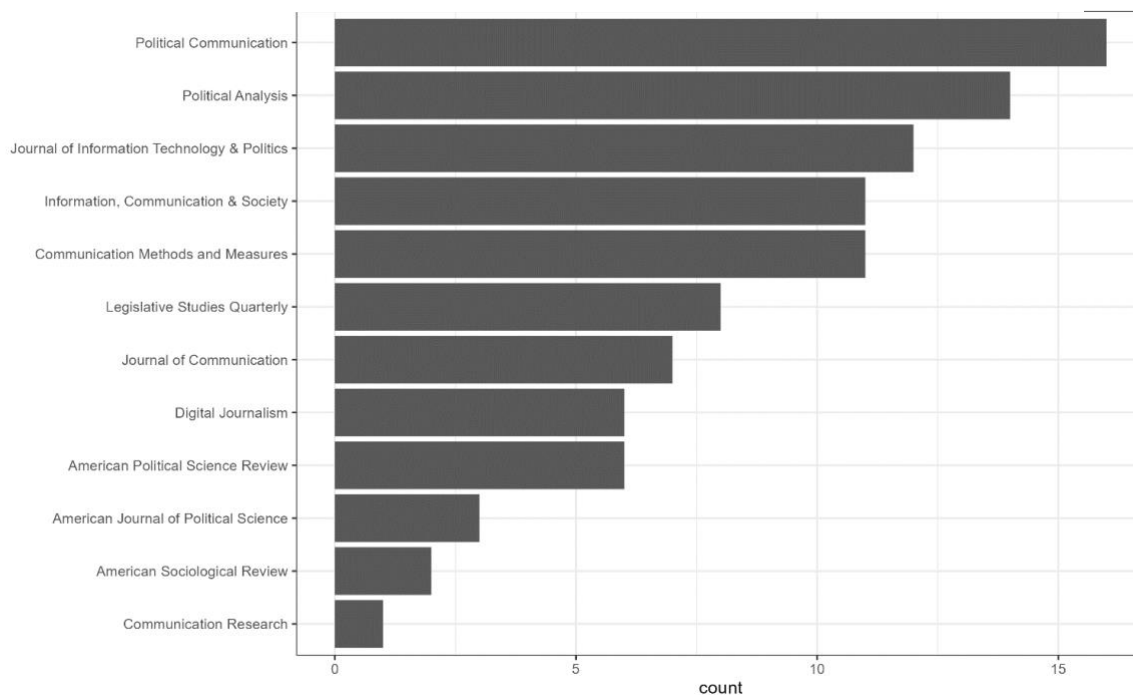
2.9.4 Appendix: List of Studies Included in the Systematic Review

Please follow this link <https://figshare.com/s/7418783cfa9f75c984f8?file=39635053> to get

- A complete list of studies included in the systematic review
- A complete list of validation steps coded for each study included in the systematic review

2.9.5 Appendix: Overview of Studies per Journal

Figure 10: Overview of Studies per Journal



2.9.6 Appendix: Interview Guideline

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 CTAM	What are the barriers that hamper the validation of computer-assisted text methods? (<i>Costs, missing guidelines etc.</i>)
	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?

3. Developing Conceptual and Practical Guidance for Validation

The second part of the dissertation develops a unified validation framework for CTAM measures of social constructs. It addresses the guiding research question: How can researchers validate their measurements? Building on the insights from the systematic review, this part of the dissertation offers both conceptual and practical guidance. It introduces a clear and transferable conceptualization of validity, grounded in the philosophy of measurement within the social sciences, and proposes shared terminology to foster consistency in validation practices. In addition, it provides practical tools in the form of templates and checklists to support researchers in planning, conducting, and documenting validation procedures.

This chapter is based on the following publication: Birkenmaier, L., Wagner, C., & Lechner, C. (2025). ValiText--a unified validation framework for CTAM measures of social constructs. arXiv preprint arXiv:2307.02863. <https://doi.org/10.48550/arXiv.2307.02863>

3.1 Introduction

CTAM measures of social constructs are difficult to validate (Baden et al., 2020; Kantner & Overbeck, 2020). However, validation is necessary because measures derived from textual data are often used as variables in downstream analysis. Hence, if a text-based measure inaccurately represents the concept or phenomenon it aims to capture, this can introduce systematic bias and errors in the research process. Therefore, in order to determine whether a measure is a valid operationalization of the underlying construct, it is crucial to thoroughly validate the relationship between text-based measures and theoretical constructs (Hughes, 2018). To do so, the most established method is comparing the output scores of computational methods with human annotations that serve as a gold-standard benchmark (Grimmer et al., 2022).

In this paper, we draw attention to critical opinions suggesting that the mere comparison of output scores with human annotations is insufficient for thorough validation. This view is supported by a long-standing body of research in the social sciences, which emphasizes the need to present multiple forms of validation evidence to validate measures of social constructs effectively (Flake et al., 2017). In computational text analysis, social scientists routinely employ a broad spectrum of

validation strategies going beyond simple human annotation comparison (Birkenmaier et al., 2023). However, strategies for validation are often diffuse, or lack justification altogether. For instance, methodological frameworks and best-practices are often transferred from specific social science subfields without fully accounting for the specific nature of text-based measures (e.g., Adcock & Collier, 2001). Additionally, inconsistencies in terminology further complicate the matter, as variability in language and definitions can cause misunderstandings, or lead to validation steps being deliberately left out in the reporting of results (Birkenmaier et al., 2023a).

To address these concerns, this paper introduces *ValiText* - a unified validation framework for CTAM measures of social constructs. As such, the framework aims to fulfil three objectives: a) to stress the importance for researchers to provide different kinds of validation evidence for computational text analysis, b) to define the key types of validation evidence that are required for sufficient validation, c) to guide researchers which concrete validation steps to apply for their text-based research, and d) to provide a documentation template in the form of a checklist that can be used to document validation efforts effectively.

3.2 Bias and Error in Computational Text Analysis

Validation is a necessary requirement for computational text analysis (Grimmer et al., 2022; Krippendorff, 2018). Essentially, the purpose of validation is to uncover and correct mistakes in the measurement process, ensuring that the output scores are an accurate reflection of the true nature of the construct being studied. Mistakes in measurement can arise from different sources, with varying consequences for the validity of the measurement.

On the one hand, random mistakes (*errors*) can be seen as uncontrolled and random errors that occur due to the limitations of the research design. These errors are not inherently problematic for the measurement because it's assumed they will neutralize each other. For text analysis, errors can be attributed to the inherent complex and multi-dimensional nature of text, as well as to the uncertainty about how social constructs are depicted in texts. For instance, language can often be interpreted in different ways depending on the readers perspective, knowledge, or cultural background. Examples of these characteristics are the presence of sarcasm or irony (Ravi & Ravi, 2015), ambiguity and polysemy (Boxman-Shabtai, 2020; Roberts, 1989), or context-specific references and interpretations (Krippendorff, 2018; Mayring, 2004). Furthermore, the construct that is to be measured (i.e., an emotion or a political

topic) can often be expressed differently in text, thus allowing for multiple ways to operationalize the construct of interest (Drisko & Maschi, 2016; Krippendorff, 2018).

On the other hand, systematic errors (*biases*) are more problematic, as biases can severely compromise the validity of measurements. This is because social constructs are typically used for further analysis as dependent or independent variables to draw insights about real-world relationships. Biases arising from incorrect measurements of constructs can negatively affect the clarity of these relationships, thus severely affecting subsequent analytical processes. For text analysis, such biases may result from methodological flaws or inherent prejudices in the methods or strategies applied by researchers. For example, large language models (LLMs) may introduce bias in the measurement process via the data used to train and test the models, either via data leakage (Gibney, 2022; Kapoor & Narayanan, 2022), non-representative datasets in respect to the measured construct (Cai et al., 2022), or model-inherent human stereotypes (van Giffen et al., 2022).

Likewise, the text analysis pipeline is usually characterized by high researchers' degrees of freedom (i.e., the various choices of collecting and analyzing data, see Bakker et al. (2020)). Because computational methods require a variety of analytical choices, starting from construct operationalization, model selection, data pre-processing and model hyperparameters, researchers' decisions often affect the outcomes of analyses significantly (Arnold et al., 2023).

Ultimately, validation enables the researcher to identify bias – thereby reducing the uncertainty around the measurement process. In other words, without sufficient validation, researchers can never completely rely on the accuracy of their empirical findings, regardless of whether they turn out to be correct or inaccurate.

3.3 A Critical Perspective on Human Annotations as a Gold-Standard

The most common way to ensure valid measurements is comparing text-based measures with human annotations, often referred to as “gold-standard” data (Grimmer et al., 2022). These human annotated labels are assumed to be entirely accurate and objective to provide a sufficient benchmark for the true underlying concepts. The foundation of this approach rests on the belief that human comprehension of texts surpasses that of text models and that, when adequately trained, humans can achieve the most accurate and valid text judgments (Song, Tolochko, et al., 2020b). Although it's widely accepted that human annotations are essential to demonstrate validity, there

is significant critique challenging the notion that human annotations alone are sufficient for validation.

First and foremost, we must constantly be reminded that objective “gold-standard” data does not exist. For instance, the content analysis literature, which focuses on manually describing communication data using trained coders, usually neglects the notion of a true underlying “gold standard” labels. Despite meticulous efforts to ensure objectivity- such as providing comprehensive construct definitions, utilizing double-blind coding, conducting inter-coder reliability assessments, and thoroughly training coders (Krippendorff, 2018) - the literature emphasizes the inherent variability in human perception. The differing interpretations suggest that the context in which a text's content is perceived plays a crucial role in determining its meaning (Drisko & Maschi, 2016). Thus, we need to detach ourselves from the idea that gold-standard targets in the form of objective truths exist after all for social constructs in texts.

Building upon this observation, human annotations frequently fall short of acceptable quality standards. In particular when annotations are sourced from crowd workers or untrained annotators, the quality of annotations can diminish due to factors like varied levels of expertise, differing interpretation of guidelines, and potential lack of motivation (Chmielewski & Kucker, 2020). Relying on Monte-Carlo Simulation and simulated data, Song et al. (2020) vividly illustrate that low-quality human annotations can impact the results derived from empirical measures. In their study, Song et al. systematically varied factors such as the sampling variability, the number of human coders, the size of the validation data, or the required level of intercoder reliability, and showed that these factors significantly affected the classification accuracy on the “true” labels of their simulated textual data.

Likewise, when researchers rely on surface-level correlations, it often prevents them from gaining context-specific insights into the measurement. This can lead to flawed conclusions, potentially caused by systematic bias or low quality data (see Daikeler et al., 2023). A striking illustration of this is provided by Hirst et al. (2014), who showed that a machine classifier designed to identify political ideologies erroneously focused on incumbency. This error occurred because incumbency was more easily detected by the model and coincidentally correlated with ideology in their dataset. To identify these more fine-grained limitations, more nuanced validation strategies beyond output comparison are required.

3.4 Deriving the Validation Framework

We suggest a validation framework for text analysis that complements the mere comparison of output scores with human annotations. We build upon earlier work of a systematic review conducted in the field of political communication (as outlined in Birkenmaier et al. (2023)). In the systematic review, we documented more than 20 different types of validation evidence and grouped them into overarching categories. Furthermore, we conducted expert interviews to further document additional activities that researchers undertake during validation, but do not usually report.

Ultimately, the development of the ValiText framework followed two main stages: first, we defined the fundamental types of validation evidence needed for text-based measures of social constructs. To do so, we extended the broad categories of validation evidence identified in the systematic review and embed them into well-established practices within the social science literature. Second, we gather an extensive list of empirical validation steps (i.e., specific tests that can be executed to produce validation evidence) for each type of validation evidence. Each step is meticulously detailed in a checklist, providing practical guidance to researchers for different use cases.

3.4.1 Types of Validation Evidence

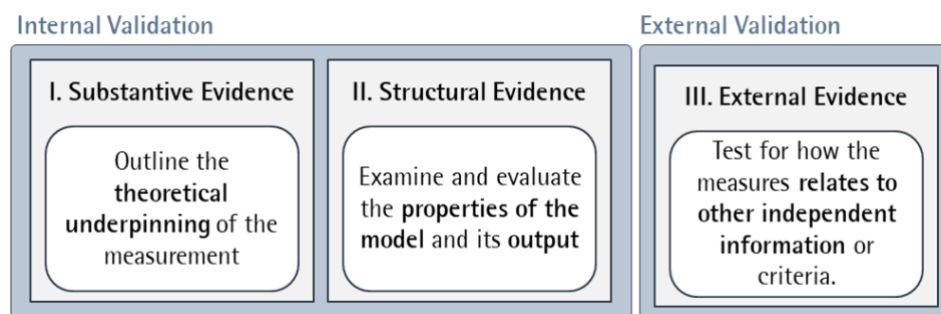
There are multiple approaches on how to separate different dimensions of validation evidence for text analysis conceptually (Fang et al., 2022; Grimmer et al., 2022; Quinn et al., 2010).²⁵ We initially build on the conceptualization introduced in Birkenmaier et al. (2023) who differentiate between *internal* and *external* validation evidence. Whereas internal validation relates to the systematic assessment of the measurement model and its output, external validation relates to the comparison of output scores with independent information. Empirically, these two fundamental types of validation evidence serve as a useful distinction of the underlying motivations in the validation of text-based measures, i.e., understanding the characteristics and limitations of the measurement process (internal validation) and demonstrating that the output scores are empirically linked to alternative measures or criteria (external validation).

In order to integrate our framework more deeply with established measurement theory, we build upon and expand Loevinger's (1957) conceptual validation framework

²⁵ For a detailed discussion, see Birkenmaier et al. (2023)

for use in computational text analysis. At its core, the conceptualization of Loevinger

Figure 11: Three Fundamental Types of Validation Evidence



(1975) incorporates three fundamental types of validation evidence, i.e., *substantive*, *structural*, and *external* evidence. Traditionally, these types of evidence are associated with the developmental phases of creating measurement instruments for social constructs: constitution of the item pool based on theory and construct definition, analysis and selection of the internal item structure, and correlations of the construct with external constructs (Clark & Watson, 2019). As such, these three types of evidence are mutually exclusive and exhaustive to all types of validation evidence, and mandatory to demonstrate measurement validity. Figure 11 presents these fundamental types of validation evidence adapted to the context of text analysis. As described in more detail below, *substantive* and *structural* evidence pertain to internal validation, and *external* evidence pertains to external validation. For text analysis, to provide *substantive evidence*, researchers need to outline the theoretical underpinning of the measurement. To provide *structural evidence*, researchers should examine and evaluate the properties of the text model and its output. To provide *external evidence*, researchers should test for how the empirical measures relate to other independent information or criteria.

One advantage of adapting this tripartite structure of substantive, structural, and external evidence is to avoid confusion with established, but often confusingly used validity terms such as “discriminant validity” or “face validity” in the context of text analysis. Nonetheless, we map established validity terms to each stage of the framework and provide a complete overview in Appendix 1.

Lastly, we also add a separate section about robustness checks (i.e., reporting alternative specifications of the measurement). As Grimmer et al. (2023) note, small changes across the measurement design can dramatically change measurement

outcomes, and a transparent reporting of alternative measurement designs effectively provides evidence for measurement validity.

3.4.2 Selecting Validation Steps

For the selection of empirical validation steps to the checklist, we initially start by the documentation of validation steps identified in Birkenmaier et al. (2023). For instance, examples for internal validation are *feature importance analysis*, and *comparison with human annotations* for external validation. In addition, we also consider validation steps that were brought up in the expert interviews with domain experts as validation activities that are rarely reported, such as *manual pre-coding of human annotators* or *qualitative strategies for error analysis*.

One aspect to overcome is that some validation steps only apply to a specific type of text-based method, such as inspection of *topic coherence metrics* for topic modelling, or require different techniques for different methods, such as *feature importance analysis*. Therefore, we initially define the validation steps in a broad way to capture the underlying idea of each validation step. In the checklists, however, we provide a more detailed overview of different use cases, also including potentially context-specific, though meaningful validation steps. To identify validation steps beyond the ones reported in social science publications, one of the authors with a background in computer science further complemented the existing list of validation steps with an evaluated opinion on state-of-the-art steps validation strategies in the broader research field of natural language processing (NLP).

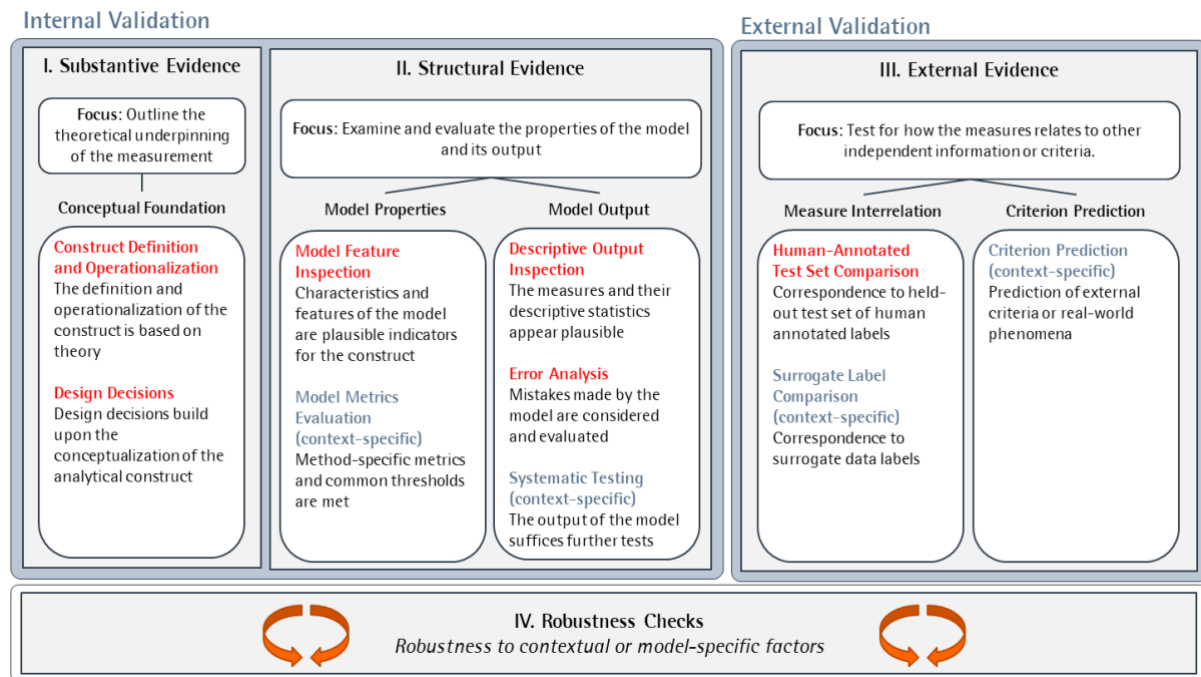
After selecting the validation steps, we carefully classify and document each validation step in the checklist. For each validation step, we include its name, a brief description, implementation methods, and a classification indicating whether the validation step is generally applicable or context-dependent. Additionally, we provide references to practical applications and further literature.

Nonetheless, it is important to note that our current selection and description of validation steps does not claim to be complete, nor that new strategies and tools should not be included. Therefore, we implement a participatory process that will allow researchers to add and discuss specific validation steps via an open repository on GitHub that allows researchers to subsequently suggest new validation steps and practical tools that can be documented using the frameworks' guiding structure.

3.5 Overview Framework

Figure 12 presents the final framework distinguishing between different subtypes of validation evidence. In the following chapters, we will primarily discuss the core subtypes of validation evidence and the respective validation steps in the checklist in greater detail. In Appendix 2 we further discuss the context-specific subtypes of validation evidence that can further supplement validation efforts.

Figure 12: ValiText Framework



3.5.1 Substantive Evidence

Before conducting any measurements, researchers should outline the theoretical underpinning of the measurement to demonstrate *substantive evidence*. This can be done by documenting and providing justification for all conceptual considerations prior to measurement. Validation steps for substantive evidence should therefore demonstrate that the measurement is based on a strong conceptual foundation and that all design decisions are sufficiently justified based on the conceptualization of the construct.

Table 4 displays the validation steps that provide substantive evidence.

Table 4: Validation Steps for Substantive Evidence (Context-Independent)

ID	Validation Steps	Considerations	Implementation	Reference
Conceptual Foundation				
Construct Definition & Operationalization				
I.1	Documentation of the conceptual background	Have I conducted a literature review or consulted with domain experts to gain a sufficient understanding of conceptual background of the construct?	Summarizing existing literature on the conceptual background of the construct	Krippendorff (2018)
I.2	Documentation of the operationalization and codebook development	Have I sufficiently explained how the construct should manifest itself in the textual data? Have I documented my operationalization in a codebook?	Providing definition and conceptualization of the construct	Krippendorff (2018)
I.3	Manual Precoding	Have I reached sufficient interrater agreement for a subsample of the textual data? Have I ensured that the construct can be detected in the textual data? Have I outlined my rules of coding uncertainty across coders?	Reaching sufficient interrater agreement (e.g., Krippendorff's alpha α)	Krippendorff (2018), Plank (2022)
Design Decisions				
I.4	Justification of data collection decisions	Have I selected a dataset that is representative and relevant to the research question and population of interest? Have I justified the data selection decisions (e.g., using keywords)? Have I assessed the quality and completeness of the dataset and checked for potential biases or inconsistencies?	Outlining the rationale behind data selection / collection decisions; Documenting potential limitations and data quality issues	Krippendorff (2018)
I.5	Justification of method choice	Have I selected the appropriate type of method based on the operationalization of the construct and data characteristics? Have I justified the concrete selection of a particular model?	Outlining the rationale behind method selection; Documenting potential limitations in comparison to alternative methods	Grimmer et al. (2022)
I.6	Justification of the level of analysis	Have I selected the appropriate level of analysis? Have I considered potential problems when aggregating measures from lower to higher levels (e.g., sentence to paragraph level)?	Outline the rationale behind the selected level of analysis (e.g., token, sentence, or paragraph level).	Jankowski & Huber (2022)
I.7	Justification of preprocessing decisions	Have I justified relevant changes to the text prior to the analysis, such as removing certain words or phrases?	Outlining the rationale behind preprocessing decisions	Grimmer et al. (2022)

Note: An interactive version of the checklist is available online at <https://www.valitex.info/>

Construct definition and operationalization. Usually, researchers should start with *documenting the conceptual background* (I.1) of the construct. This may involve referencing previous definitions or research on the construct's

dimensionality and its manifestation in language. Next, researchers should provide a *justification for the operationalization* (I.2), creating a connection between the construct definition and the textual data. As an example, if researchers aim to measure polarization in political discourse, they could make the argument that heightened polarization is likely to result in the use of more distinct vocabulary, as supported by previous studies (Lowe & Benoit, 2013; Peterson & Spirling, 2018). To test their operationalization, researchers then need to conduct *manual precoding* (I.3) where two or more coders manually label a subset of the data to determine if the construct is sufficiently defined and discernible by human coders.²⁶ If researchers see that the coding task is too subjective, they can adjust their operationalization, possibly considering “learning from disagreement” methods (see Uma et al., 2021).

Design Decisions. Based on the construct definition, researchers must demonstrate that their design decisions are rooted in the conceptualization of the analytical construct. Crucial design decisions involve outlining the criteria for data collection (I.4), method selection (I.5), level of analysis (I.6) and preprocessing decisions (I.7). Broadly construed, design decisions should primarily be derived from the construct definition and its operationalization. This is to navigate through what Gelman and Loken (2016) call a “garden of forking paths” – having to take and justify multiple decisions in a setting of high researchers degree of freedom.

For instance, for the *justification of data collection decisions* (I.4), researchers can critically discuss general considerations on the origin of the data, such as the level of knowledge about the author(s) and the data generation process. Usually, researchers are interested that their text corpus comes from the units of analysis they want to make statements about (i.e., social media users) and not about automated bots or other unknown units.

Another example of relevant design decisions relates to the *justification of the level of analysis* (I.6) at which the construct should be measured, such as the sentence, paragraph, document, or corpus level. Here, researchers need to rule out a misalignment between the presumed manifestation of the construct and the level of measurement (as defined by the operationalization), which will likely lead to inconsistent and theory-agnostic measures (Jankowski & Huber, 2022; McKenny et al., 2013).

²⁶ If human coders are not able to correctly identify a construct, any computational method is likely to fail as well, possibly by taking up spurious relations in the data not connected to the construct of interest.

3.5.2 Structural Evidence

For structural evidence, researchers can conduct validation steps to examine and evaluate the properties of the model and its output. Thus, the underlying motivation behind structural evidence is to understand the characteristics of the measurement to eliminate biases and identify limitations. Structural evidence thus enables researchers to gain a deeper understanding of the measurement process. This may require an iterative procedure, where the initial model or features of the textual data are adapted.

Table 5 displays the validation steps that provide structural evidence.

Table 5: Validation Steps for Structural Evidence (Context-Independent)

ID	Validation Steps	Considerations	Implementation	Applications
Model Properties				
Model Feature Inspection				
II.1	Inspection of predictive model features	Have I inspected the predictive features for my model? Have I assured they are conceptually aligned with the construct being measured?	Qualitative evaluation of top-ranked model features using feature-importance methods like e.g., LIME or ICE	Molnar (2020), Küpfer & Meyer (2023)
Model Output				
Descriptive Output Inspection				
II.2	Visual inspection of output	Have I visualized my output descriptively? Have I identified and visualized outliers and extreme values?	Plotting descriptive statistics; discussing the plausibility of the observed distribution	Goet (2019)
II.3	Comparison of aggregated measures across known groups	Have I aggregated the output scores across known groups (e.g., mean share of sexist sentences across social media user demographics)?	Plotting aggregated measures across groups; discussing the plausibility of the observed distribution	Goet (2019)
II.4	Qualitatively assess top documents with the highest overall scores for each output category	Have I assessed the most outstanding documents for each type of output, such as labels with the highest confidence, or highest and lowest scores on a numerical scale?	Qualitative evaluation to ensure that the top-ranked texts align with the construct	Goet (2019)
Error Analysis				
II.5	Error analysis using data grouping	Have I conducted error analysis to compare the performance of my model across known subgroups?	Comparing performance metrics (i.e., F1) across subgroups	Wu et al. (2019)
II.6	Error analysis of outstanding or deliberately chosen observations	Have I conducted error analysis to qualitatively evaluate the sources and types of errors associated with the measures?	Exploring the underlying causes of misclassifications by qualitatively screening misclassified examples	(Wu et al., 2019)

Note: An interactive version of the checklist is available online at <https://www.valitex.info>

Model feature inspection. The goal of *model feature inspection* (II.1) is to assess whether the characteristics and features of the model are plausible indicators

for the construct. For instance, a model aiming to measure populism should prioritize meaningful words like “elite” or “establishment,” rather than irrelevant ones unrelated to the construct of interest. Whenever possible, validation can therefore include the inspection of the learned weights of features and distributions (see Linardatos et al., 2020; Molnar, 2020).

Descriptive output inspection. Researchers should conduct validation steps to ensure that the measures and their descriptive statistics look plausible. For text analysis, this form of validation steps is often referred to as “face validity” as it solely includes argument-based validation steps that require rather qualitative reasoning and less formal analysis (Goet, 2019; Temporão et al., 2018).²⁷

For example, descriptive output inspection can include the *visual inspection of the output* (II.2), in particular outliers and extreme values to get a feeling on the distribution of output scores. Based on this initial examination, researchers can already discuss the plausibility of observed distribution, and identify inconsistencies. Furthermore, researchers can also inspect the *aggregated output across known groups* (II.3) visually. These groups can relate to various variables typically included in the data, such as specific time periods or demographic characteristics (e.g., gender or age groups of social media users). For example, if prior research suggests that men are more likely to exhibit sexist behavior than women, the average proportion of sexist tweets should be higher for men than for women in a sample of social media users. Likewise, it can also be useful to *qualitatively assess the top documents with the highest overall scores for each output category* (II.4), such as for distinct groups or topics, or highest and lowest scores on a numerical scale.

Error analysis. Additionally, researchers should apply error analysis to ensure that systematic biases and errors are considered and evaluated. We here use error analysis as an umbrella term for a set of exploratory analyses that attempt to analyze where mistakes in the text analysis pipeline emerge and how they might affect the outcomes (Wu et al., 2019).

There are broadly two approaches to error analysis. On the one hand, researchers can conduct *error analysis using data grouping* (II.5) to identify especially problematic categories (per-label performance) (Alsallakh et al., 2014). On the other hand, researchers can *conduct qualitative error analysis of outstanding or*

²⁷ In other domains, face validity often refers to the subjective appearance of whether a method appears to be valid (Flake et al., 2017). Due to this ambiguity, we step back from using the term face validity for ValiText, but rather stick to the label “visual inspection of model output”

deliberatively chosen observations (II.6) to manually identify problematic text characteristics (for an application, see Wadhwa et al., 2018).

In Appendix 2, we further outline additional context-specific types of validation evidence for structural evidence, in particular the evaluation of *model metrics* (i.e., reporting of method specific metrics and thresholds) and the implementation of *systemic tests* (i.e., various strategies and experimental approaches to demonstrate that the models can generate meaningful measures for the construct being measured).

3.5.3 External Evidence

For external evidence, researchers should conduct validation that test for how the measures corresponds to independent information or criteria. Thus, information outside the scope of the textual data in which the measure was constructed serves as an external benchmark (hence “external” evidence). However, it is important to note that external validation alone is not sufficient because even a model with strong bias (lacking substantive and structural evidence) may exhibit meaningful but misleading associations with external criteria. Hence, focusing excessively on output optimization might result in overlooking errors without a complete understanding of their nature.

Table 6 displays the validation step that provide external evidence.

Table 6: Validation Steps for External Evidence (Context-Independent)

ID	Validation Steps	Considerations	Implementation	Reference
Measure Interrelation				
Human-annotated test set comparison				
III.1	Comparison of measures with human-annotated test set (“gold-standard data”)	Have I reached sufficient predictive performance on a test set of held-out human annotations? Did I apply cross-validation to calculate average performance metrics?	Evaluating performance metrics (i.e., F1) for dataset of human annotations	(Samory et al., 2021)

Note: An interactive version of the checklist is available online at <https://www.valitex.info>

Primarily, researchers should *compare measures with a human-annotated test set* (III.1), often referred to as benchmark or “gold-standard” data (Lowe & Benoit, 2013). Generally, human annotations are considered the most reliable form of validation for text-based research, regardless of the measurement design. In the checklist, we offer additional references to methodological standards and best practices, including coding procedures (using a codebook or pairwise comparison) and reporting classification metrics, such as (average) accuracy, precision, recall and F1-score.

In Appendix 2, we further outline additional context-specific types of validation evidence for external evidence, in particular *surrogate label comparison* (i.e., correspondence to labels that can be linked with the textual data) and *criterion prediction* (i.e., correspondence to external criteria or real-world phenomena).

3.5.4 Robustness Checks

Next to the three types of validation evidence outlined above, the ValiText framework also recommends the continuous test of robustness checks to assess the impact of researchers' degree of freedom on the measurement outcome. On a general note, one could see robustness checks as additional means to test whether decisions regarding the measurement design might have a sustainable effect on the measure's outcome.

Table 7 provides an overview over relevant aspects which might be tested.

Table 7: Robustness Checks (Context-Independent)

ID	Validation Steps	Considerations	Implementation	Reference
Robustness Checks				
IV.1	Rerunning the analysis using alternative text models	Have I rerun the analysis with alternative text-based methods, such as a baseline model?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative measurements	(Samory et al., 2021)
IV.2	Rerunning the analysis using different hyperparameter settings	Have I rerun the analysis with alternative hyperparameter settings?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative hyperparameter settings	(Arnold et al., 2023)
IV.3	Rerunning the analysis using different cutoff thresholds	Have I rerun the analysis with alternative cutoff thresholds?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative cutoff thresholds	Grimmer et al. (2022)
IV.4	Rerunning the analysis using different text cleaning and preprocessing steps	Have I rerun the analysis using alternative data cleaning and preprocessing settings (e.g., removing certain phrases or features of the data)?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative preprocessing steps	Grimmer et al. (2022)

Note: An interactive version of the checklist is available online at <https://www.valitex.info>

In general, factors to consider for robustness checks relate to rerunning the analysis using *alternative text models* (IV.1), *hyperparameter settings* (IV.2), *cutoff thresholds* (IV.3) and *text cleaning and preprocessing steps* (IV.4) and displaying variance in measurement outcomes and performance metrics.

3.5.5 Checklist

For each of the validation steps presented, we complement our framework with a checklist that is adaptable for different use cases (see Table 8). In particular, we provide individual checklists for applying dictionaries (A), (semi-) supervised classification (B), prompt-base classification using generative LLMs (C) and topic modelling (D).

Table 8: Use Cases

#	Use Case	Training / finetuning data required	Known output categories	Description	Example
A	Dictionaries	No	Yes	Assign scores to text units using predefined word lists	A dictionary assigns polarity values ranging from -1 to 1 to each known text unit
B	(Semi-) Supervised Classification	Yes	Yes	Train a model to predict known output categories based on labelled training/fine-tuning data	A pretrained BERT model is fine-tuned on labelled social media posts and predicts “offensive” and “non-offensive” posts
C	Prompt-based classification using LLMs	Yes	Yes	Ask (“prompt”) LLMs to assign labels to texts	A Mistral model is prompted to determine whether a text is “sexist” or “non-sexist”
D	Topic Modelling	No	No	Assign topics without any labeled data	An LDA topic model generates 13 coherent topics

An important condition for all use cases is that researchers must have full control over their models and be theoretically capable of replicating findings by setting random seeds and having access to model parameters. As a result, we deliberately decided not to include a use case for closed-source blackbox APIs (such as Perspective API, ChatGPT²⁸ interface, among others) and we generally avoid using them to measure social constructs (Rauchfleisch & Kaiser, 2020; Schoch et al., 2023).

The checklists are provided as fillable templates in Appendix 3 and can also be downloaded from a publicly available shiny app.²⁹ In Appendix 4, we demonstrate the applicability of the checklists by revisiting and documenting the validation steps from a study by Samory et al. (2019), which measured sexism in social media comments for the use case “B: (Semi-) Supervised Classification”. Ultimately, the checklists can be used to it can function as a methodological tool, offering structured guidance for researchers to navigate validation. On the other hand, can also serve as a

²⁸ Lately, providers such as OpenAI allow researchers to set random seeds and system fingerprints when using models via API (<https://platform.openai.com/docs/guides/text-generation/reproducible-outputs>) to improve reproducibility. However, as the model weights of most models are not public, severe challenges for reproducibility still persist.

²⁹ Link (anonymized)

documentation scheme, streamlining communication between researchers engaged in validation and research consumers.

3.6 Critical Reflection

We developed the ValiText framework with the intention of tackling the manifold challenges associated with validating text-based measures of social constructs. To advance the overall quality of validation practices beyond the framework proposed, however, we want to emphasize some general conditions for successful validation in greater detail.

Transparency and Reproducibility. First and foremost, we want to stress the role of transparency in the validation process. Given that CTAM measures are susceptible to errors and biases inherent to the models themselves, we argue that researchers should openly acknowledge and embrace the potential limitations and biases of their measurement methods. In practical terms, transparency should entail rigorously documenting the validation process and providing access to the data and code utilized. For example, there is a considerable drive in the NLP literature towards documentation standards, both for the textual data (Bender & Friedman, 2018; Gebru et al., 2021; Heger et al., 2022) and the respective measurement model and its parameters (Derczynski et al., 2023; Dodge et al., 2019; Rogers et al., 2021). Moreover, to address the inherent research freedom associated with text analysis, researchers should anchor their work within robust open science standards, particularly embracing preregistration and registered reports, while also establishing the necessary infrastructure to facilitate replication studies (Schoch et al., 2023).

Human-in-the-loop. Second, we are convinced that, at least for now, any CTAM method should still rely on human semantic understanding of language as an absolute benchmark (van Atteveldt et al., 2021; Weber et al., 2018). Human judgement is crucial because too often, computational methods are prone to rely on spurious relations or noise in the data, thereby lacking a deeper ontological sense of error which should prevent trusting empirical measures blindly (see Jankowski & Huber, 2022). However, human annotations do not automatically lead to accurate labels for (training and) testing models. On the contrary, without adequate guidance and training, human annotation bias might result in low-quality data labels (see Geva et al., 2019; Liu et al., 2021), which might be further aggravated by coders' limited focus, fatigue, and evolving interpretations of the underlying conceptual categories (Neuendorf, 2017).

Nevertheless, until today, human understanding remains the most dependable form of annotation that is not easily replaced in the validation process.

Adaptability and Flexibility. Although our proposed framework offers a general validation workflow that is broadly applicable, there may be a need to customize validation strategies and have differing viewpoints on what constitutes sufficient validation evidence in specific contexts. For instance, the way meaning is encoded can vary significantly between languages in multilingual settings (for a detailed discussion of multilingual challenges in validation, see Lind et al., 2023). In this case, we would recommend providing additional validation evidence to demonstrate cultural and linguistic sensitivity.

In addition, adaptability also applies to determining appropriate cut-off values and metrics, which unfortunately lack universal interpretability. As a case in point, consider the F1 score, with values above 0.8 generally considered good, and values below 0.5 considered poor. However, the interpretation of the metric depends heavily on the nature of the construct and contextual factors within the measurement design. Thus, researchers must be able to interpret and adapt their specific cut-off values, such as by comparing it to related work or picking the best-performing model out of several competing models.

3.7 Conclusion

We do not claim that our proposed framework provides a “holy grail” that sweeps away all concerns in the process of validating text-based measures. Nevertheless, we emphasize that our framework offers a much-needed structure for discussions circling around validation, benefiting both researchers conducting validation and research consumers seeking solid documentation. Likewise, we hope that our taxonomy of validation steps provides a starting point for a more normative discussion on validation. Because, as evidenced by the decade-long history of validity in the social sciences, methodological discussions concerning validation commonly echo dominant perspectives and best practices within research communities. Therefore, we strongly advocate for community efforts and critical discussion to agree on generally acknowledged standards for the validation of text-based measures of social constructs. Because increased confidence about the validity of text-based measures will not only aid researchers in presenting their research, but also enable various stakeholders – from public institutions to individual researchers– to evaluate and rely on empirical

evidence derived from text-based research, thus enhancing the overall credibility and impact of empirical studies.

3.8 Appendix

3.8.1 Appendix: Glossary of Validity Terms

Validity Term	(Sub-)type(s) of Evidence	Validation steps
Face Validity	Substantive Evidence - Construct Definition and Operationalization - Design Decisions Structural Evidence - Descriptive Output Inspection	I.1, I.2, I.4, I.5, I.6, I.7
Content Validity	Substantive Evidence - Construct Definition and Operationalization - Design Decisions	I.1, I.2, I.3
Semantic Validity	Substantive Evidence - Construct Definition and Operationalization - Design Decisions	I.1, I.2, I.3
Discriminant Validity	Structural Evidence Systematic Testing	II.10
Convergent Validity	Structural Evidence - Systematic Testing External Evidence - Human-Annotated Test Set Comparison - Surrogate Label Comparison	II.11, III.1, III.2
Concurrent Validity	External Evidence - Human-Annotated Test Set Comparison - Surrogate Label Comparison	II.11, III.1, III.2
Predictive Validity	External Evidence Criterion Prediction	III.3
Criterion Validity	External Evidence Criterion Prediction	III.3
Hypothesis Validity	External Evidence Criterion Prediction	III.3
Construct Validity	Substantive Evidence - Construct Definition and Operationalization - Design Decisions	I.1, I.2, I.4, I.5, I.6, I.7

3.8.2 Appendix: Context-dependent Types of Validation Evidence

Table 6 and 7 present an overview over context-specific validation steps that can further supplement validation efforts. Broadly construed, the *context* can be determined by either the research interest characteristics or the method employed in a particular use case.

3.8.2.1 Context: Data Characteristics

Table 6 displays the context-specific validation steps for varying *research interest* characteristics.

Table 9: Validation Steps for research interest characteristics (context-specific)

ID	Validation Steps	Context	Considerations	Implementation	Applications
Structural Evidence					
Systematic Testing (context-specific)					
V.1	Counterfactual tests	Research Interest: Demonstrating sensitivity of the measurement	Have I tested that my model is sensitive to meaningful changes in the text data?	Evaluating performance metrics (i.e., F1) for new dataset of counterfactual examples	(Garg et al., 2019)
V.2	Adversarial tests	Research Interest: Demonstrating stability of the measurement	Have I tested that my model is resilient to slight perturbations in the text data?	Evaluating performance metrics (i.e., F1) for new dataset of adversarial examples	(Ribeiro et al., 2018)
V.3	Discriminant tests	Research Interest: Demonstrating low correspondence to different constructs	Have I tested that my model is able to distinguish between the construct of interest and similar, but unrelated concepts (e.g., and sexist language)?	Inspecting output scores for a sample of “discriminant” examples	Fang et al. (2023)
V.4	Out of domain tests	Research Interest: Focus on demonstrating generalizability	Have I tested that my model is able to generalize to out-of-domain examples?	Evaluating performance metrics (i.e., F1) for new dataset of out-of-domain examples	(Sen et al., 2022)
External Evidence					
Surrogate Label Comparison (context-specific)					
V.5	Comparison of measures with surrogate labels	Research Interest: Demonstrating correspondence to surrogate labels	Have I reached sufficient predictive performance on the surrogate labels?	Evaluating performance metrics (i.e., F1) for the surrogate labels	Grimmer et al. (2022)
Criterion Prediction (context-specific)					

V.6	Criterion Prediction	Research Interest: Demonstrating predictive performance to related criteria	Have I been able to accurately predict real-word phenomena?	Evaluating predictive metrics (i.e., regression coefficient) for the criteria	Grimmer et al. (2022)
------------	-----------------------------	---	---	---	-----------------------

Structural Evidence

Systematic Testing. Depending on their research interest, researchers may apply systematic tests that encompass various strategies and experimental approaches, all aimed at demonstrating that the model can generate meaningful measures for the construct being measured.

For example, researchers can conduct *counterfactual tests* (V.1), i.e., strategically synthesized versions of the original texts to put the model to test. For example, counterfactual examples could include changing the polarity of sentences, or adding negations (Ilyas et al., 2019). Thus, increased vulnerability to these changes could highlight important shortcomings and, thus, threats to validity. Likewise, *adversarial tests* (V.2) can be utilized to ensure that the model is resilient to input specifically designed to fool a text model, i.e., by adding slight variations to the original data.³⁰ In a similar direction, researchers can also draft *discriminant tests* (V.3) to ensure that the model is able to distinguish the construct of interest from similar, but unrelated concepts. For instance, a classifier that is designed to detect benevolent sexism (i.e., subjectively positive view towards men and woman, see (Jha & Mamidi, 2017)) should not necessarily be correlated with hostile sexism (i.e., explicitly negative attitude towards men and woman), as both concepts are assumed to be independent subdimensions of the overall construct of sexism. On a different note, research can also apply *out of domain tests* (V.4) to explore the model’s ability to generalize to examples come from a different data distribution than the ones a model has seen before.

External Evidence

Surrogate Labels Comparison: Often, the data comes with relevant information that can be linked to the construct being measured. Therefore, researchers might further compare their output scores with *surrogate labels* (V.5) that constitute similar, but closely related characteristics of the data, such as the party label as a proxy measure for a politician’s speech ideology.

Criterion Prediction. In the same vein, researchers might include a validation step to verify that their output scores can forecast *external criteria or real-*

³⁰ <https://textattack.readthedocs.io/en/master/>

world phenomena (V.6). This aims to confirm hypothesized relationships with real-world phenomena, like predicting voting behavior based on textual measures of political ideology (Lauderdale & Herzog, 2016; Rheault & Cochrane, 2020) – much like work performance or school achievement have served as external criteria for tests of cognitive ability or personality in the psychometric tradition.

3.8.2.2 Context: Use-Case Characteristics

Table 7 displays the context-specific validation steps for varying *use cases* characteristics.

Table 10: Validation Steps for use case characteristics (context-specific)

ID	Validation Steps	Context	Considerations	Implementations	A	B	C	D	Applications
Substantive Evidence									
Design Decisions									
C.1	Including examples into the prompt	Method: Selecting Use Case C: Prompting	Have I supplied clear examples of the construct in the prompt?	Outlining the rationale behind the provided example	-	-	✓	-	(Gilardi et al., 2023)
Structural Evidence									
Model Metrics Evaluation									
D. 1	Evaluation of topic coherence metrics	Method: Selecting Use Case D: Topic Modelling	Have I evaluated topic coherence metrics to review co-occurrence of top words?	Calculating topic coherence metrics (i.e., CV) to select the most coherent topics	-	-	-	✓	(Röder et al., 2015)
A.1	Evaluation of number of (N-)tokens matched	Method: Selecting Use Case A: Dictionaries	Have I checked the numbers of matches for my dictionary?	Calculating the mean share of matched words; implementing methods to increase matching (e.g., stemming)	✓	-	-	-	Goet (2019)
Robustness Checks									
C. 2	Rerunning the analysis using different prompts	Method: Selecting Use Case C: Prompting	Have I rerun the analysis with alternative prompt designs?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative prompts	-	-	✓	-	(Gilardi et al., 2023)

Note: A: Dictionaries, B: (Semi-) Supervised Classification, C: Prompt-based classification using LLMs, D: Topic Modelling

For the use case A. *Dictionaries*, researchers can critically *evaluate the number of words matched by the dictionary* (A.1) as a model metric evaluation. If only a

fraction of relevant terms is matched, the researcher might consider implementing methods to increase the number of words matched, such as removing text features or word stemming.

For the use case *C. Prompting*, researchers should provide the *prompts with clear and concrete examples* (C.1) of the construct being measured for substantive evidence. Furthermore, in the robustness checks section, researcher should rerun the analysis using *different prompt templates* and varying prompt designs (C.2).

For the use case *D. Topic Modelling*, researchers can furthermore evaluate *topic coherence metrics* (D.1) that provide guidance on the most coherent number of topics.

3.8.3 Appendix: Illustrative Example: Measuring Sexism in Social Media Comments

Below, we demonstrate the practical application of the ValiText framework through an illustrative example of measuring sexism in social media comments. We revisit and document validation steps in a study conducted by Samory et al. (2019) that measured sexism in social media comments. Specifically, this study used different types of supervised machine learning models to measure sexism for different datasets. We chose the study by Samory et al. (2019) because of their comprehensive exploration and contemplation of the intricacies tied to measuring sexism. We proceed by scrutinizing the validation steps employed in their study, aligning them with the conceptual framework and the checklist for “B. Classification using Fine-tuned Machine-Learning Model”.

3.8.3.1 Substantive Evidence

Starting with *documentation of the conceptual background* (I.1), Samory et al. (2021) provide extensive evidence on their engagement with the relevant literature and other sources of information. For example, the authors discuss existing definitions and attempts to measure sexism using computational methods, conclude that there is definitional unclarity, and reflect on possible biases and spurious artifacts in previous research. Moreover, they evaluate survey measures of sexism (i.e., sets of questions or statements that are used to measure social constructs) from the field of social psychology.

Afterwards, they provide a *justification of the operationalization* (I.2), that is, the link between their construct and the textual data. Because they primarily apply supervised machine-learning methods, the authors’ underlying justification lies in the provision

of high-quality training data that enables the text model to autonomously learn and adapt relevant patterns in the data. To annotate sexism within their training data, the authors develop a detailed codebook based on four subdimensions of sexism identified in the previous literature (e.g., “behavioral expectations” and “endorsement of inequality”). In addition, they extend the codebook to not only differentiate between sexist content, but also between varying degrees of sexist phrasing. To test their operationalization using a codebook empirically, they rely on *manual precoding* (I.3) to label the sexist items from the survey scales into relevant subcategories, finding considerable agreement.

Following this initial stage of construct definition, the authors then discuss their design decisions. They outline their *justification of data collection decisions* (I.4), in particular using different textual datasets. They rely on Twitter data collected through various keywords and strategies and survey scales, while acknowledging the strengths and weaknesses associated with each dataset. In addition, they create a subset of adversarial examples with minimal lexical changes that switches the meaning of sentences from sexist to non-sexist.

Furthermore, for the *justification of method choice* (I.5), the authors rely primarily on a supervised approach. Although implicitly, their argumentation is that only supervised models can replicate human codings that distinguish between different subdimensions of sexism. However, they do not only rely on one specific type of method but rather select a variety of models, such as a Logit model, a CNN and a fine-tuned BERT model with increasing complexity to systematically compare their performance, allowing for a thorough evaluation of measurement performance across different models. Likewise, they also decide to include dictionary baseline models (see the section on robustness checks *rerunning the analysis using alternative text-based methods* (IV.1)).

For justifying the *level of analysis* (I.6), they select the sentence level as the unit of analysis, which is aligned with the literature and the structure of the survey items. For the *justification of preprocessing decisions* (I.7), the authors only provide a detailed description for the Logit model, omitting such details for the other methods.

I D	Validation Step	Documentation	Considerations	Performance Criteria	Source / References
Construct Definition and Operationalization					
I.1	Documentation of the conceptual background	- Reference to existing definitions for sexism - Reference to previous attempts to measure sexism (misogyny, benevolent vs. hostile sexism, etc.)	Have I conducted a literature review or consulted with domain experts to gain a sufficient	Summarizing existing literature on the conceptual	Krippendorff (2018)

		• Discussion of implications of definitional unclarity • Reflection on previous models capturing spurious artifacts of the datasets instead of sexist language. • Systematic engagement with survey scales to measures sexist language. • Considerations of sexist phrasing (not only content), i.e., offensive language (no explanation how this decision might relate to literature)	understanding of conceptual background of the construct?	background of the construct	
I. 2	Justification of the operationalization	• Development of a detailed codebook based on four dimensions identified in the psychological literature. • Discussion of coding inconsistencies + Adaption of the coding instructions	Have I sufficiently explained how the construct should manifest itself in the textual data? Have I documented my operationalization in a codebook?	Providing definition and conceptualization of the construct	Krippendorff (2018)
I. 3	Manual Precoding	• Test of the codebook using 5 MTURKERS (86% agreement for majority verdict (at least 3 out of 5 agreement) on the survey scales	Have I reached sufficient interrater agreement for a subsample of the textual data? Have I ensured that the construct can be detected in the textual data? Have I outlined my rules of coding uncertainty across coders?	Reaching sufficient interrater agreement (e.g., Krippendorff's alpha α)	Krippendorff (2018), Plank (2022)
Design Decisions					
I. 4	Justification of data collection decisions	• Combination of different data sources with different characteristics ○ Scale items from psychological scales ○ Twitter data collected by keywords and human-annotated. ○ Twitter data collected by “call me sexist but” phrase (quite experimental approach) ○ Creation of adversarial examples (“crowd workers to generate adversarial examples, i.e., examples that are a valid input for a machine learning model, strategically synthesized to put the model to test”) • Discussion on the features of the annotated datasets	Have I selected a dataset that is representative and relevant to the research question and population of interest? Have I justified the data selection decisions (e.g., using keywords)? Have I assessed the quality and completeness of the dataset and checked for potential biases or inconsistencies?	Outlining the rationale behind data selection / collection decisions; Documenting potential limitations and data quality issues	Krippendorff (2018)
I.5	Justification of method choice	• Application of different models with increasing complexity (Logit/CNN/Bert) including state of the art methods • Systematic Comparison of these methods	Have I selected the appropriate type of method based on the operationalization of the construct and data characteristics? Have I justified the concrete selection of a particular model?	Outlining the rationale behind method selection; Documenting potential limitations in comparison to alternative methods	Grimmer et al. (2022)

I.6	Justification of the level of analysis	Not explicitly mentioned, but the focus on the sentence level (based on the survey scales) appears plausible in regard to the literature.	Have I selected the appropriate level of analysis? Have I considered potential problems when aggregating measures from lower to higher levels (e.g., sentence to paragraph level)?	Outline the rationale behind the selected level of analysis (e.g., token, sentence, or paragraph level).	Jankowski & Huber (2022)
I.7	Justification of preprocessing decisions	Only for the Logit model, a short reference to the adoption of preprocessing decisions similar to Jha and Mamidi is provided.	Have I justified relevant changes to the text prior to the analysis, such as removing certain words or phrases?	Outlining the rationale behind preprocessing decisions	Grimmer et al. (2022)

3.8.3.2 Structural Evidence

To demonstrate structural evidence, Samory et al. (2021) proceed with a combination of validation steps to examine and evaluate the properties of the model and its output. To evaluate the model properties, they conduct an *inspection of predictive model features* (II.1). To do so, they evaluate the most predictive words for each sexism category (unigrams) and compare them across their data sets and methods applied. Thus, they observe that some models, which are trained on slightly adapted adversarial examples (see Wallace et al., 2019; Zhang et al., 2019), exhibit more general features, which indicates increased model robustness and, partially, performance.

To evaluate the model output, they furthermore conduct extensive *error analysis using data grouping* (II.5) of misclassified examples to identify systematic errors on their most promising BERT model. They specifically investigate the influence of various factors to assess where the model misclassifies messages using a logistic regression model. Relevant factors which they consider are the type of model used (i.e., whether it was trained on original or adversarial examples), or the origin of the training data. Furthermore, they also evaluate the impact of the initial agreement among the coders on the probability of misclassifying errors.

ID	Validation Step	Documentation	Considerations	Performance Criteria	Source / References
Model Feature Inspection					
I.1	Inspection of predictive model features	Conducting of feature importance analysis for predictive unigrams (Table 4)	Have I inspected the predictive features for my model? Have I assured they are conceptually aligned with the	Qualitative evaluation of top-ranked model features using feature-	Molnar (2020), Küpfer & Meyer (2023)

			construct being measured?	importance methods like e.g., LIME or ICE	
Descriptive Output Inspection					
II. 2	Visual inspection of output	Not provided	Have I visualized my output descriptively? Have I identified and visualized outliers and extreme values?	Plotting descriptive statistics; discussing the plausibility of the observed distribution	Goet (2019)
II. 3	Comparison of aggregated measures across known groups	Not provided	Have I aggregated the output scores across known groups (e.g., mean share of sexist sentences across social media user demographics)?	Plotting aggregated measures across groups; discussing the plausibility of the observed distribution	Goet (2019)
II. 4	Qualitatively assess top documents with the highest overall scores for each output category	Not provided	Have I assessed the most outstanding documents for each type of output, such as labels with the highest confidence, or highest and lowest scores on a numerical scale?	Qualitative evaluation to ensure that the top-ranked texts align with the construct	Goet (2019)
Error Analysis					
II. 5	Error analysis using data grouping	detailed discussion of misclassified examples, identification of systematic errors (e.g., varying performance of baseline model for topicality)	Have I conducted error analysis to compare the performance of my model across known subgroups?	Comparing performance metrics (i.e., F1) across subgroups	Wu et al. (2019)
II. 6	Error analysis of outstanding or deliberately chosen observations	Not provided	Have I conducted error analysis to qualitatively evaluate the sources and types of errors associated with the measures?	Exploring the underlying causes of misclassifications by qualitatively screening misclassified examples	(Wu et al., 2019)
Systematic Testing (context-specific)					
V. 1	Counterfactual tests	Conducting counterfactual tests; providing new training samples of counterfactual tests and displaying performance metrics (F1 score).	Have I tested that my model is sensitive to meaningful changes in the text data?	Evaluating performance metrics (i.e., F1) for new dataset of counterfactual examples	(Garg et al., 2019)
V. 2	Adversarial tests	Not provided	Have I tested that my model is resilient to slight	Evaluating performance metrics	(Ribeiro et al., 2018)

			perturbations in the text data?	(i.e., F1) for new dataset of adversarial examples	
V. 3	Discriminant tests	Not provided	Have I tested that my model is able to distinguish between the construct of interest and similar, but unrelated concepts (e.g., and sexist language)?	Inspecting output scores for a sample of “discriminant” examples	Fang et al. (2023)
V. 4	Out of domain tests	Not provided	Have I tested that my model is able to generalize to out-of-domain examples?	Evaluating performance metrics (i.e., F1) for new dataset of out-of-domain examples	(Sen et al., 2022)

3.8.3.3 External Evidence

Do demonstrate external evidence, Samory et al. (2019) primarily rely on the *comparison of measures with a human-annotated test set* (III.1). To calculate classification performance, they apply k-fold cross-validation and report F1 scores. The evaluation of F1 score is widely regarded as the most viable metric, as alternative metrics such as accuracy (i.e., the overall ratio of positive predictions) can be misleading when dealing with imbalanced data (Spelmen & Porkodi, 2018).

ID	Validation Step	Documentation	Considerations	Performance Criteria	Source / References
Construct Definition and Operationalization					
III.1	Comparison of measures with human-annotated test set ("gold-standard data")	Systematic comparison with hand-annotated test set, report of F1 scores	Have I reached sufficient predictive performance on a test set of held-out human annotations? Did I apply cross-validation to calculate average performance metrics?	Evaluating performance metrics (i.e., F1) for dataset of human annotations	(Samory et al., 2021)
Surrogate Label Comparison (context-specific)					
V. 5	Comparison of measures with surrogate labels	Not provided	Have I reached sufficient predictive performance on the surrogate labels?	Evaluating performance metrics (i.e., F1) for the surrogate labels	Grimmer et al. (2022)
Criterion Prediction (context-specific)					

V. 6	Criterion Prediction	Not provided	Have I been able to accurately predict real-world phenomena?	Evaluating predictive metrics (i.e., regression coefficient) for the criteria	Grimmer et al. (2022)
------	----------------------	--	--	---	-----------------------

3.8.3.4 Robustness Checks

Besides validation steps that provide validation evidence, Samory et al. (2019) conduct a series of robustness checks. For instance, the authors rerun their analysis using alternative text-based methods (IV.1). The inclusion of a simple baseline model (toxicity and gender word dictionary) demonstrates that their supervised models achieve higher overall performance on the hand-annotated test set, providing evidence that their supervised models are more successful in replicating human-annotated measures of sexism.

ID	Validation Step	Documentation	Considerations	Performance Criteria	Source / References
Construct Definition and Operationalization					
IV .1	Rerunning the analysis using alternative text models	Comparison with several baseline models	Have I rerun the analysis with alternative text-based methods, such as a baseline model?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative measurements	(Samory et al., 2021)
IV .2	Rerunning the analysis using different hyperparameter settings	Not provided	Have I rerun the analysis with alternative hyperparameter settings?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative hyperparameter settings	(Arnold et al., 2023)
IV .3	Rerunning the analysis using different cutoff thresholds	Not provided	Have I rerun the analysis with alternative cutoff thresholds?	Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1))	Grimmer et al. (2022)

				for alternative cutoff thresholds	
IV .4	Rerunning the analysis using different text cleaning and preprocessi ng steps	Not provided	Have I rerun the analysis using alternative data cleaning and preprocessing settings (e.g., removing certain phrases or features of the data)?	Displaying performan ce metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative preprocess ing steps	Grimmer et al. (2022)

4. Framework Application 1: Personality Traits

The third part of the dissertation demonstrates how the proposed validation framework can be applied to improve the validation of CTAM measures of social constructs. It addresses the guiding research question: How can the framework be applied?

To that end, this chapter presents the first of two application chapters: a multi-method validation study of text-based measures of politicians' public personalities, operationalized through the presence of personality cues in their public communication. In this study, I develop and validate fine-grained measures for the traits of agency and communion, using models such as RoBERTa, DeepSeek-V3, and GPT-4o. The validation relies on multiple strategies, including comparisons with human annotations, the design of systematic tests, and sensitivity analyses of political statements by German politicians as well as the 2024 U.S. presidential debate between Donald Trump and Kamala Harris.

This chapter is based on the following publication:

Birkenmaier, L., & Lechner, C. (2025). Measuring Politicians' Public Personality Traits using Computational Text Analysis: A Multi-Method Feasibility Study for Agency and Communion https://doi.org/10.31219/osf.io/7nhjz_v1

Birkenmaier, L., Sieber, L., & Bergstein, F. (2025). PolInterviews--A Dataset of German Politician Public Broadcast Interviews. arXiv preprint arXiv:2501.04484. <https://doi.org/10.48550/arXiv.2501.04484>

4.1 Introduction

The way citizens judge politicians is influenced by how they evaluate politicians' character or personality traits (Bittner & Peterson, 2018; Hayes, 2009; Laustsen & Bor, 2017). A more favorable evaluation of a politician's personality increases the likelihood of voting for them and contributes to the formation of positive partisan attitudes (Funk, 1997; Miller et al., 1986). The increasing personalization of politics renders voters' perceptions of politicians' personality even more important, as individual political actors become more prominent at the expense of collective actors and institutions such as political parties (Pedersen & Rahat, 2021; Rahat & Kenig, 2018).

Whereas a vast literature has studied which personality traits voters look for in politicians (see Van Aelst et al., 2012 for an overview), less is known about the ways in which politicians project their own personality traits in public communication, such as through speeches or social media posts. Previous studies have explored various aspects of politicians' strategic self-presentation, including their style of representation (Birkenmaier, Sältzer, et al., 2025; Fenno, 2003; Gulati, 2004; Schürmann & Stier, 2022) or their private lives (Colliander et al., 2017; Perloff, 2013; Stanyer, 2008; Van Zoonen, 2005). However, the role of personality traits in politicians' self-presentation to voters has received limited attention in research.

Generally, psychology conceptualizes personality as relatively stable patterns of thoughts, feelings, and behaviors that distinguish individuals from one another and that influence their interactions with others. Personality is typically conceptualized as a set of global traits organized within frameworks such as the Big Five (Costa Jr. & McCrae, 1995) or HEXACO (Ashton & Lee, 2007). When it comes to studying politicians' personalities, the task becomes more complex because observers typically only have access to the politician's *public* personality. This public personality is not solely a reflection of inherent personality traits but is also influenced by (unobservable) strategic considerations to emphasize or downplay certain traits to appeal to the public. This does not necessarily present a problem for studying politicians' public personalities, as voters' perceptions are primarily shaped by these public personas, making them a relevant and meaningful subject of research. However, accurately measuring these public personality traits based on politicians' public behavior—indicated through verbal or non-verbal *personality cues*—is challenging. This is due not only to the inherently fuzzy nature of personality traits but also to the fact that the cues reflecting them are often ambiguous and context-dependent. Consequently, we still know relatively little about how validly these public personalities can be measured.

Leveraging recent advances in computational text analysis and political psychology research, our study seeks to validate the detection of *verbal* personality cues as measurable indicators of the politician's public personality. We build upon a validated framework and survey inventory to assess politicians' personalities and apply different methods and strategies to operationalize personality cues using a corpus of three distinct types of data sources, namely social media posts, political speeches, and political interviews. Our results suggest prompt-based approaches using state-of-the-art models, such as DeepSeek-V3 and GPT-4o, show the strongest evidence of measurement validity.

This paper makes five important contributions: (1) it provides a conceptual model of the relationship between politicians' intrinsic and public personality, alongside the personality cues through which the public personality can be measured; (2) it introduces a methodological approach to detect *verbal* personality cues for two traits – agency and communion – using computational text analysis; (3) it systematically compares multiple models and measurement strategies and provides extensive validation using a systematic framework; (4) it presents compelling evidence that computational text analysis can be used to measure politicians' public personality from verbal cues; and (5) it provides materials, including trained models, code, and data, and discusses how the methods applied can advance political psychology.

4.2 Literature

4.2.1 Politicians' Personality Traits Matter

Politicians' personality plays a central role in explaining political actions, such as partisan attitudes or voting behavior (Federico & Malka, 2018; Schoen & Schumann, 2005). Whereas politicians and their characteristics have always played a key role in politics, there are growing claims that a new, increasingly personalized political culture of “candidate-centered politics” has emerged (McAllister, 2007; Pedersen & Rahat, 2021). This shift is attributed to a decline in the traditional emotional connections between voters and political parties (Dalton et al., 2000; Mair, 2005) and the transformative impact of a media landscape that prominently includes decentralized social media (Stier et al., 2018b). The growing interest in the characteristics of politicians has led to an increased focus on analyzing politicians' personalities.

For example, previous research has examined the role of politicians' personality as a means to understand their ideological orientations and political behavior (George & George, 1998; Walker et al., 1998).³¹ This strand of research indicates that conservative politicians often show lower levels of agreeableness and openness but higher levels of conscientiousness, whereas the link to other personality traits is less clear (Hanania, 2017; Schumacher & Zettler, 2019). Overall, evidence also indicates that politicians, compared to the general population, appear to show higher levels of

³¹ This branch of research is based on the observation that personality traits are often linked to ideology (Gerber et al., 2010). For instance, traits like conscientiousness are often linked with conservative ideologies, whereas openness tends to correlate with liberal views (Carney et al., 2008).

agreeableness, emotional stability and openness (Caprara et al., 2003). Likewise, in-office performance (Joly et al., 2019; Lilienfeld et al., 2012; Watts et al., 2013), media visibility (Amsalem et al., 2020), or electoral success (Nai, 2019; Pillai et al., 2003; Scott & Medeiros, 2020) are also related to personality, particularly through higher levels of extraversion and lower levels of agreeableness.

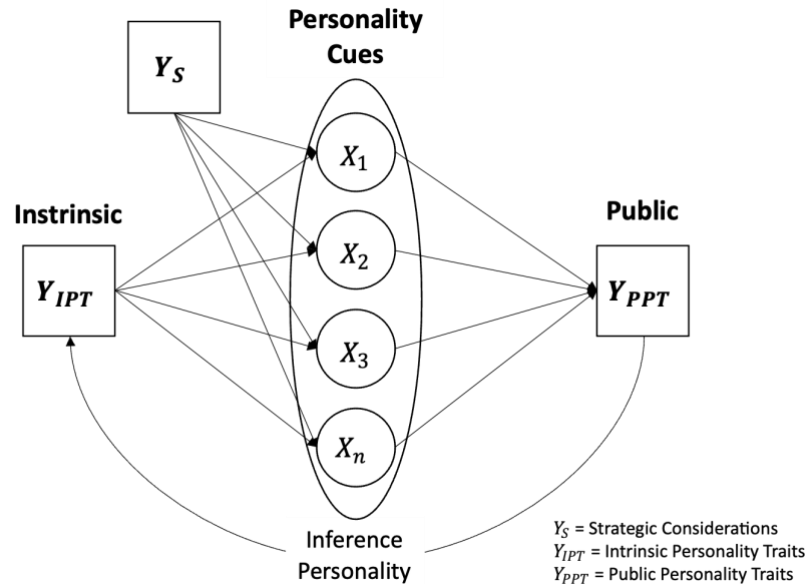
Voter preferences also reflect personality-based judgments. For instance, voters appear to consistently prefer characteristics such as competence, integrity, or energy in political figures (Aichholzer & Willmann, 2020; Mondak, 1995). Caprara and Zimbardo (2004) further elaborate that voters' evaluation of politicians often depends on how closely politicians resonate with the voters' own traits and values. This can manifest through emotional attachments (likability heuristic) or the attribution of positive characteristics (dispositional heuristics). Cross-national studies confirm varying preferences for traits such as emotional stability, extraversion, and conscientiousness in ideal politicians (Aichholzer & Willmann, 2020).

4.2.2 Politicians Strategically Display Personality Cues

The study of politicians' personalities differs fundamentally from personality research in the general population (Immelman, 1993). Whereas personality research in the general population assumes that personality traits are relatively stable and consistently expressed across different contexts, the key assumption in the political domain is that politicians engage in impression management, that is, they actively manage and adjust their public persona depending on the audience and situation in an order to appeal to their constituency (Fenno, 2003; Goffman, 2016). Fenno (2003), for instance, describes how politicians show different behaviors when they interact with voters in their district ("home style") or when they interact with other members of the political system ("Washington style"). The rise of social media further contributes to this dynamic, as social media not only enhances politicians' visibility but also allows them to shape their public persona in a more nuanced and controlled way (Colliander et al., 2017; Meeks, 2016; Stier et al., 2018b). Thus, as the electorate's political attitudes are continually monitored by public opinion polling, politicians are all too aware of the need to strategically emphasize or downplay some aspects of their displayed personality to appeal to voters (Nai et al., 2022; Stier et al., 2018b). This poses a fundamental measurement problem: how can we validly and reliably capture the personality that politicians communicate publicly, knowing it results from both their

intrinsic personality and deliberate self-presentation? Figure 13 outlines this problem by conceptualizing the relationship between the true *intrinsic* personality and the *public* personality using an adaption of the lens model by Brunswik (1952).

Figure 13: Conceptual Model (Adapted from Brunswik, 1955)



In the case of a politician's personality, the unknown criterion is the actual intrinsic personality (Y_{IPT}) of a politician. What voters and researchers alike have access to, however, is the publicly displayed personality (Y_{PPT}) which results from the observable personality cues $X_1...X_n$ communicated by politicians. These cues can be verbal, such as statements in speeches or debates, or non-verbal, such as rhetorical gestures or facial expressions.

Crucially, the public personality Y_{PPT} is not a mere reflection of intrinsic personality Y_{IPT} but is typically shaped by a combination of intrinsic personality traits and strategic considerations Y_S , as politicians consciously emphasize or suppress certain cues to appeal to voters. While the exact mixture of intrinsic dispositions and strategic behavior remains unknown, this is not a limitation studying the way politician present themselves. On the contrary, the public personality, shaped by both intrinsic and strategic factors, is precisely what voters perceive and use as a basis for judgment formation. The key challenge is, rather, to assess whether the personality cues $X_1...X_n$ can be reliably defined, identified, and interpreted as valid indicators of the public

personality Y_{PPT} and whether they can be systematically aggregated into coherent personality traits.

4.2.3 Personality Cues are Inherently Ambiguous

Identifying personality cues and linking them to coherent traits is inherently challenging due to two main reasons.

First, personality traits themselves are not sharply defined or neatly separable—they have the nature of fuzzy concepts (Thielmann & Hilbig, 2019). Although structural models like the Big Five or HEXACO provide a useful framework, these models are based on factor-analytic approaches that reduce correlated questionnaire items into higher-order dimensions. The resulting traits are statistical factors that are theoretically orthogonal (i.e., uncorrelated), but in practice, they often show substantial correlations and overlapping indicators (Zettler et al., 2020). This is evident in substantial cross-loadings within factor structures, indicating that certain indicators reflect more than one trait at a time. As such, traits like extraversion and agreeableness do not always have clear-cut boundaries, making it difficult to determine where one ends and another begins.

Second, the observable manifestations of these traits - in our case the personality cues $X_1...X_n$ - are themselves ambiguous. A single behavior can be driven by multiple traits (a phenomenon known as equifinality), and conversely, a single trait can result in a wide range of behaviors (multifinality). This means that very few behaviors can be considered “pure” indicators of a single trait; instead, they tend to remain ambiguous stimuli open to interpretation. In political contexts, these ambiguities are amplified by the viewer's own perspective. Partisan or cultural biases can shape how observers interpret the same cue—what one supporter sees as “bold” or “charismatic,” an opponent might perceive as “reckless” or “showboating”. In addition, not all traits are equally observable. Traits such as extraversion often have relatively clear behavioral cues—like a politician’s public speaking style or social engagement—whereas others, such as openness or honesty, are harder to detect. Research on “thin-slice” judgments finds that observers can judge some cues that relate to traits like

extraversion much more accurately than others (like neuroticism) from brief interactions (Tackett et al., 2017) or verbal cues (Mairesse et al., 2007).³²

In sum, personality as a construct is both conceptually and empirically fuzzy, and its behavioral indicators are often ambiguous and context-dependent. This makes the reliable measurement of personality in public discourse particularly demanding.

4.2.4 How Can We Measure Personality Cues?

Due to these challenges, methods for measuring politicians' public personalities based on personality cues face considerable limitations (McDonald, 2008). Broadly speaking, these methods can be categorized into self-report and observer-report approaches.

Self-report approaches include participants being asked to fill out surveys (Caprara et al., 2003; Joly et al., 2019; J. Maier et al., 2023).³³ However, politicians – especially high-profile politicians – are a hard-to-survey population, and surveys among politicians often suffer from low response rates, which not only make data access challenging but may also lead to biased response patterns. Likewise, time limitations frequently lead researchers to use brief personality inventories, which may fail to fully represent the complexity of a politician's personality profile. On top of that, there is an inherent limitation that only contemporary legislative periods – where politicians can be contacted – can be studied.

On the other hand, observer-report approaches typically rely on observer ratings derived from various forms of observational data, such as everyday interactions or digital traces from social media. In the study of politicians' personalities, this can be done using ratings by experts or the general public (de Vries & van Prooijen, 2019; Rubenzer et al., 2000; Visser et al., 2017; J. D. Wright & Tomlinson, 2018). Although these evaluations are explicitly based on observable personality cues ($X_1 \dots X_n$), they can be skewed by the raters' political beliefs, and these raters have, by definition, only a partial and biased view of the politicians they are evaluating (J. D. Wright & Tomlinson, 2018). In addition, the interpretation and aggregation of personality cues into traits by

³² Mairesse et al. also experimented with multiple computational methods, finding consistent evidence that cues for traits such as extraversion are more easily identified.

³³ Although these surveys are designed to provide a direct assessment of politician's intrinsic personality, in the political domain, they are highly susceptible to social desirability bias and "faking good", meaning politicians are likely to portray themselves in a more favorable light unintentionally or intentionally. Consequently, we argue that politicians' self-reports can also be seen as a special type of self-statement personality cues rather than as accurate reflections of intrinsic traits.

raters often hides the direct connection between specific cues and the final evaluation, hampering the observation of short-term changes and specific settings.

4.2.5 The Need for Computational Approaches

In this paper, we argue that computational text analysis provides a valuable yet underutilized tool for assessing politicians' self-presentation. Generally, political communication serves as the primary medium for self-presentation, with every statement made by a politician potentially being a personality cue that voters perceive. Although research has aimed to study political self-presentation for verbal cues, much of it relies on content analysis that usually inhibits a targeted, case-study-specific angle for politicians like Donald Trump and Hillary Clinton (Lee & Lim, 2016) or George Bush and Mikhail Gorbachev (Winter et al., 1991). However, while manual content analysis likely provides the most valid research approach in a case-study design, it is unnecessarily effortful for analyzing politicians' public images on a large scale, given that automated personality prediction from digital footprints has demonstrated performance comparable to traditional approaches (see meta-analysis by Azucar et al., 2018).

A notable exception is the study by Ramey et al. (Ramey et al., 2019), who used computational techniques to measure the personalities of U.S. Congress members based on their floor speeches. Their approach builds on a rich literature that uses tools like Linguistic Inquiry and Word Count (LIWC) to capture linguistic features-such as the use of second-person pronouns, punctuation, or word length-and link them to personality traits (Chen et al., 2020; Mairesse et al., 2007; G. Park et al., 2015). Using the pretrained Support Vector Machines for Regression models from Mairesse et al. (2007) and applying it to U.S. Congress floor speeches, they found that their text-based measures were correlated with survey response rates for members of Congress.

While the work of Ramey et al. represents a significant contribution to the field by offering a computational method to extract politicians' personality cues, it also has limitations. First, their reliance on a pretrained support vector machine model raises concerns about the validity of their findings, particularly since the authors did not report a validation against human-annotated data-a gold standard in computational text analysis (Birkenmaier et al., 2023a; Grimmer et al., 2022). Furthermore, their methodological choice is challenged by the recent rapid advancements in natural language processing, where more sophisticated large language models (LLMs) have

consistently shown to provide greater capabilities. Finally, although the Big Five personality traits used in their analysis are widely recognized for general personality assessment, they lack the nuanced, domain-specific application for evaluating politicians' personality characteristics.

Building on the work of Ramey et al., we, therefore investigate whether computational methods can be used to develop more valid, domain-specific automated measures of politicians' public personalities based on the personality cues they exhibit.

4.3 Method

4.3.1 Data

We compiled a corpus comprising a diverse set of public statements from leading German politicians covering the period from 2017 to 2024, encompassing two full legislative terms of the German Bundestag. Our definition of "leading politicians" includes those who served as leaders or co-leaders of both legislative factions and political parties, covering all six parties represented in the German Bundestag. For these 46 politicians (see Appendix A), we collected data from three sources commonly used to study politicians' public communication: interviews, social media posts, and parliamentary speeches.

Interviews: First, we rely on a novel dataset comprising 99 interviews with political leaders broadcasted in five different media formats (Birkenmaier, Sieber, and Bergstein, 2025). Interview statements provide one of the most natural and authentic sources of politicians' interactions and self-presentation to the public, ideally lending themselves to capture the public personality of politicians (Schoonvelde et al., 2019).

Social Media Posts: Second, we rely on social media data as another highly popular resource for politician's public communication. The initial data consists of a subsample of 300.000 Facebook posts from a raw corpus of 6.7 million politicians' posts utilized in Birkenmaier, Sältzer and Wurthmann (2025) based on a list of Twitter Accounts by Sältzer, Stier and Bäuerle (2023).

Parliamentary Speeches: Finally, we rely on a corpus of politicians' speeches in the German Bundestag. We utilize a subset of German MP speeches provided by the Parlspeech database, which provides transcripts of parliamentary speeches of nine representative democracies (Rauh & Schwalbach, 2020). We used only the German subset for our present analysis.

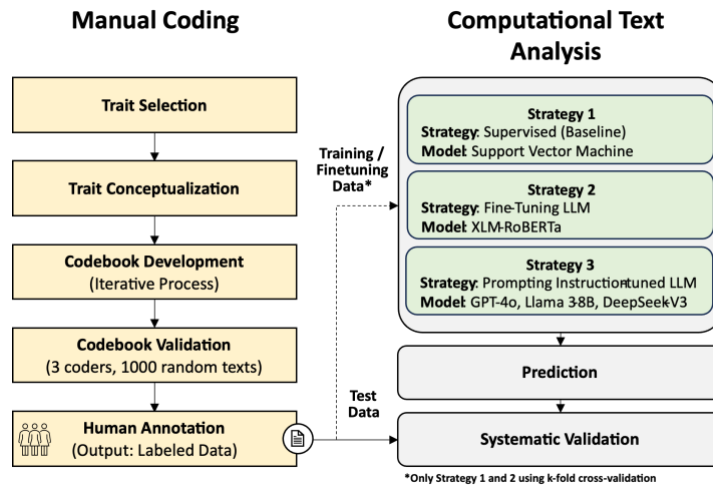
Across all data sources, we defined individual sentences and statements uttered or written by politicians as our unit of analysis. Sentences represent personality cues at a granular level that can be detected and potentially aggregated across texts. Because our primary focus is on validating the detection and measurement of personality cues in text rather than analyzing explanatory factors influencing changes or differences in the public personality, we draw a random sample of 15,000 sentences from each data source as our main corpus.

4.3.2 Measurement

4.3.2.1 Workflow

Figure 14 illustrates our workflow. As detailed below, we begin with manual coding to develop and validate a codebook and create a set of labelled data. We then employ computational text analysis using different models and measurement strategies and assess the validity of our measures using a comprehensive validation framework (Birkenmaier, Wagner, et al., 2024).

Figure 14: Workflow



4.3.2.1.1 Trait Selection

To select personality traits that can be measured using computational text analysis, we focus on two core psychological traits that underpin most personality frameworks: **agency** and **communion** (Bakan, 1966). These so-called “meta-traits” provide a higher-order structure that captures common variance among lower-order traits, including the Big Five, and consistently emerge across diverse personality models (Partsch et al., 2022; Saucier et al., 2014). Adopting the terminology of Hogan (1982),

agency can be thought of as the drive to “get ahead,” whereas communion reflects the drive to “get along.” These dimensions not only underpin widely recognized personality frameworks like the Big Five, but they also constitute key behavioral determinants in both personal and political environments. In a U.S. political context, for instance, research has found that Republicans often place greater emphasis on agency, while Democrats place greater emphasis on communion (Eriksson, 2018).

To guide our measure of agency and communion, we rely on a validated framework and survey inventory to assess politicians’ personalities by Lechner et al. (2025). This inventory comprises 15 domain-specific personality facets specifically tailored to the personas of politicians. Compared to global and general personality frameworks such as the Big Five (Costa Jr. & McCrae, 1995), the dimensions in the Framework by Lechner et al. offer a validated, domain-specific framework for politicians’ personality traits, providing both a clear definition and a conceptual basis for subsequent computational text analysis. In our analysis, we focus on the two prototypical traits from their framework that capture the essence of agency and communion, respectively. This keeps our method interpretable, theoretically grounded, and capable of focusing on the validation of measuring these traits using personality cues in textual data.

Agency: In the Lechner et al. inventory, the trait *Assertiveness & Will to Take Control* is the key trait for agency. It reflects the *willingness to be a leader as well as to achieve power and one’s assertiveness*. The items for this trait characterize a politician who is “*power-hungry, claims leadership*”, “*assertive, dominant*,” and “*a born leader*.”

Communion: In the Lechner et al. inventory, *Empathy & Compassion* constitutes the key trait representing communion. It reflects the ability to express and act on feelings of empathy and compassion toward others. The items for this trait characterize a politician who is “*empathetic, shows genuine compassion*”, “*understands the needs of disadvantaged people*” and “*is compassionate, has a soft heart*.”

We focus on the positive poles of the prototypical traits introduced above, but not the opposing poles “Passivity & Submissiveness” (-communion) and “Apathy & Indifference” (-agency). This reflects our assumption that politicians usually aim to project personality cues that reinforce a strong and favorable public image.³⁴

³⁴ For readability purposes, we subsequently refer to them to “agency” and “communion” instead of the prototypical traits “Empathy & Compassion” and “Assertiveness & Will to Take Control”

4.3.2.1.1.2 Trait Conceptualization

For each trait, we define both *weak* and *strong* cues that reflect varying degrees of explicitness in the representation of personality cues. Personality cues can differ in both strength and relevance, as they vary in how clearly and accurately they reflect the traits being judged in real-world context (Funder, 1995). In our case, strong cues refer to clear and direct statements that contain self-disclosure or direct references to a personality trait ascribed to oneself (e.g., “I am a very stable genius”), whereas weak cues are more ambiguous or indirect and might be more open to subjective interpretation by the receiver of the cue. **Agency, for example,** can be explicitly signalled through strong statements such as “*As the leader of this party, I will ensure that this policy is implemented, regardless of who stands in the way*”. In contrast, weaker agency cues might include phrases that can potentially be subject to human interpretation, such as “*We will push this policy forward and ensure its implementation*”. While the latter statement suggests a call to action, it is much more passive and does not clarify whether the action entails the speaker actively overcoming resistance.

4.3.2.1.1.3 Codebook Development & Validation

We developed the codebook in multiple steps. Initially, we limited coding to sentences containing elements of personal attributes using self-referential words as filtering keywords (e.g., “I,” “we,” “mine”; see Appendix B). Coders annotated these sentences while providing continuous feedback to refine the codebook. Once coders achieved a consensus on trait assignments, we expanded the coding to a random selection of sentences from the larger dataset to improve generalizability. Ultimately, a team of three annotators (i.e., the first author and two research assistants) annotated 1000 sentences. We computed Krippendorff’s Alpha as a measure of multi-label agreement, with values above $\alpha > .70$ deemed as acceptable (Krippendorff, 2018). After demonstrating sufficient agreement on the 1000 sentences (see Results section), we proceeded to generate a labeled corpus containing at least 200 examples per class (“agency,” “communion,” or “none”) for both weak and strong cues.

4.3.2.2 Computational Analysis

With the growing application of LLMs in the social sciences, there is an ongoing debate concerning the best strategies for computational text classification (Gilardi et al., 2023; Ziems et al., 2024). For the reasons outlined earlier (i.e., the fuzziness of personality

traits as such and the ambiguity of verbal cues), classifying sentences based on personality trait cues is especially challenging, and no clear methodological standards exist. Therefore, in this paper, we systematically test different strategies and computational models to determine the best-performing architecture to detect personality cues in text data.

For all models, we test their performance on different dataset variants. In the first dataset variant, referred to as *strong*, we only include examples of strong cues for agency and communion. In the second variant, referred to as *weak+strong*, we include an equal proportion of weak and strong cues for agency and communion, representing a more inclusive (but potentially less clear-cut) definition of personality trait expressions. To address class imbalance, we assess performance on both a balanced dataset, where the three classes (“agency”, “communion”, and “none”) are equally represented, and an unbalanced dataset, where “none” appears twice as often as the other classes.

We compare three major strategies across multiple model architectures (see Figure 14) and assess performance on different dataset variants. All hyperparameters and computational resources used in our experiments are documented in Appendix D and E.

For the supervised strategy (Strategy 1), we follow the approach by Ramey et al. (2019) and train a **Support Vector Machine** (SVM) as a baseline model on the human-annotated data. SVM are fast, cheap, and computationally efficient, as they are only trained on the document-term feature matrix of the training data. We compute 5-fold cross-validation, using a train-test split of 80/20.

For fine-tuning a pretrained LLM (Strategy 2), we fine-tune a 279M parameters **XLNet-RoBERTa (base)** model (Conneau et al., 2020) using common hyperparameters. We compute 3-fold cross-validation, using a train-test split of 80/20.

For the prompting approach (Strategy 3), we prompt three different generative instruction-tuned LLM, following current research exploring generative models for classification (Gilardi et al., 2023; Ziems et al., 2024). We rely on three models: a proprietary instruction-tuned **GPT-4o**, and two open-source models **Llama 3-8B** (Touvron, Lavril, et al., 2023) and **DeepSeek-V3** (DeepSeek-AI et al., 2024). For each model, we utilize different prompting templates tailored to dataset variants featuring either exclusively strong or a combination of strong + weak cues, as outlined in our codebook. Additionally, our approach distinguishes between zero-shot and few-shot

settings, with the latter including class-specific examples (see Appendix F for full prompts). Each prompt directs the respective model to classify politicians’ statements into one of three categories (“agency”, “communion”, and “none”) and to provide a brief, one-sentence explanation for its decision in a structured format.

To mitigate the effects of randomness in training and prompting, we run each of our models with three different random seeds and report the macro-averaged F1 performance across runs.

4.3.3 Validation

Due to the foundational nature of our research, we take a systematic approach to validation and rely on the ValiText framework (Birkenmaier et al., 2024), which distinguishes between different types of validation evidence for computational text analysis. We start by reporting and documenting the theoretical framework used to operationalize our constructs, provide interrater agreement measures for human coders, and document major design decisions, such as the rationale behind our choice of methods (substantive validation). We then evaluate how our measures correspond to human annotations (external validation) and further explore the properties and the output (structural validation). Specifically, this involves designing functional tests and analyzing the sensitivity for partisan differences across (a) the complete unlabeled corpus of 45,000 randomly selected sentences from the German Bundestag and (b) around 2,000 translated statements from the 2024 U.S. presidential debate between Donald Trump and Kamala Harris. We document all our validation steps using checklist templates (see Appendix G).

4.4 Results

4.4.1 Substantive Validation

As outlined in our method section, we employ a domain-specific personality framework and attendant survey inventory as the conceptual basis for operationalizing personality cues. As detailed in Lechner et al. (2025), the substantive validity of this personality framework was ensured during its development process through multiple means, including an analysis of extant work on personality assessment in general and politicians’ personality traits more specifically; input and reviews by experts from

psychometrics and political science; as well as extensive tests of structural and external validity of the resulting inventory in multiple samples of voters.

We then chose the prototypical traits of the framework to operationalize personality cues for agency and communion. To do so, we developed a codebook and conducted manual coding to evaluate whether human annotators could reliably identify corresponding personality cues. Our analysis demonstrated sufficient agreement, with Krippendorff's Alpha of .73 (strong) and .72 (weak + strong) for the overall identification of the presence of personality cues, and agreement ranging from .70 to .78 for “agency” and “communion” (see Table 11).

Table 11: Interrater Agreement for Human Coding (n=1000)

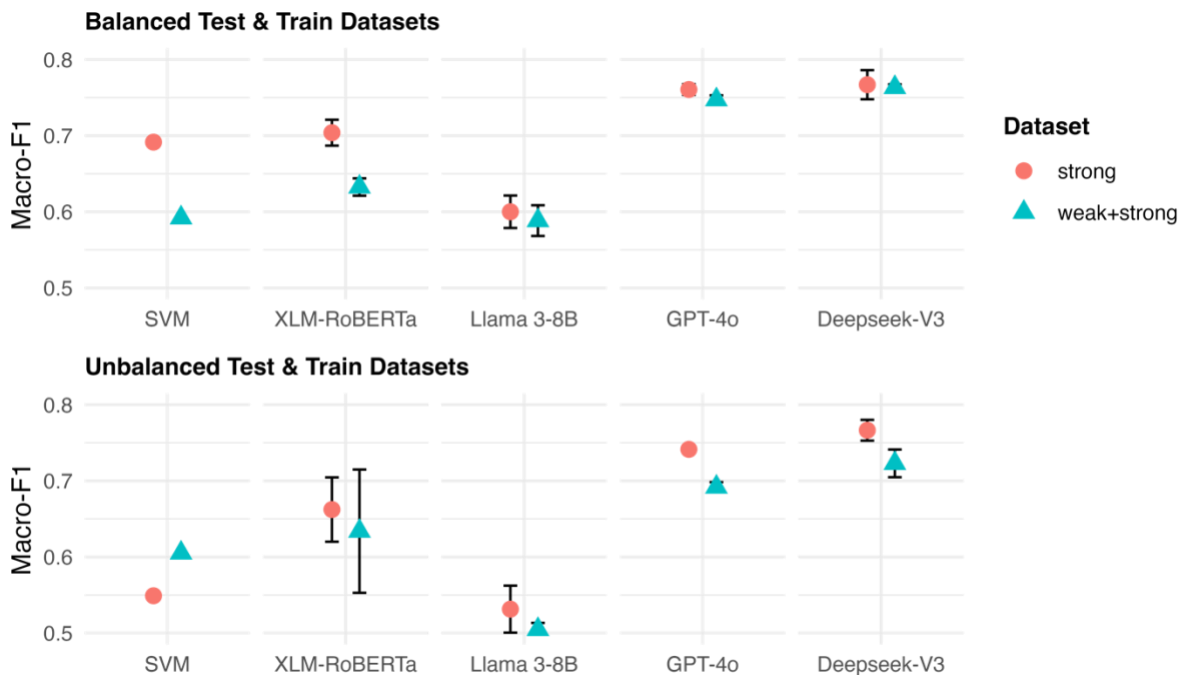
Category	Strong Cues	Weak + Strong Cues
Detecting Personality Cue Personality (A or C) vs. No Personality Cue	0.73	0.72
Agency Agency vs. No Agency Cues	0.70	0.71
Communion Communion vs. No Communion Cues	0.78	0.74

Regarding label distribution, most sentences were classified as “none,” with 91% for strong cues and 83% for combined weak + strong cues. Among sentences containing personality cues, “communion” appeared more frequently (approximately 5% of strong cues and 8% of weak cues) than “agency” (around 3% of strong cues and 6% of weak cues).

4.4.2 Comparison with Human-annotated Test Set (External Validation)

Figure 15 displays the performance of all models across balanced and unbalanced test and training datasets.³⁵ **Red points** indicate results on the *strong* datasets that only contain strong personality cues and **blue triangles** indicate results on the **combined** *weak + strong* datasets with an equal proportion of weak and strong personality cues. Overall, our results show that DeepSeek-V3 and GPT-4o

Figure 15: Classifier Performance for Balanced and Unbalanced Datasets



demonstrated the highest performance across dataset variants. As shown in Table 12, DeepSeek-V3 performed best on both balanced and unbalanced datasets, attaining a macro-F1 score of 0.77. GPT-4o performed similarly well, achieving a macro-F1 score of 0.76 on the strong-balanced dataset, and slightly worse performance on the balanced and unbalanced dataset. In contrast, the Llama 3-8B model showed significantly lower performance, with the highest F1 scores reaching approximately 0.60 for the balanced dataset and 0.55 for the unbalanced dataset, along with greater variability across training runs.

³⁵ As outlined in the method section, only supervised (SVM) / semi-supervised (XLM-RoBERTa) models relied on training / finetuning data, whereas prompting only included a selection of examples in the prompt (few-shot prompting)

Table 12: Overall Best-Performing Models on the Test Data (3 Random Seeds)

Model	Strategy	Test Dataset	Prompt-Strategy	Macro F1	SD
Balanced					
DeepSeek-V3	3. Prompting	Strong	few	0.77	0.02
DeepSeek-V3	3. Prompting	Weak + Strong	few	0.76	0.01
GPT-4o	3. Prompting	Strong	few	0.76	0.01
Unbalanced					
DeepSeek-V3	3. Prompting	Strong	few	0.77	0.01
GPT-4o	3. Prompting	Strong	few	0.74	0.00
DeepSeek-V3	3. Prompting	Strong	zero	0.74	0.01

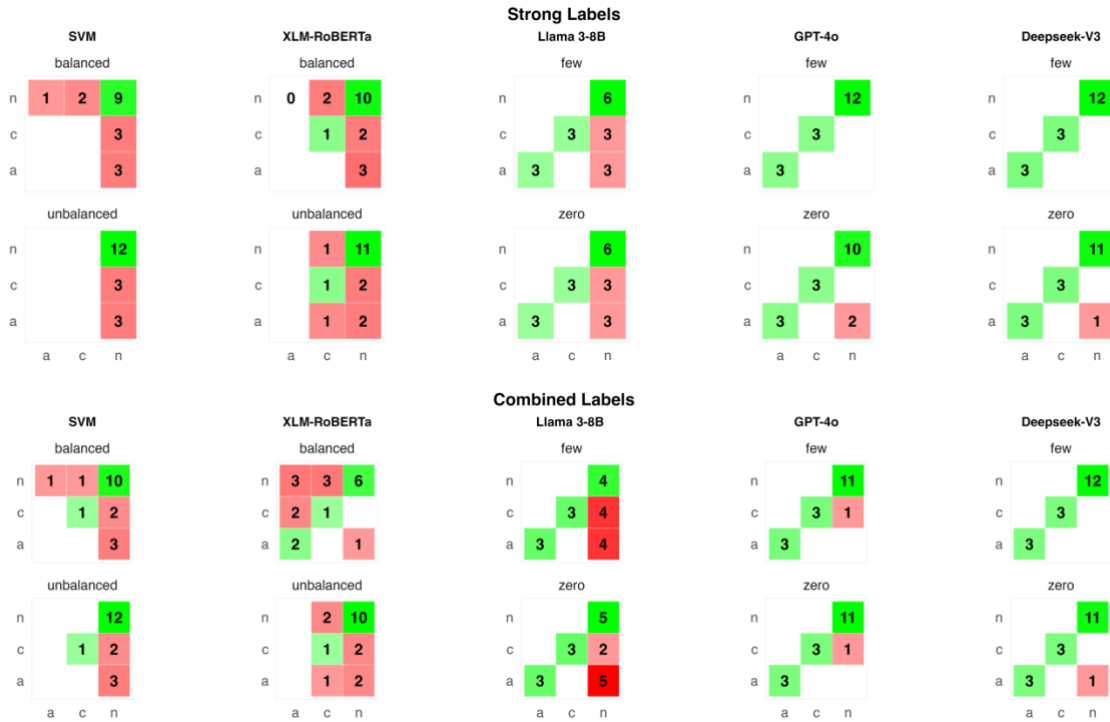
Likewise, models using a semi-supervised (XLM-RoBERTa) or supervised (SVM) strategy exhibited also lower macro-F1 scores. The SVM model performed reasonably well on the balanced dataset with strong labels ($F1 = 0.69$) but performed significantly worse on the *weak+strong* dataset ($F1 = 0.59$) and both unbalanced datasets ($F1 = 0.61$ and 0.55 , respectively). The XLM-RoBERTa model showed only slight improvements over SVM for the balanced dataset, with its best performance recorded for the balanced strong-labeled dataset ($F1 = 0.70$).

4.4.3 Functional Tests (Structural Validation)

In addition to the human-label comparison, we conduct systematic tests on deliberately chosen sentences and the survey items drawn from Lechner et al.’s (2025) original personality inventory to evaluate whether models can reliably predict explicitly defined test cases (units). Specifically, each model was tested with 18 test units: (a) the three self-report items designed to reflect each associated trait (for instance, “I show understanding for the needs of disadvantaged people” for communion or “I am a born leader” for agency); (b) the three opposing items per trait, expected to be classified as “none” (for example, “I am hard, cold-hearted” as the opposite pole for communion or “I prefer to stay in the background” for the opposite pole for agency); and (c) six sentences from our original corpus, which we identified as unrelated to any personality cues and that also should be labeled as “none” (for a full list of test units, see Appendix H).

Our results are displayed in Figure 16.³⁶ A well-performing model's confusion matrix should ideally display three green predictions in the lower-left corner (representing three “agency” items), three green predictions in the center (corresponding to three “communion” items), and twelve green predictions in the upper-right corner (covering the six “none” sentences from the original corpus and the three opposing poles for both “agency” and “communion”, totalling twelve “none” predictions). For strong labels, only few-shot prompting with GPT-4o and DeepSeek-V3 correctly classified all test items, whereas zero-shot prompting for both models was still satisfactory with only minor prediction errors. For combined weak + strong labels, encompassing both weak and strong cues together, only DeepSeek-V3 was able to correctly classify all test cases. For zero-shot prompting, the performance of GPT-4o and DeepSeek-V3 was also still satisfactory, again with only minor prediction errors. In contrast, SVM, XML-RoBERTa, and Llama 3-8B showed more overall wrong

Figure 16. Confusion Matrices for the Test Units (n=18).



predictions for both strong and combined weak + strong labels, often failing to reach a passing rate of 50 conclude that validation evidence for accurate and consistent classification was demonstrated exclusively by GPT-4o and DeepSeek-V3.

³⁶ The y-axis corresponds to the true values, and the x-axis corresponds to the predicted values, with n = none, c = “communion” and a = “agency”.

4.4.4 Exploring Sensitivity to Partisan Differences (External Validation)

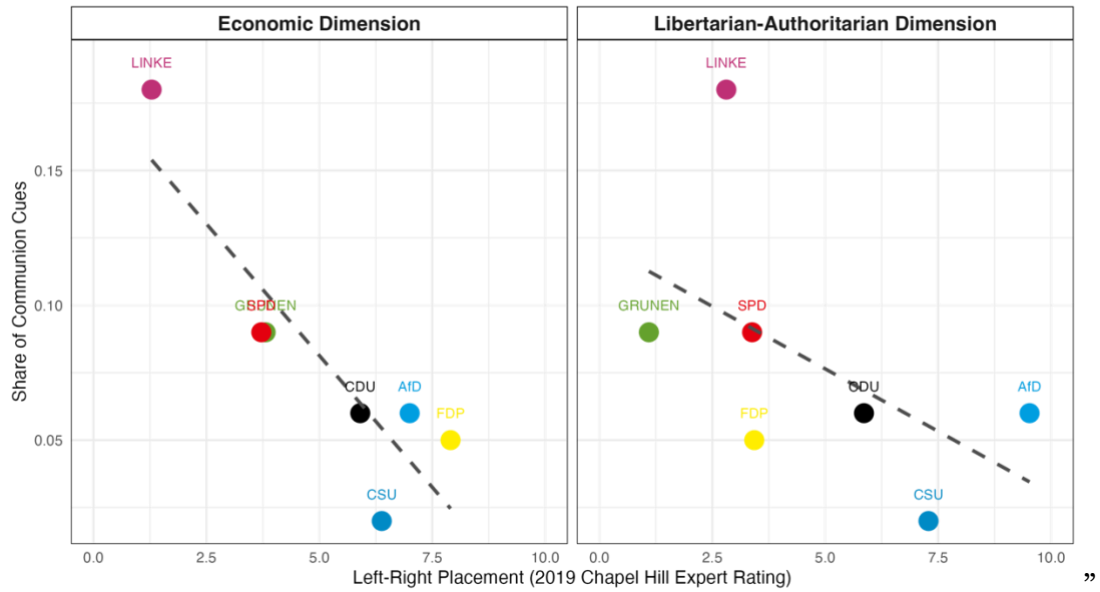
As a further validation step, we evaluate the sensitivity to partisan differences of our measures. Toward that end, we examine partisan differences in public personality across two use cases: (1) aggregated communion scores of politicians across political parties in the German Bundestag, and (2) individual agency and communion scores for the U.S. Presidential Debate between Donald Trump and Kamala Harris on September 10, 2024. In both cases, we expect, in line with previous research, that liberal or left-leaning politicians generally display higher and more universal levels of communion compared to their conservative counterparts, who often tend to limit empathy to deliberately chosen (in-)groups (Hasson et al., 2018; McCue & Gopoian, 2000). In the U.S. context, there is even consistent politician-specific evidence suggesting that Donald Trump is regularly rated lower on empathy-related dimensions reflecting communion and higher on conflict-related dimensions representing agency (Hyatt et al., 2018; Nai et al., 2021; J. D. Wright & Tomlinson, 2018). Therefore, we expect that (1) German politicians from left-leaning parties will be more likely to signal personality cues that indicate communion, and (2) Donald Trump will be more likely to signal personality cues that indicate agency, whereas Kamala Harris will be more likely to signal personality cues that indicate communion.

For clarity, we report results exclusively for the best-performing model that successfully passed all previous validation steps, namely DeepSeek-V3.

4.4.4.1 Left-Right Communion Distribution in the Bundestag

Figure 17 presents scatterplots showing the relationship between politicians' shares of communion with values aggregated at the party level (y-axis) and their parties' expert-rated positions on the Chapel Hill Expert Survey (CHES, see Jolly et al. (2022)) for two dimensions that make up the left-right continuum: the economic left-right (left panel) and the cultural libertarian-authoritarian (right panel) dimension. Generally, the highest shares of sentences containing communion cues can be found for the left party "LINKE" (18%), followed by social-democrat SPD and the Greens (both 9%), whereas politicians from more conservative parties like the CDU and right-wing AfD show overall lower communion values of around 4-5%.

Figure 17: Correspondence of Aggregated Politicians Shares for each Party



As indicated by the regression lines, we see a clear and consistent negative relationship between the parties CHES score and the politicians' share of communion. This association is particularly pronounced for the economic dimension (left panel), but also present in the cultural dimension (right panel). This pattern suggests that the signaling of communion is systematically more prevalent among politicians from economically and culturally left-leaning parties, which supports our prior assumption. These results support the validity of our agency–communion measures because it vividly demonstrates its correspondence with established ideological dimensions.³⁷

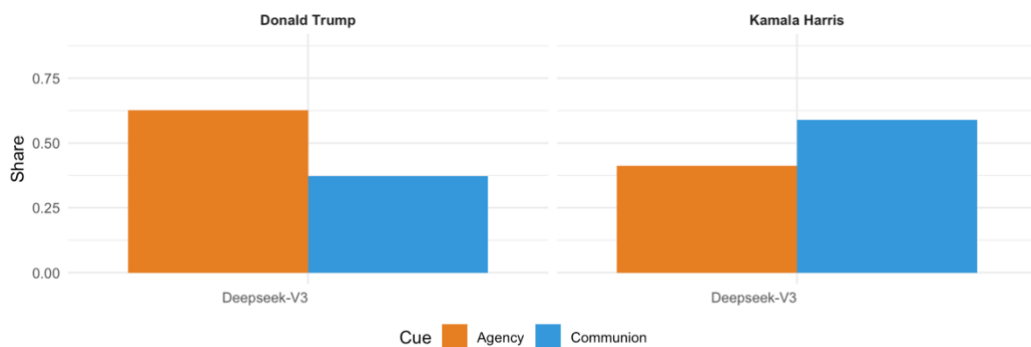
4.4.4.2 US Presidential Debate Corpus

Next, we measure the signalling of personality cues for the 2024 U.S. Presidential Debate between Donald Trump and Kamala Harris on September 10, 2024, to demonstrate that our measure is also sensitive to individual-level measures. Transcripts for the debate were derived from the American Presidency Project (Woolley & Peters, 1999), translated from English to German using the Deepl-API and subsequently processed using the same steps as in our previous analysis.

³⁷ Although not part of our initial analysis due to less apparent theoretical connections in a multi-party parliamentary system, Appendix I presents the aggregated proportions of “agency” and the CHES dimensions, revealing a strong positive relationship with the economic dimension and a moderate positive relationship with the cultural dimension.

Our findings, visualized in Figure 18, confirm our expectations that Trump should signal more agency compared to Kamala Harris. On average, Trump’s sentences are notably shorter, leading to nearly twice as many utterances ($n = 655$) compared to Harris ($n = 320$). Notably, the overall share of personality cues is higher for Harris (approx. 23%) than for Trump (approx. 10%). However, examining the within-person distribution reveals a clear pattern: Trump displays a stronger emphasis on agency (63% agency vs. 37% communion), whereas Harris shows the reverse pattern, with a

Figure 18: Ratio of Agency and Communion



higher share of communion (59% vs. 41% agency). This means that the relative prominence of agency and communion is nearly reversed between the two candidates.

These results show that our agency–communion measure was able to capture differences in the two candidates’ public personalities that were widely discussed in the media during the election campaign (Hyatt et al., 2018).

4.5 Discussion

4.5.1 Summary and Discussion

This study explored the feasibility of using computational text analysis to measure politicians’ publicly displayed personality traits from various textual sources (interviews, speeches and social media posts) using the two metatraits agency and communion as a case in point. Measuring personality traits from verbal behavior is highly challenging - not only because personality traits are inherently fuzzy constructs, but also because their linguistic cues are often ambiguous and context-dependent. To address these challenges, we compared multiple measurement strategies and models, varying dataset compositions (balanced vs. unbalanced) and personality cue conceptualizations (strong cues only vs. a combination of weak and strong cues). We conducted extensive validation using a systematic framework as outlined in

Birkenmaier et al. (2023). Our findings suggest that prompt-based methods, particularly GPT-4o and DeepSeek-V3, allowed the most valid measurement of agency and communion from politicians' public utterances-evaluated through comparisons with human labeling, deliberately designed functional tests, and sensitivity to partisan differences in real-world applications.

Regarding the conceptualization of personality traits, models tasked to detect only strong cues for the traits in question performed better, as expected. This aligns with intuition, as detecting personality cues becomes more challenging when weak cues are included. The fact that personality traits are fuzzy and concepts with ambiguous verbal cues was reflected in the moderate agreement among human coders (between .70 and .78 Krippendorff's Alpha) and the fact that none of the computational models achieved a macro-F1 score above .80 (see Table 12).

Nevertheless, the results suggest that the best available computational methods can indeed capture meaningful personality-related cues in text, particularly for the two examined dimensions agency and communion. These findings provide the foundation for more substantive research in the field of political psychology. For instance, this could include the investigation of changes in politicians' expressed personality across environmental factors that can indicate how political figures adapt their public personas in response to evolving circumstances. Potential explanatory factors include crisis events, such as economic downturns or global conflicts, which may lead politicians to fulfil the demand for dominant political leaders (Petersen & Laustsen, 2020; Sprong et al., 2019). Additionally, individual role transitions-such as taking over the position of party secretary or becoming a lead candidate-could influence the way personality traits are displayed.

Beyond individual-level analysis, there is also a need for more substantive research on the broader cultural and institutional factors that shape politicians' personality displays. The political system designs the arena of political interactions, with different governmental structures, such as presidential and parliamentary systems, shaping how politicians behave (Schürmann & Stier, 2022). For instance, it remains an open question whether presidential leaders exhibit higher dominance and assertiveness than their parliamentary counterparts. Likewise, cultural norms should also be investigated to determine how regional differences impact the perception and display of traits. Finally, future methodological research could investigate the extent to which computationally derived personality cues align with politicians' self-reported traits. This would deepen our understanding of the strengths and limitations of various

measurement approaches and highlight which personality traits are more likely to show consistency across methods.

While this study demonstrates the potential of computational methods for measuring politician’s personality cues, several limitations remain. The analysis focused on only two personality traits – agency and communion - which should be expanded to other relevant traits. Next, keep in mind that personality cues can be expressed in multiple ways – such as paraverbal and nonverbal features like rhetoric, gestures or interactions. Thus, future research should complement our approach by using multi-modal approaches, incorporating video, audio, and behavioral data (Achmann-Denkler et al., 2024). Connected to this, our study operated at the sentence level; extending it to paragraph-level or token-level analysis would provide deeper insights into personality expression in political speech.

In conclusion, our study demonstrates that while measuring personality trait expressions from text remains a daunting task, it is not an impossible one. By combining robust validation strategies with recent advances in language models, we demonstrated that even abstract traits like agency and communion leave detectable traces in political discourse. This opens the door to a new generation of research: one where personality is not just surveyed or inferred, but measured at scale, over time, and in context - from speeches to debates, from parliaments to press conferences. If personality is indeed what politicians project to the public, then computational methods may soon become essential tools in decoding the psychological choreography of power.

4.6 Appendix

4.6.1 Appendix: Overview of politicians in the German Raw Corpus

Table 13: Overview Politicians

#	Full Name	wikidata_id	wikidata_link
1	Annalena Baerbock	Q564328	https://www.wikidata.org/wiki/Q564328
2	Katarina Barley	Q15117357	https://www.wikidata.org/wiki/Q15117357
3	Dietmar Bartsch	Q99450	https://www.wikidata.org/wiki/Q99450

4	Nicola Beer	Q1986069	https://www.wikidata.org/wiki/Q1986069
5	Marco Buschmann	Q1893672	https://www.wikidata.org/wiki/Q1893672
6	Tino Chrupalla	Q40895143	https://www.wikidata.org/wiki/Q40895143
7	Özlem Demirel	Q307920	https://www.wikidata.org/wiki/Q307920
8	Saskia Esken	Q15813082	https://www.wikidata.org/wiki/Q15813082
9	Nancy Faeser	Q1288810	https://www.wikidata.org/wiki/Q1288810
10	Sven Giegold	Q84687	https://www.wikidata.org/wiki/Q84687
11	Robert Habeck	Q109943	https://www.wikidata.org/wiki/Q109943
12	Hubertus Heil	Q71534	https://www.wikidata.org/wiki/Q71534
13	Björn Höcke	Q18176756	https://www.wikidata.org/wiki/Q18176756
14	Katja Kipping	Q76342	https://www.wikidata.org/wiki/Q76342
15	Lars Klingbeil	Q1806261	https://www.wikidata.org/wiki/Q1806261
16	Annegret Kramp-Karrenbauer	Q71359	https://www.wikidata.org/wiki/Q71359
17	Tilman Kuban	Q62079592	https://www.wikidata.org/wiki/Q62079592
18	Günther Lachmann	Q1561612	https://www.wikidata.org/wiki/Q1561612
19	Richarda Lang	Q42841188	https://www.wikidata.org/wiki/Q42841188
20	Armin Laschet	Q78205	https://www.wikidata.org/wiki/Q78205
21	Karl Lauterbach	Q105676	https://www.wikidata.org/wiki/Q105676
22	Steffi Lemke	Q1502854	https://www.wikidata.org/wiki/Q1502854
23	Christian Lindner	Q85806	https://www.wikidata.org/wiki/Q85806
24	Heiko Maas	Q108447	https://www.wikidata.org/wiki/Q108447
25	Friedrich Merz	Q566257	https://www.wikidata.org/wiki/Q566257
26	Jörg Meuthen	Q15907132	https://www.wikidata.org/wiki/Q15907132
27	Omid Nouripour	Q71635	https://www.wikidata.org/wiki/Q71635
28	Cem Özdemir	Q12839	https://www.wikidata.org/wiki/Q12839
29	Boris Pistorius	Q893758	https://www.wikidata.org/wiki/Q893758
30	Bernd Riexinger	Q69346	https://www.wikidata.org/wiki/Q69346

31	Norbert Röttgen	Q61903	https://www.wikidata.org/wiki/Q61903
3 2	Martin Schirdewan	Q17124693	https://www.wikidata.org/wiki/Q17124693
3 3	Olaf Scholz	Q61053	https://www.wikidata.org/wiki/Q61053
3 4	Martin Schulz	Q17905	https://www.wikidata.org/wiki/Q17905
3 5	Horst Seehofer	Q44612	https://www.wikidata.org/wiki/Q44612
3 6	Markus Söder	Q50664	https://www.wikidata.org/wiki/Q50664
3 7	Jens Spahn	Q86294	https://www.wikidata.org/wiki/Q86294
3 8	Frank-Walter Steinmeier	Q76658	https://www.wikidata.org/wiki/Q76658
3 9	Marie-Agnes Strack-Zimmermann	Q15391841	https://www.wikidata.org/wiki/Q15391841
4 0	Peter Tauber	Q124622	https://www.wikidata.org/wiki/Q124622
41	Johannes Vogel	Q105426	https://www.wikidata.org/wiki/Q105426
4 2	Sahra Wagenknecht	Q77193	https://www.wikidata.org/wiki/Q77193
4 3	Alice Weidel	Q27975838	https://www.wikidata.org/wiki/Q27975838
4 4	Volker Wissing	Q1278715	https://www.wikidata.org/wiki/Q1278715
4 5	Janine Wissler	Q1682774	https://www.wikidata.org/wiki/Q1682774
4 6	Paul Ziemiak	Q18169686	https://www.wikidata.org/wiki/Q18169686

4.6.2 Appendix: General Workflow

This procedure outlines a **generalizable, step-by-step workflow** for detecting personality cues in textual data and aggregating them to personality traits. The workflow is trait-agnostic and applicable across research fields.

Step 0: Understand Your Research Context & Data

- Familiarize yourself with the empirical and theoretical landscape of the research context (e.g., research about individual, public personalities, etc.).
- Reflect on the nature of the data source, structure, etc.

Step 1: Trait Selection

- Identify one or more personality traits based on a coherent personality framework (e.g., Big Five, HEXACO, domain-specific inventories). Use a domain-specific trait model when appropriate.
- Specify which pole(s) of the trait you are interested in measuring (e.g., “Assertiveness” vs. “Submissiveness” or only the positive pole).

Step 2: Trait Conceptualization

- Define the link of personality cues to the selected trait(s) in conceptual terms using prior psychological literature.
- Identify what constitutes a strong cue (direct, explicit statements) and a weak cue (indirect, implicit indicators).

Step 3: Codebook Development

- Build a codebook that contains all information for the annotation process for personality cues in text.
- Develop coding categories (e.g., “Trait X – strong,” “Trait X – weak,” “None”) and annotate a small sample
- Refine the codebook iteratively with coder feedback.

Step 4: Codebook Validation

- Annotate a larger, randomly drawn text sample using at least two independent coders.
- Evaluate inter-coder agreement (e.g., Krippendorff’s Alpha, with $\alpha > .70$ as acceptable).
- Revise definitions and examples to improve reliability and conceptual clarity.

Step 5: Human Annotation of Full Dataset

- Annotate the full dataset

4.7.1 *Appendix: Filter Terms*

German

ich | meine | meiner | unsere | wir | persönlich | man | selbst | mich | mein | mir | uns
| unsere

English Translation

I | my / mine | mine / of mine | our / ours | we | personal / personally | one | self /
oneself | me | my | me (dative) | us | our / ours

4.7.2 Appendix: Codebook

This is the codebook for the coding of politician 's personality (PoliPers). The coding was conducted using the software "Label Studio".

Facet	Description	Strong	Weak+Strong	Anmerkungen	Survey Items ("I ...")
Empathy & Compassion (Empathie & Mitgefühl)	A politician who sympathizes with (disadvantaged) people's problems and feelings, who is caring and compassionate in his behavior, shows empathy vs. someone who is callous, unemotional or remorseless.	When referring to concrete (disadvantaged) people or groups "The number of homeless people is too high, this is a humanitarian catastrophe." "We increased child benefits [...] which was a significant financial improvement for many families." "Poor people are worried that others will take advantage of their situation." When compassion or empathy is explicitly called for "It is important to understand people's concerns and hardships in order to find real solutions that leave no one behind." "Well, I wouldn't kill anything with two eyes to feed myself."	When moral judgment is not explicitly empathetic, but fact-based / argumentative "The number of homeless people is high." Reference to global crises (wars, etc.) without directly naming the consequences for people (or lives); "Russia has attacked Ukraine." "We must continue to support Ukraine." Focus on political entities "We must support the municipalities."	No generally positive empathy expressed toward other politicians (e.g., "Karl Lauterbach is a good colleague") No general references to acts of war between countries (e.g., "The conflict between Russia and Ukraine is escalating") When referencing people in vague terms (e.g., "they are disadvantaged", "we need to help them") – If the meaning is implicit and allows for multiple	am empathetic, show genuine compassion. understand the needs of disadvantaged people. am compassionate, have a soft heart.

		Reference to global crises (wars, etc.) with consequences for people / human lives “Russia attacked the people of Ukraine.” “The ISIS reign of terror over 3.2 million Syrians and 4.5 million Iraqis was broken last year.”	„The CDU is suffering under demographic changes“	interpretations, it is considered implicit. If there is a clear and unambiguous interpretation pointing to empathetic behavior (e.g., "And the state still treats them as if something is wrong with them."), then it is considered explicit.	
Assertiveness & Will to take control (Durchsetzungsfähigkeit und Führungsanspruch)	A politician who is assertive, willing to act as a leader, who takes charge, seeks for greater power vs. someone who is more submissive, unobtrusive, even feels awkward in positions of power	Concrete power-related plans in the context of personal assertiveness against resistance or concerns, etc. "That was a very concrete example, and we will continue to apply pressure." "We want to push this through despite resistance." Willingness to take on and actively perform leadership "We want to be elected in order to bring about a change in policy."	General references to actions with a disruptive / active character in contrast to the status quo "And I want us to take the first step, to intervene now, to offer support now." "We want to move the country forward — in education, digitalization, and the economy — and do something for the middle class, for example:	No description of role-specific tasks or actions that are not explicitly framed as responsibilities (e.g., “We will introduce a new regulation for the XX law” → unclear if this is not just part of the standard role description) No general statements about desirable conditions	am power-hungry, claim leadership. am assertive, dominant. am a born leader.

		"I am fighting to become party leader." Reference to one's own perceived leadership role "In decisive moments, I take responsibility and resolutely stand up for what is right."	completely abolish the solidarity surcharge, establish a digital ministry, implement an immigration code with a points system." General references to formal or actual leadership or power positions "I very much enjoy being party leader." "We are a party that can change politics." "We are using the resources of the state."	or ideological perspectives "Capitalism is the root of all problems, but I would say that patriarchy and racism are equally systems of oppression that we must fight equally." "We must fight for human rights." Reference to engagement without a dominant or power-related character "We advocated for the regulation." "In my greeting speech, I pointed out the revised noise ordinance for sports facilities, which I supported."	
--	--	---	--	--	--

4.7.3 *Appendix: Hyperparameters*

Study 1 (Baseline SVM): We use default hyperparameters from the caret packet in R.

Study 2 (Fine-Tuning XLM-RoBERTa): We use the "roberta-base" model from the transformer library, with 12-layer, 768-hidden, 12-heads, 279M parameters. We use common hyperparameters: batch size of 16, learning rate of $2e-5$, and 5 epochs. Unless otherwise noted, we utilize standard hyperparameters

Study 3 (Prompting Instruction-Tuned LLM). We use default hyperparameters, with temperature of .7, max-tokens of 150,

4.7.4 Appendix: Computational Resources

Study 1 (Baseline): Training the SVM was implemented in R using the *caret* package (Kuhn, 2008). The computations were carried out on a standard CPU. Cross-validation and feature selection were performed on moderate-sized DFMs, and the models (trained over three different seeds) required on average 30 seconds (cross-validation training and evaluation) and 2 min (unlabeled data) of processing time.

Study 2 (Fine-Tuning XLM-RoBERTa): For model training, we used a A100 GPU from Google Colab running Python 3.10.12 using Jupyter Notebooks. Fine-tuning XLM-RoBERTa-base and classifying the datasets took approximately 20 min (cross-validation training and evaluation) and 2 hours (unlabeled data).

Study 3 (Prompting Instruction-Tuned LLM). Models were accessed via the openai python package. GPT-4o was directly accessed via the OpenAI-API, whereas the other models were accessed via the DeepInfra API. API calls were done concurrently using the *concurrent.futures* python module. Classification took approximately 20 min (cross-validation training and evaluation) and 3 hours (unlabeled data). Costs for the labeling were around 500 EUR for GPT-4o, 80 EUR for DeepSeek-V3, and 5 EUR for Llama.

4.7.5 Appendix: Prompts

Prompt 1 — Zero-Shot (Only Strong)

Task Instruction:

Du hast die Aufgabe, Sätze von Politikern anhand von konzeptuellen Kategorien zu klassifizieren:

1. **Empathie & Mitgefühl:**

Eine Aussage zeigt, dass ein Politiker empathisch ist, mit den Problemen und Gefühlen (benachteiligter) Menschen mitfühlt, Fürsorge und Mitgefühl zeigt sowie die Bedürfnisse von benachteiligten Gruppen versteht.

Beispiele:

- Wenn bezogen auf konkrete (benachteiligte) Menschen(gruppen) oder Personen
- Wenn konkret Mitgefühl oder Empathie gefordert wird
- Wenn Verweis auf globale Krisen (Kriege, etc.) mit Folgen für Menschen(-leben)

2. **Durchsetzungsfähigkeit & Führungsanspruch:**

Eine Aussage zeigt, dass ein Politiker entschlossen ist, die Führung zu übernehmen, Verantwortung zu tragen und gegen Widerstände aktiv zu handeln.

Beispiele:

- Wenn konkrete machtbezogene Planungen, die im Kontext von persönlicher Durchsetzungsfähigkeit gegen Widerstände oder Bedenken / etc. stehen
- Wille zur Übernahme und aktive Ausführung von Führung
- Verweis auf eigene wahrgenommene Führungsrolle

3. **None:**

Die Aussage macht keine Rückschlüsse über eine Verhaltensweise des Politikers, die auf die ersten beiden Kategorien „Empathie und Mitgefühl“ oder „Durchsetzungsfähigkeit & Führungsanspruch“ zurückzuführen ist.

Aufgabe: Klassifiziere die folgende Aussage und gib eine Erklärung ab.
Gib deine Antwort im folgenden Format. Wähle eine der drei Kategorien aus:

Ergebnis: [emp, dur, none]

Erläuterung: [Erläuterung (max. 1 Satz)]

Prompt 2 — Zero-Shot (Strong + Weak)

Task Instruction:

Du hast die Aufgabe, Sätze von Politikern anhand von konzeptuellen Kategorien zu klassifizieren:

1. **Empathie & Mitgefühl:**

Eine Aussage zeigt, dass ein Politiker empathisch ist, mit den Problemen und Gefühlen (benachteiligter) Menschen mitfühlt, Fürsorge und Mitgefühl zeigt sowie die Bedürfnisse von benachteiligten Gruppen versteht.

Beispiele:

- Wenn bezogen auf konkrete (benachteiligte) Menschen(gruppen) oder Personen
- Wenn konkret Mitgefühl oder Empathie gefordert wird
- Wenn Verweis auf globale Krisen (Kriege, etc.) mit Folgen für Menschen(-leben)
- Wenn moralische Bewertung nicht explizit empathisch, sondern faktenbasiert / argumentierend
- Verweis auf globale Krisen (Kriege, etc.) ohne direkte Benennung der Folgen für Menschen(-leben)
- Empathischer Fokus auf politischen Einheiten

2. **Durchsetzungsfähigkeit & Führungsanspruch:**

Eine Aussage zeigt, dass ein Politiker entschlossen ist, die Führung zu übernehmen, Verantwortung zu tragen und gegen Widerstände aktiv zu handeln.

Beispiele:

- Wenn konkrete machtbezogene Planungen, die im Kontext von persönlicher Durchsetzungsfähigkeit gegen Widerstände oder Bedenken / etc. stehen
- Wille zur Übernahme und aktive Ausführung von Führung
- Verweis auf eigene wahrgenommene Führungsrolle
- Generelle Verweise auf Handlungen, die einen disruptiven / aktiven Charakter gegenüber des Status Quo haben
- Generelle Verweise auf formale oder faktische Führungs- oder Machtposition

3. **None:**

Die Aussage macht keine Rückschlüsse über eine Verhaltensweise des Politikers, die auf die ersten beiden Kategorien „Empathie und Mitgefühl“ oder „Durchsetzungsfähigkeit & Führungsanspruch“ zurückzuführen ist.

Aufgabe: Klassifiziere die folgende Aussage und gib eine Erklärung ab.
Gib deine Antwort im folgenden Format:

Ergebnis: [emp, dur, none]

Erläuterung: [Erläuterung (max. 1 Satz)]

Prompt 3 — Few-Shot (Only Strong)

Task Instruction:

Du hast die Aufgabe, Sätze von Politikern anhand von konzeptuellen Kategorien zu klassifizieren:

1. **Empathie & Mitgefühl:**

Eine Aussage zeigt, dass ein Politiker empathisch ist, mit den Problemen und

Gefühlen (benachteiligter) Menschen mitfühlt, Fürsorge und Mitgefühl zeigt sowie die Bedürfnisse von benachteiligten Gruppen versteht.

Beispiele:

- Wenn bezogen auf konkrete (benachteiligte) Menschen(gruppen) oder Personen
- Wenn konkret Mitgefühl oder Empathie gefordert wird
- Wenn Verweis auf globale Krisen (Kriege, etc.) mit Folgen für Menschen(-leben)

2. Durchsetzungsfähigkeit & Führungsanspruch:

Eine Aussage zeigt, dass ein Politiker entschlossen ist, die Führung zu übernehmen, Verantwortung zu tragen und gegen Widerstände aktiv zu handeln.

Beispiele:

- Wenn konkrete machtbezogene Planungen, die im Kontext von persönlicher Durchsetzungsfähigkeit gegen Widerstände oder Bedenken / etc. stehen
- Wille zur Übernahme und aktive Ausführung von Führung
- Verweis auf eigene wahrgenommene Führungsrolle

3. None:

Die Aussage macht keine Rückschlüsse über eine Verhaltensweise des Politikers, die auf die ersten beiden Kategorien zurückzuführen ist.

Aufgabe: Klassifiziere die folgende Aussage und gib eine Erklärung ab.
Gib deine Antwort im folgenden Format:

Ergebnis: [emp, dur, none]

Erläuterung: [Erläuterung (max. 1 Satz)]

Beispiele:

„Die Anzahl der Wohnungslosen ist zu hoch, das ist eine humanitäre Katastrophe“

→ Ergebnis: [emp]

→ Erläuterung: [Aussage bezieht sich auf konkrete (benachteiligte) Menschen(gruppen) oder Personen]

„Wir wollen das auch gegen Widerstände durchsetzen“

→ Ergebnis: [dur]

→ Erläuterung: [Aussage bezieht sich auf konkrete machtbezogene Planungen im Kontext von Durchsetzungsfähigkeit]

Prompt 4 — Few-Shot (Strong + Weak)

Task Instruction:

Du hast die Aufgabe, Sätze von Politikern anhand von konzeptuellen Kategorien zu klassifizieren:

1. **Empathie & Mitgefühl:**

Eine Aussage zeigt, dass ein Politiker empathisch ist, mit den Problemen und Gefühlen (benachteiligter) Menschen mitfühlt, Fürsorge und Mitgefühl zeigt sowie die Bedürfnisse von benachteiligten Gruppen versteht.

Beispiele:

- Wenn bezogen auf konkrete (benachteiligte) Menschen(gruppen) oder Personen
- Wenn konkret Mitgefühl oder Empathie gefordert wird
- Wenn Verweis auf globale Krisen (Kriege, etc.) mit Folgen für Menschen(-leben)
- Wenn moralische Bewertung nicht explizit empathisch, sondern faktenbasiert / argumentierend
- Verweis auf globale Krisen ohne direkte Benennung der Folgen für Menschen(-leben)
- Empathischer Fokus auf politischen Einheiten

2. **Durchsetzungsfähigkeit & Führungsanspruch:**

Eine Aussage zeigt, dass ein Politiker entschlossen ist, die Führung zu übernehmen, Verantwortung zu tragen und gegen Widerstände aktiv zu handeln.

Beispiele:

- Wenn konkrete machtbezogene Planungen, die im Kontext von persönlicher Durchsetzungsfähigkeit gegen Widerstände stehen
- Wille zur Übernahme und aktive Ausführung von Führung
- Verweis auf eigene wahrgenommene Führungsrolle
- Generelle Verweise auf Handlungen, die einen disruptiven / aktiven Charakter gegenüber des Status Quo haben
- Generelle Verweise auf formale oder faktische Führungs- oder Machtposition

3. **None:**

Die Aussage macht keine Rückschlüsse über eine Verhaltensweise des Politikers, die auf die ersten beiden Kategorien zurückzuführen ist.

Aufgabe: Klassifiziere die folgende Aussage und gib eine Erklärung ab.
Gib deine Antwort im folgenden Format:

Ergebnis: [emp, dur, none]

Erläuterung: [Erläuterung (max. 1 Satz)]

Beispiele:

„Anzahl der Wohnungslosen ist hoch“ → Ergebnis: [emp]

„Wir wollen das auch gegen Widerstände durchsetzen“ → Ergebnis: [dur]

„Und ich möchte, dass wir den ersten Schritt gehen, dass wir jetzt eingreifen, dass wir jetzt unterstützen.“ → Ergebnis: [dur]

4.7.6 *Appendix: Validation Strategy*

Due to the foundational nature of our research, we take a systematic approach to validation and rely on the ValiText framework (Birkenmaier et al., 2024), which distinguishes between different types of validation evidence for computational text analysis. We start by reporting and documenting the theoretical framework used to operationalize our constructs, provide interrater agreement measures for human coders, and document major design decisions, such as the rationale behind our choice of methods (substantive validation). We then evaluate how our measures correspond to human annotations (external validation) and further explore the properties and the output (structural validation). Specifically, this involves designing functional tests and analyzing the sensitivity for partisan differences across (a) the complete unlabeled corpus of 45,000 randomly selected sentences from the German Bundestag and (b) around 2,000 translated statements from the 2024 U.S. presidential debate between Donald Trump and Kamala Harris. We document all our validation steps using checklist templates (see Appendix G).

4.7.7 Appendix I: Filled-Out Checklist

Substantive Evidence

Before conducting any measurements, researchers need to outline the theoretical underpinning of the measurement to demonstrate substantive evidence. Validation steps for **substantive evidence** should therefore demonstrate that the measurement is based on a strong conceptual foundation, including the operationalization of the construct and the design decisions around the measurement process.

ID	Validation Step	Documentation	Considerations	Validation Output	Source / References
Construct Definition and Operationalization					
I.1	Documentation of the conceptual background	Review of previous approaches to measure politicians' personality, both for politicians' personality in general, and computational text analysis (theory section)	Conducting a literature review to understand the conceptual background of the construct; consulting domain experts to validate and refine the understanding.	Discussion of existing literature / knowledge on the conceptual background of the construct	Krippendorff (2018)
I.2	Justification of the operationalization	Outlining the theoretical relationship between politician's public and intrinsic personality; outlining the importance of personality cues as measurable indicators of personality (theory section)	Explaining how the construct should manifest in the textual data; documenting the operationalization of the textual features for the construct, e.g., in an annotation codebook	Definition of the construct; Rationale of how the construct should express itself in textual features	Krippendorff (2018)
I.3	Manual Precoding	We reached sufficient interrater agreement above .70 Krippendorff's Alpha (Table 1)	Ensuring the construct and its operationalization are clearly defined and suitable for coding; testing and refining the codebook using a random subsample of the data and multiple coders.	Sufficient interrater agreement (e.g., Krippendorff's alpha α , with values above .70 to .80 deemed as minimum, depending on the context)	Krippendorff (2018), Plank (2022)
Design Decisions					
I.4	Justification of data collection decisions	We selected three different data sources (interviews, speeches, and social media posts) as they have been discussed in the literature	Outlining the rationale for the data collection, including the representativeness and relevance to the research question and target	Discussion of relevant data collection decisions	Krippendorff (2018)

			population; explaining data selection decisions (e.g., using keywords); assessing the dataset for quality, completeness, and potential biases or inconsistencies.		
I.5	Justification of method choice	We select three different measurement approaches and rely on SOTA models for each strategy	Outlining the rationale for method selection based on the construct's operationalization and data characteristics; justifying the choice of a specific model	Discussion of method selection decisions	Grimmer et al. (2022)
I.6	Justification of the level of analysis	"Across all data sources, we defined individual sentences and statements uttered or written by politicians as our unit of analysis. Sentences represent personality cues at a granular level that can be detected and potentially aggregated across texts. Because our primary focus is on validating the detection and measurement of personality cues in text rather than analyzing explanatory factors influencing changes or differences in the public personality, we draw a random sample of 15,000 sentences from each data source as our main corpus."	Outlining the rationale for selecting the appropriate level of analysis (e.g., corpus, document, paragraph, sentence, word, or token level); considering potential issues when aggregating measures from lower to higher levels (e.g., sentence to paragraph).	Discussion of selecting level of analysis	Jankowski & Huber (2022)
I.7	Justification of preprocessing decisions	We did not apply particular preprocessing decisions beyond standard practices	Outlining the rationale for any text preprocessing decisions, such as removing specific words or phrases prior to analysis.	Discussion of preprocessing decisions	Grimmer et al. (2022)

Structural Evidence

For **structural evidence**, researchers should conduct validation steps to examine and evaluate the properties of the model and its output. Structural evidence enables the researcher to gain a deeper understanding of how the measurement process functions, as well as to identify any biases or errors introduced by the computational workflow.

ID	Validation Step	Documentation	Considerations	Validation Output	Source / References
Model Feature Inspection					
I.1	Inspection of predictive model features		Ensuring that top-ranked predictive features are conceptually aligned with the intended construct (e.g., a classifier trained to detect	Qualitative evaluation of top-ranked model features using feature-	Molnar (2020), Küpfer & Meyer (2023)

			ideology should be most influenced by “ideological” words and phrases)	importance methods like e.g., LIME or ICE	
Descriptive Output Inspection					
II.2	Visual inspection of output		Describing the distribution of model outputs; detecting anomalies by visualizing descriptive statistics and identifying outliers	Visual or descriptive display of data distribution and summary statistics	Goet (2019)
II.3	Comparison of aggregated measures across known groups	We compare aggregated measures of agency and communion across - Gender of politicians - Data Sources	Validating group-level differences in outputs by comparing aggregated scores (e.g., mean, median) across relevant subgroups	Group-wise aggregation (e.g., by age, gender, platform); visualizations (e.g., bar plots, error bars); comparison with expected trends from literature	Goet (2019)
II.4	Qualitatively assess top text units (document, sentence, etc.) with the highest overall scores for each output category		Ensuring that text units with the highest or lowest predicted scores reflect the underlying construct accurately	Display and Interpretation of text units with the highest (confidence) scores, e.g., in a table	Goet (2019)
Error Analysis					
II.5	Error analysis using data grouping		Checking for bias and subgroup-specific performance disparities by comparing metrics across demographic or contextual groups	Calculation and comparison of performance metrics (e.g., F1, precision, recall) by subgroup; visualizations and discussion of disparities	Wu et al. (2019)
Systematic Testing (context-specific)					
V.1	Systematic Tests	We apply systematic tests using test sentences & display them in a confusion matrix (see Result section)	Evaluating the model's sensitivity, robustness, discriminative validity, and generalizability by testing it on	Performance metrics (e.g., F1, precision, recall) and/or output	(Garg et al., 2019)

			systematically altered or novel input data (e.g., counterfactual examples, adversarial perturbations, discriminant contrasts, out-of-domain texts)	score distributions for curated test sets; qualitative review of selected predictions to assess expected behavioral changes; identification of systematic weaknesses or strengths	
--	--	--	--	---	--

External Evidence

For external evidence, researchers should conduct validation steps that test for how the measures correspond to independent information or criteria. Thus, information outside the scope of the textual data in which the measure was constructed serves as an external benchmark (hence “external” evidence).

ID	Validation Step	Documentation	Considerations	Validation Output	Source / References
Construct Definition and Operationalization					
III.1	Comparison of measures with human-annotated test set ("gold-standard data")	We conduct human-annotation and report performance metrics of all models on the test set (see results section)	Assessing whether the model achieves sufficient predictive performance on a held-out dataset with human annotations, using robust evaluation practices (e.g., cross-validation, multiple random seeds, error bars)	Reporting averaged performance metrics (e.g., macro-F) across random seeds; visualization of performance estimates and error bars; interpretation of results	(Samory et al., 2021)
Surrogate Label Comparison (context-specific)					
V.5	Comparison of measures with surrogate labels		Evaluating whether model predictions align with alternative or proxy labels	Reporting averaged performance metrics (e.g., macro-F) across random seeds;	Grimmer et al. (2022)

				visualization of performance estimates and error bars; interpretation of results	
Criterion Prediction (context-specific)					
V.6	Criterion Prediction	We conduct sensitivity analysis for two cases - Unlabeled German Corpus - US presidential debates	Validating whether model outputs are significantly and meaningfully associated with relevant external criteria or real-world outcomes (e.g., engagement metrics, voting behavior, survey responses)	Reporting effect sizes (e.g., regression coefficients, R ²) and significance tests for associations between model predictions and outcome variables; interpretation of results	Grimmer et al. (2022)

Robustness Checks

Next to the three types of validation evidence outlined above, the ValiText framework also recommends the continuous test of robustness checks to assess the impact of researchers' degree of freedom on the measurement outcome. On a general note, one could see robustness checks as additional means to test whether decisions regarding the measurement design might have a sustainable effect on the measure's outcome.

ID	Validation Step	Documentation	Considerations	Performance Criteria	Source / References
Construct Definition and Operationalization					
IV.1	Rerunning the analysis using alternative text models		Rerunning the analysis with alternative text-based methods, such as a baseline model	Observing significant changes in the validation output, e.g., macro-F1 scores (see III.1)	(Samory et al., 2021)
IV.2	Rerunning the analysis using different		Rerunning the analysis with alternative hyperparameter settings	Observing significant changes in the validation output,	(Arnold et al., 2023)

	hyperparameter settings			e.g., macro-F1 scores (see III.1)	
IV.3	Rerunning the analysis using different cutoff thresholds		Rerunning the analysis with alternative cutoff thresholds	Observing significant changes in the validation output, e.g., macro-F1 scores (see III.1)	Grimmer et al. (2022)
IV.4	Rerunning the analysis using different text cleaning and preprocessing steps		Rerunning the analysis with alternative text preprocessing steps	Observing significant changes in the validation output, e.g., macro-F1 scores (see III.1)	Grimmer et al. (2022)
C.2	Rerunning the analysis using different prompts	We apply two different prompt designs and display all results next to each other (zero-shot and few-shot prompting, see Appendix F)	Rerunning the analysis with alternative prompt designs	Observing significant changes in the validation output, e.g., macro-F1 scores (see III.1)	(Gilardi et al., 2023)

4.8.1 Appendix: Test Units

text	expected_label
Ich bin einfühlsam, zeige echtes Mitgefühl.	emp
Ich zeige Verständnis für Bedürfnisse benachteiligter Menschen.	emp
Ich bin mitfühlend, warmherzig.	emp
Ich bin eher hart, kaltherzig.	None (opposite pole of Empathy & Compassion)
Ich zeige wenig Verständnis und Mitgefühl für Schwächere.	None (opposite pole of Empathy & Compassion)
Ich bin gegenüber Benachteiligten herablassend, abwertend.	None (opposite pole of Empathy & Compassion)
Ich bin machthungrig, erhebe den Führungsanspruch.	dur
Ich bin durchsetzungsstark, dominant.	dur
Ich bin eine geborene Führungsperson.	dur
Ich bin zögerlich, halte mich im Hintergrund.	None (opposite pole of Assertiveness & Will to Take control)
Ich bin in Diskussionen unaufdringlich, zurückhaltend.	None (opposite pole of Assertiveness & Will to Take control)
Ich bleibe lieber in der zweiten Reihe.	None (opposite pole of Assertiveness & Will to Take control)
An Kolleginnen nervt mich nichts.	none
Ende April schloss er die Nutzung der nagelneuen Gasleitung Nord Stream 2 endgültig aus.	none
Aber es ist halt so, wir leben im Risiko und wir können uns schützen.	none
Und dann fange ich an mit den weiteren Nutztierarten Schritt für Schritt.	none
Und das war noch nicht mal so, Blunt.	none
Schaut man sich bei unseren europäischen Nachbarn um, sieht die Situation dort ganz anders aus.	none

4.9.1 Appendix: Personality and Ideology

A potential concern in our study is that, instead of measuring personality, our approach might inadvertently capture ideological positions. To address this, we compared our codebook for the personality trait “communion” with the codebook for the right–left position (RILE) score of the Manifesto Project (Lehmann et al., 2024). The rile score is an aggregate measure consisting of different categories outlined in Figure 9. We included this comparison to directly assess the extent of conceptual and textual overlap between personality- and ideology-based measures. This step allows us to demonstrate that, while certain ideological categories contain language relevant to communion, our codebook also captures personality-specific cues that are clearly independent of ideological positioning.

Figure 19: Screenshot of the Respective Codebook Section for the Rile-Score

rile	Right-left position of party as given in Michael Laver/Ian Budge (eds.): Party Policy and Government Coalitions, Houndmills, Basingstoke, Hampshire: The MacMillan Press 1992: (per104 + per201 + per203 + per305 + per401 + per402 + per407 + per414 + per505 + per601 + per603 + per605 + per606) - (per103 + per105 + per106 + per107 + per403 + per404 + per406 + per412 + per413 + per504 + per506 + per701 + per202).
------	---

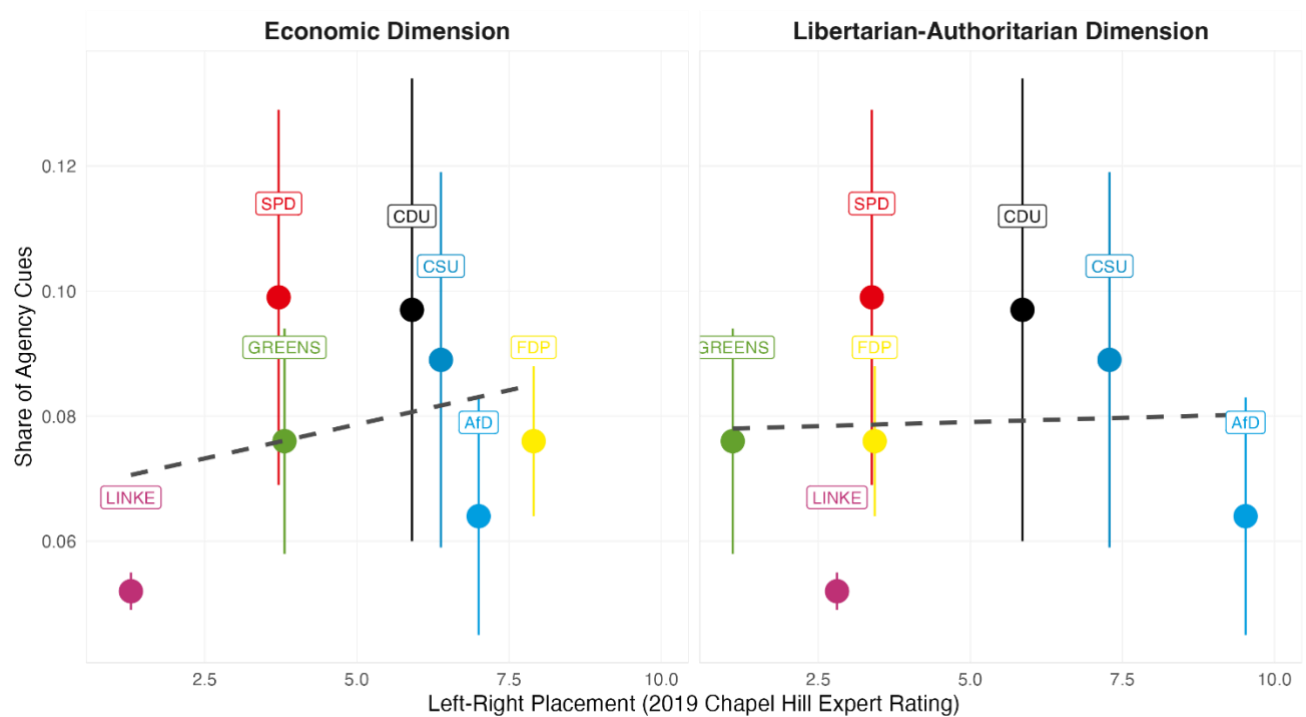
We acknowledge that some categories clearly target communion-related perspectives, such as **Per105 Military: Negative**, which includes negative references to the military or to the use of military power to solve conflicts, as well as references to the “evils of war.” Another example is **Per106 Peace**, which encompasses any declaration of belief in peace and peaceful means of solving crises, without reference to the military.

However, other categories such as **Per404 Economic Planning**, referring to favourable mentions of long-standing economic planning by the government, and **Per406 Protectionism: Positive**, which includes favourable mentions of extending or maintaining the protection of internal markets (either by the manifesto country or other countries), cannot be meaningfully linked to communion-related concepts and clearly target their distinct ideological perspective.

This argument is further strengthened as some themes associated with right-wing positions could also be linked to communion — even though, theoretically, these should oppose each other if we were measuring the same construct (e.g., our personality-based measures vs. ideology). Examples include **Per201 Freedom and Human Rights**, which contains favourable mentions of the importance of personal freedom and civil rights, and **Per606 Civic Mindedness: Positive**, which refers to appeals for national solidarity and the need for society to see itself as united, as well as calls for solidarity with and help for fellow people, both familiar and unfamiliar.

4.9.2 Appendix: Sensitivity Analysis for Agency and CHES Dimensions

Figure 20: Correspondence of Aggregated Agency Politicians' Shares

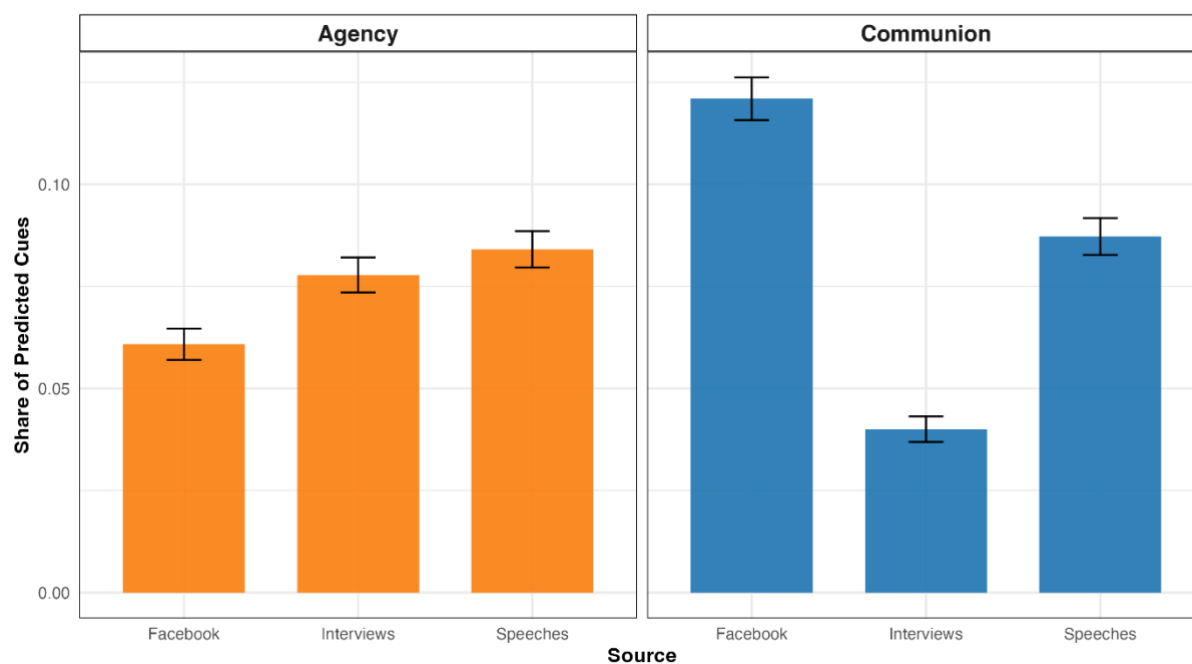


Error Bars indicate Standard Errors around each Party Mean

4.9.3 Appendix: Distribution of Agency and Communion across Data

While agency scores are relatively consistent, speeches exhibit slightly higher levels. In contrast, communion is most pronounced in Facebook posts, with notably lower levels in interviews.

Figure 21: Mean Levels of Agency and Communion across Communication Formats

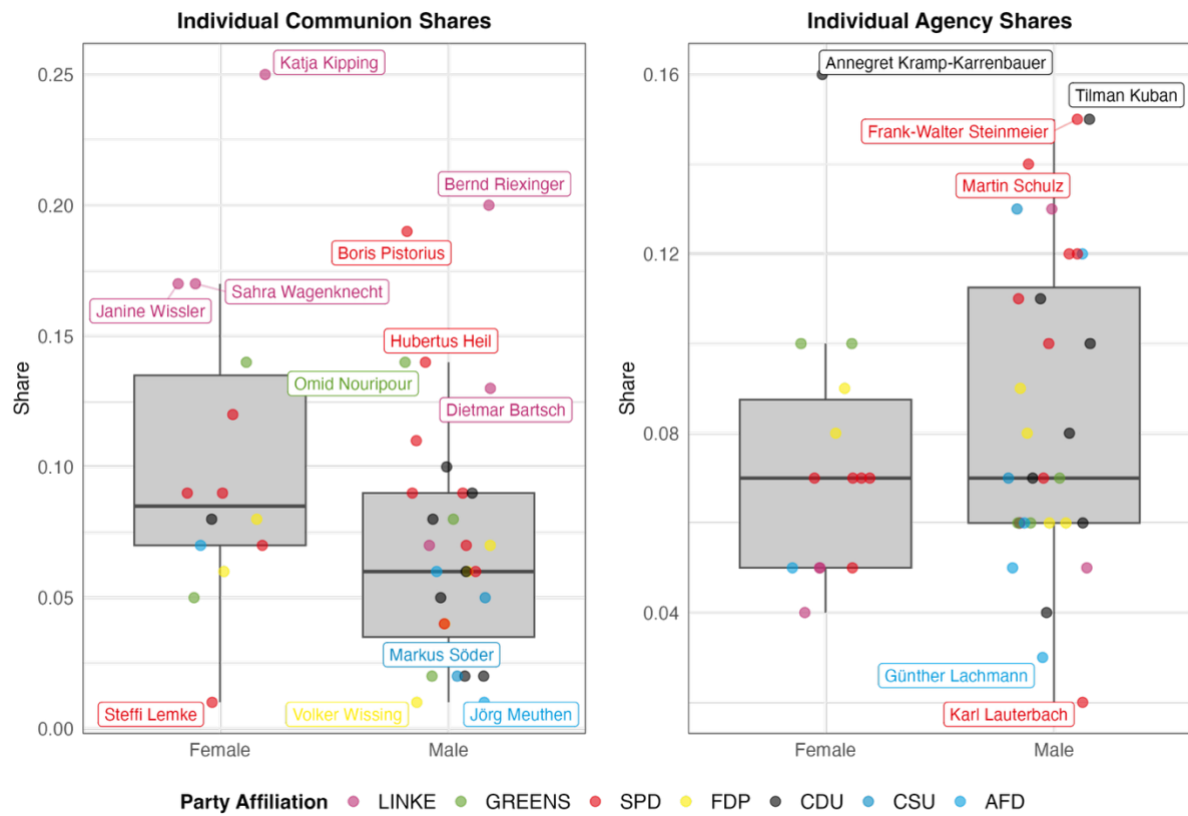


Error bars represent 95% confidence intervals generated via binomial distribution

4.9.4 Appendix: Personality Scores and Gender Across Parties

Figure 9 displays the Boxplot of the Distribution of agency and communion scores by politician and gender for our German Sample. Female politicians (left) generally show higher levels of communion compared to their male counterparts (right). Agency scores vary more widely within each gender group. Boxplots indicate the interquartile range and median, while individual dots represent politician-level means, colored by party affiliation.

Figure 22: Distribution of Scores by Gender



5. Framework Application 2: Geographic Representation

This chapter introduces the second application chapter of the framework: a study on the validation of measures of politicians' geographic representation. Rather than following a conventional construct classification pipeline, the study focuses on validating location references directly, using techniques such as named entity recognition and data linkage.

This chapter is based on the following publication: Birkenmaier, L., Stroppe, A.-K., Wurthmann, C. L., & Sältzer, M. (2025). Between Home Turf and Hinterland: Directly Elected MPs Focus More on Local and Deprived Places than List Candidates on Social Media. *Working Paper. Currently under review in Research & Politics*

5.1 Introduction

Spatial divides, such as urban–rural and center–periphery cleavages, remain defining features of democratic politics. Urban centers concentrate economic activity, political institutions, and cultural capital, whereas peripheral and rural areas often perceive themselves as neglected or “left behind.” Electoral institutions have historically evolved to balance these inequalities by ensuring that geographically distinct communities gain political voice and representation (Bäck & Debus, 2018; Cramer, 2016; Rodríguez-Pose, 2018; Smrek, 2023). In most democracies, representation is organized through territorially defined constituencies that tie elected officials to specific places and create incentives to demonstrate local responsiveness (Norris, 2004).

However, it remains unclear how such institutional arrangements influence equal territorial representation. Mixed-member electoral systems (MMES) offer a valuable setting to examine this question, as they combine single-member districts with proportional party lists, generating two distinct representational logics. While directly elected members of parliament (MP) depend on geographically bounded constituencies, list MPs derive their mandates from parties and are accountable to broader regional or national electorates (Haffert, 2021; Patzelt, 2013).

These institutional incentives shape politicians' communication and which places they choose to highlight. Constituency salience, i.e., the explicit mention of places, serves as a key indicator of local responsiveness (Schürmann & Stier, 2022; Zittel et al., 2019). Naming a place signals attention and commitment, particularly in decentralized and personalized environments such as social media, where politicians decide whether to emphasize their home district, regional ties, neglected areas, or nationally salient locations.

Recent research has broadened the focus to more diffuse and strategic forms of territorial representation (Harmening et al., 2025; Nyhuis et al., 2025; Pratas et al., 2025). Nyhuis et al. (2025) introduce an automated cross-national approach using Wikipedia and Geonames to detect geographic references in parliamentary questions, while Pratas et al. (2025) show that MPs concentrate attention on large cities and small rural localities, neglecting medium-sized and suburban areas. At the local level, Harmening et al. (2025) demonstrate that geographic references align closely with electoral strongholds, especially under district-based systems. Despite recent advances, many approaches still rely on manually curated location lists or proprietary tools like the Google Places API that limit scalability and reproducibility. Building on Schürmann and Stier (2023), we apply named entity recognition (NER) to extract place references from MPs' posts and link them to Wikidata via the Wikidata Query Service, enabling matching places with all corresponding characteristics in the Wikidata database.

We furthermore extend previous approaches of identifying places by developing continuous measures of localness, i.e., the distance of referenced locations from an MP's constituency, and deprivation, i.e., the distance from essential public services. These indicators capture two modes of political responsiveness: proximity to constituencies for local engagement and attention to "left-behind" areas to signal inclusion.

Analyzing over 50,000 tweets and 190,000 Facebook posts by all members of the 19th Bundestag (2017–2021), we find a clear mandate effect: directly elected MPs more often reference places within or near their constituencies, while list MPs mention locations 20–40 kilometers farther away on average. Directly elected MPs also refer more frequently to deprived areas, suggesting an additional effort to acknowledge underserved communities.

Overall, the findings underscore how electoral incentives and representational responsibility shape communication and how social media serves as a key arena for signaling local rootedness, or its absence.

5.2 *Spatial Representation*

Representation is a cornerstone of democracy, linking citizens to political institutions and ensuring accountability, legitimacy, and political support (Arnesen and Peters 2018; Norris 2004). Territorial representation, in particular, anchors citizens' interests in geographically defined constituencies rather than centralized structures. This territorial link brings regionally specific concerns, such as economic disparities, infrastructure, and cultural particularities, to the political agenda, connecting spatial inequality to questions of voice and responsiveness.

Electoral institutions enhance accountability by creating a direct and traceable link between representatives and the communities they serve (Cramer 2016). Representatives must justify their actions to a defined electorate, fostering responsiveness to local concerns and strengthening legitimacy (Norris 2004). Spatial representation thus institutionalizes accountability both vertically, between citizens and representatives, and horizontally, by integrating territorially distinct interests into the political system (Arnesen & Peters, 2018).

5.2.1 *Electoral Institutions and the Logic of Representation*

Accountability to voters is achieved through electoral connection, i.e., the system of incentives linking representatives to their constituents (Norris 2004). These incentives motivate politicians to address local concerns and to engage in symbolic representation, for instance by adopting regional references or invoking local traditions (Bäck and Debus 2018; Schürmann and Stier 2023; Nyhuis et al. 2025). The strength and nature of this connection depend on the institutional design that defines how representatives are elected and to whom they are accountable.

Electoral law reflects normative ideas about how interests are formed and represented. While most democracies rely on single-member districts to ensure geographically grounded representation, some combine different representational logics within the same framework. MMES deliberately balance two democratic goods: territorial responsiveness, achieved through district-based mandates, and proportional fairness, achieved through party lists (Shugart, 2005). MPs who win a

constituency seat represent a specific district (district MPs), while list MPs enter parliament on behalf of their state or national party. This institutional distinction creates two types of electoral dependency, direct and list mandates, resulting in a mandate divide (Wurthmann & Sältzer, 2025).

5.2.2 Representing the Local

Constituency MPs derive their legitimacy from a geographically defined electorate, generating strong expectations of responsiveness to local interests. List MPs, by contrast, owe their mandates to proportional representation and are expected to prioritize programmatic and national concerns. The consequences of this divide are well documented, shaping both parliamentary behavior (Bäck & Debus, 2018; Gschwend & Zittel, 2018) and candidate communication (Schürmann and Stier 2023; Wurthmann and Sältzer 2025). It also manifests in how politicians balance constituency responsiveness with party alignment, particularly regarding the geographic dimension of representation (Zittel et al. 2019). Accordingly, district MPs are incentivized to win votes in their constituencies, focusing representational behavior on places close to their electoral district. Therefore, we argue that

H1: Nominally elected MPs are more likely to mention places closer to their constituency than MPs elected via party lists.

5.2.3 Representing the Deprived

While single-member districts provide a territorial anchor, the political relevance of a place depends on the salience assigned to it by political actors. Geographic representation manifests not only in role perception but also in practice, for instance, when MPs publicly reference their constituencies in political communication (Schürmann and Stier 2023). Such references serve as signals of attentiveness and authenticity, linking representatives to the lived realities of their voters. They are particularly meaningful where citizens perceive neglect by political elites (Cramer 2016; Rodríguez-Pose 2018). Limited access to public services and economic deprivation fuel these perceptions, as the visibility of state institutions serves

as a tangible indicator of political inclusion (Cremaschi et al., 2024; Nyholt, 2024; Stroppe, 2023).

From an electoral perspective, district MPs must attend to deprived areas even when they are sparsely populated, while list MPs can focus on populous urban centers where electoral returns are higher (Gschwend and Zittel 2018). Yet, directly elected MPs perceive these places as part of their representational responsibility. Following role theory (Patzelt 2013), they are guided not only by electoral incentives but also by a sense of appropriateness linked to their territorial mandate. They see themselves as advocates for the entire district, including its most deprived and disadvantaged areas. By publicly referencing such places, they demonstrate attentiveness, empathy, and territorial inclusion.

Perceptions of neglect are not merely material but closely tied to recognition and symbolic inclusion. Voters value representatives who acknowledge their concerns and reward those who do (Cramer 2016). Hence, these localities are not only normatively relevant but also strategically meaningful for directly elected MPs, who depend on visible signals of local connectedness and care of marginalized communities. Therefore, we expect that:

H2: Nominally elected MPs are more likely to mention places that are located farther from public service facilities.

5.3 Research Design

5.3.1 Case and Data Selection

We focus on Germany because its mixed-member proportional system uniquely combines single-member districts with party lists, allowing us to disentangle the distinct incentives arising from each mandate type for geographical representation (Norris 2004; Haffert 2021). Our analysis covers the 19th legislative term (October 10, 2017–October 10, 2021) and draws on data from the public Twitter (now X) and Facebook accounts of all German MPs. These platforms play a central role in contemporary political communication, enabling personalized and targeted outreach. Their interactive nature allows political elites to strategically emphasize proximity and small-scale linkages to their constituencies (Gilardi et al., 2022).

Twitter: The raw Twitter dataset consists of 6.7 million tweets from elected officials published by 485 active accounts in the 19th legislative period. The data was

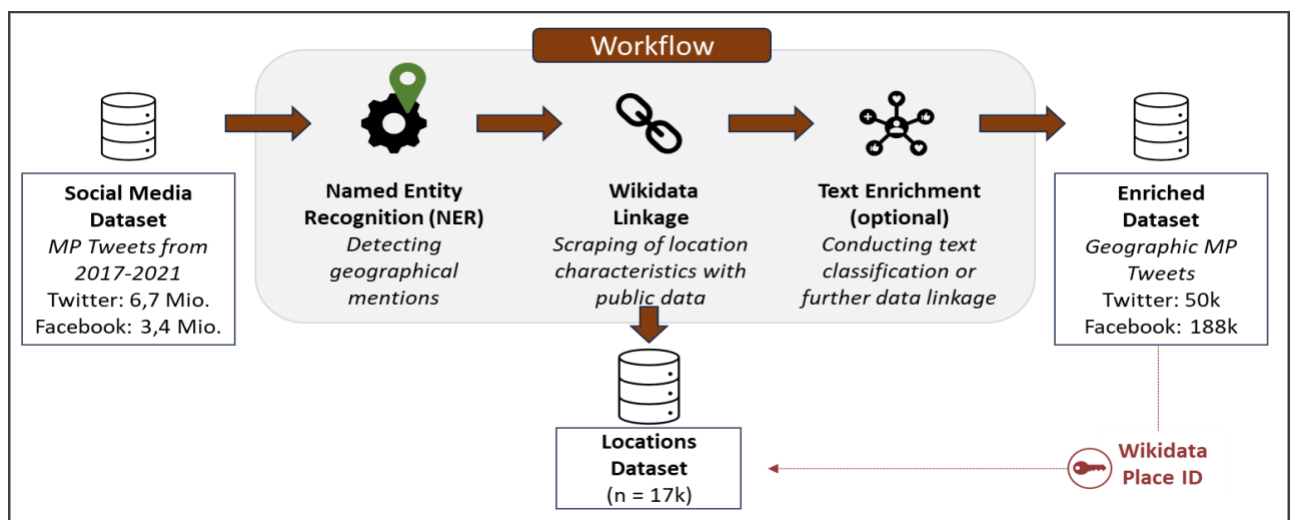
collected in 2021 using the academic Twitter API and includes tweets from MPs elected through party lists as well as direct candidates from Germany’s 299 constituencies.

Facebook: The raw dataset comprises 3.4 million posts published by 573 active accounts during the 19th legislative period. We obtained these data from the Facebook pages collected by Stier et al. (2020) and retrieved all historical posts via Meta’s CrowdTangle platform. As the platform’s terms of use limit access to public pages rather than private profiles, our sample includes only posts from public accounts—an appropriate restriction, since members of the Bundestag typically communicate through public pages

5.3.2 Place Identification

We applied a step-wise procedure to identify locations within our sample of social media posts. Figure 23 visualizes the full workflow.

Figure 23: Computational Workflow



First, we applied named entity recognition (NER) to detect geographical mentions in the initial corpus of Twitter and Facebook data. We rely on the transformer-based model "German NER" (Akbik et al., 2018). The model was trained on the CoNLL-2003 German dataset (Sang & Meulder, 2003) reaching state-of-the-art performance for the German language while also being publicly available on Huggingface³⁸. We preprocess the social media posts by removing all URLs and mentions of accounts from the posts. Ultimately, we feed the data into the model and receive a list of locations for each post.

Wikidata Linkage: For each location identified through named entity recognition, we match and link the location to its corresponding entry in the Wikidata database (Wikimedia Foundation 2024), a comprehensive knowledge base for Wikipedia with over 100 million records. Due to the non-deterministic nature of the matching process, where different locations may share the same name or be referred to by different names, we initially store the three most likely matches for each location in a list (see Appendix for an example). We then gather relevant information for each identified location from the Wikidata database, such as geographic coordinates or the type of location, and store this data in a separate dataset containing 17,685 distinct locations. Following this setup, we query each identified location for each social media post to determine if the location lies within the electoral borders of the MP's constituency. The rationale for this approach is that if multiple locations with the same name exist (e.g., there are 12 different locations named Hausen in Germany), the MP is likely referring to the location within their constituency. If the location is not located within the constituency, the most likely match determined by the query is finalized and linked as the corresponding Wikidata entry.

For our analysis, we define a place as a local municipality within Germany. Municipalities represent the most relevant geographic units because they allow politicians to signal proximity to their constituency and reflect the everyday environments in which voters live. Following this definition, we exclude posts referencing state, national, or international entities, as well as street-level locations or institutions that may cause ambiguous matches. Our analysis thus focuses on cities, towns, districts, and villages.

Measurement

³⁸ <https://huggingface.co/flair/ner-german>

Once all place references have been identified in MPs' social media posts, we process the data in R using the *sf* package (Pebesma & Bivand, 2023). MPs with fewer than ten identified places are excluded from the analysis.

Before testing our hypotheses, we validate our workflow using different strategies, including comparison with human annotations. In addition, we replicate established findings of geographic representation showing that nominated (district) candidates are more likely to mention places within their constituencies than list candidates (Schürmann and Stier 2023). To do so, we create a measure of constituency salience indicating whether each referenced place lies within an MP's constituency (Ho).³⁹ For constituencies located in major cities, a place is also coded as an exact match if it falls within another constituency of the same city (see Appendix). We then compute the share of unique places within constituency borders for each MP.

To measure district distance (H1), we calculate the mean distance of all referenced places from the outer border of each MP's constituency. To measure service distance (H2), we compute the average distance from each place centroid to the nearest general practitioner, pharmacy, elementary school, train station, post office, and supermarket. Location of service facilities stem from Deutsche Bahn's Open Data Portal and the "Points of Interest" dataset curated by the Federal Agency for Cartography and Geodesy (BKG).

5.3.3 Analysis

We use multivariate regression analysis to test our hypotheses. The main independent variable, mandate mode, is not randomly assigned: district and list MPs systematically differ in electoral success, communication focus, and other characteristics that may confound the relationship between mandate type and attention to local geography. We therefore refrain from making causal claims. Our goal is to derive correct descriptive inferences while adjusting for potential confounding factors. To this end, we account for constituency-specific and individual characteristics. Because the distance from the constituency center is naturally constrained by district size, we include a control for constituency area to avoid bias stemming from geographical variation. District size also correlates with population density—cities contain more districts and, hence, more directly elected MPs—which could otherwise attenuate the estimated relationship. In addition, we control for list position, incumbency, and demographic factors such as age

³⁹ Figure 28 in Appendix visualizes this process.

and gender, as these may influence both mandate type and communication behavior. A list of all variables with descriptive statistics is displayed in

5.4 Results

5.4.1 Measurement Validation

We validate our computational workflow for detecting and linking places in political communication using different strategies⁴⁰.

First, we compare both the NER and Wikidata linkage, i.e., the process of matching a place with its corresponding unique Wikidata ID, with human annotations (see Table 14).

Table 14: Validation - Comparison with Human Annotations

Step	Precision	Recall	F1 Score
NER	.86	.88	.87
Wikidata Linkage	.87	.93	.90

For NER, we hand-coded a subsample of 1000 randomly selected social media posts and calculated the recall (i.e., how many of the places identified using human coding were also detected by NER) and precision (i.e., how many of the places identified by the model were correctly identified), resulting in satisfactory metrics (precision = .86, recall = .88, F1 = .87)⁴¹. For Wikidata linkage, we further hand-coded a selection of 500 places with their Wikidata ID. We again reached sufficient congruence (precision = .87, recall = .93, F1 = .90), giving us confidence that the places we identified were matched correctly with their corresponding Wikidata entry.

Second, we also inspect the distribution of places across our data. To do so, we carefully screened the most often mentioned 200 places that were matched in our final dataset (see Appendix). For each of these places, we screened the first 10 social media posts that contained the place. Except for one place⁴², the places were correctly matched.

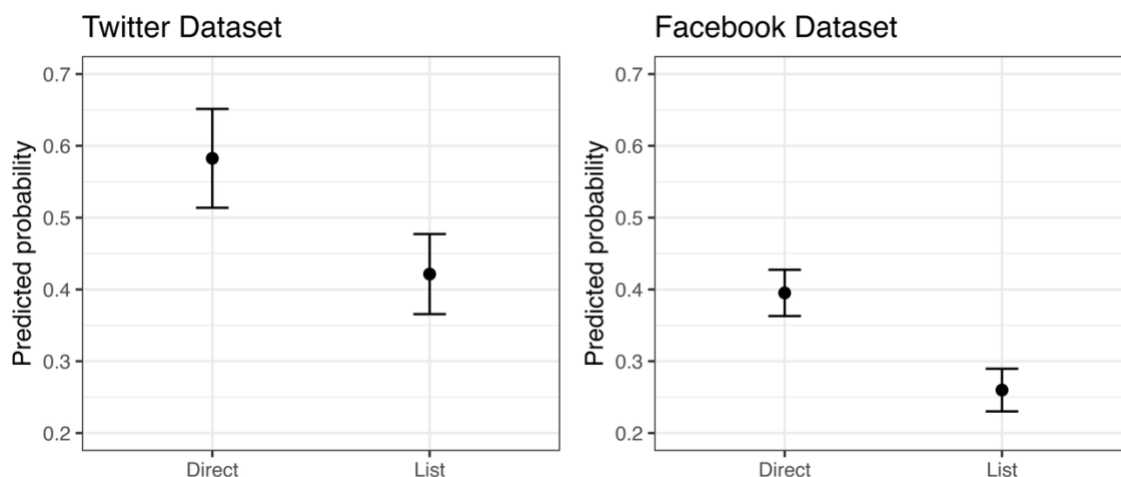
⁴⁰ For a detailed overview of all validation steps, see the checklist downloaded from the ValiText framework (Birkenmaier et al. 2023, 2024), in Appendix

⁴¹ For the NER, we were primarily interested in a high recall ("flagging all potential places even if they turn out to be wrong at the end), as false positives could be filtered out at a later stage of the analysis

⁴² We only removed one place "Thüringenhausen", which was predominantly mismatched to the German state "Thüringen"

Third, we inspect the predictive validity of our measure by replicating established findings of constituency salience, i.e., politicians who are directly elected by their constituency being more likely to refer to places within their constituency than list candidates (Schu"rmann and Stier 2023; Zittel et al. 2019). Figure 24 presents predicted values from a multivariate regression analysis, using the share of places

Figure 24: Predicted Values for Mandate Mode and District Match.



located within a politician's constituency as the dependent variable.⁴³

The Figure shows that in both datasets, directly elected candidates are more likely to mention places that are within the constituency of the MP compared to list candidates. On Twitter, the predicted probability is 0.58 for directly elected MPs versus 0.42 for list MPs, while on Facebook it is 0.40 versus 0.26.

Taken together, the validation steps in this chapter demonstrate that our computational approach to identifying and linking places yields a valid measure, which we can confidently employ in the subsequent analysis.

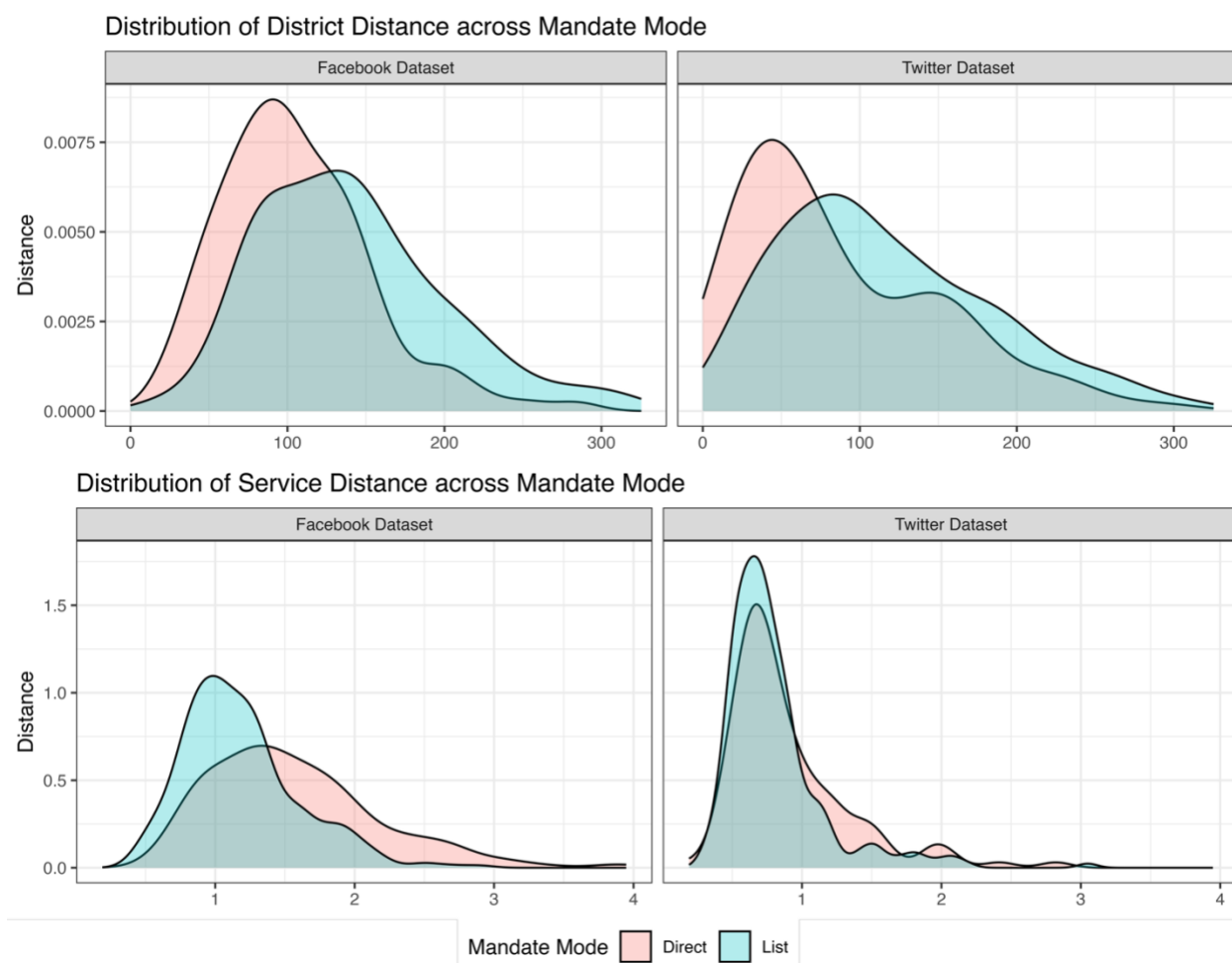
5.4.2 Descriptive Statistics

Figure 25 presents the descriptive statistics for both measures, district distance (DV H1) and service distance (DV H2), differentiated by mandate mode (IV).

⁴³ Errorbars represent 95% confidence intervals.

Figure 25 shows clear differences between the two groups in the types of places they reference. The top row (H1) illustrates district distance: directly elected MPs predominantly mention locations within or near their constituency, whereas list candidates are more likely to refer to distant

Figure 25: Distributions of District Distance and Service Distance

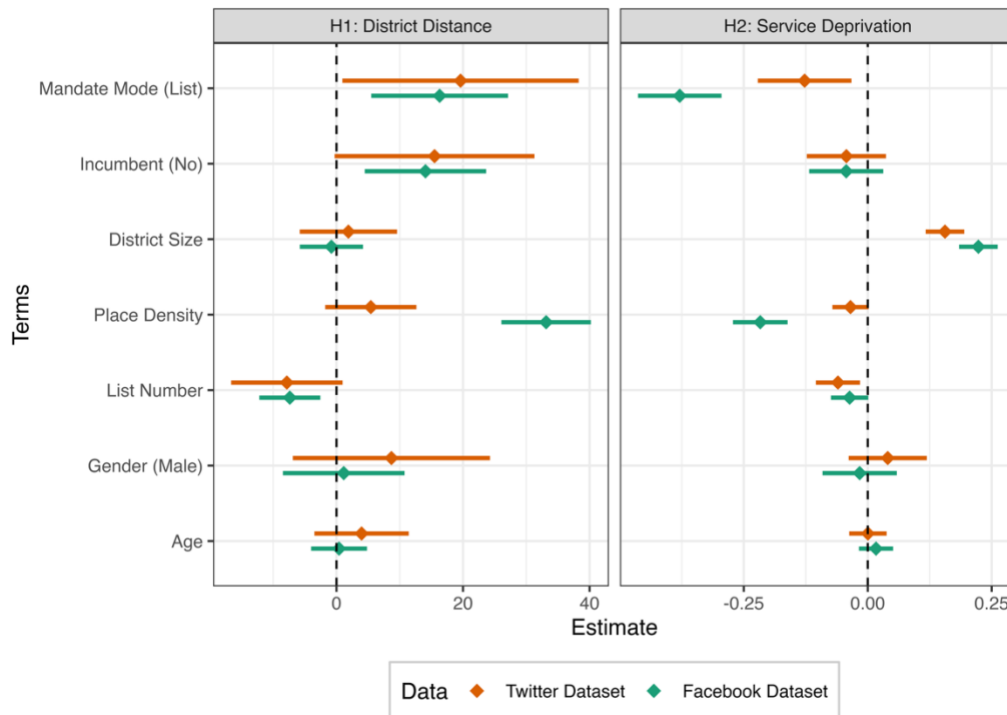


places. The bottom row (H2) depicts service distance: directly elected MPs talk more about places with lower levels of service accessibility, while list candidates tend to focus on well-served localities. This pattern is pronounced in the Facebook dataset but less evident in the Twitter data.

5.4.3 Hypotheses

Figure 26 displays the scaled regression coefficients for both outcomes: District Distance (H1, left panel) and Service Distance (H2, right panel).

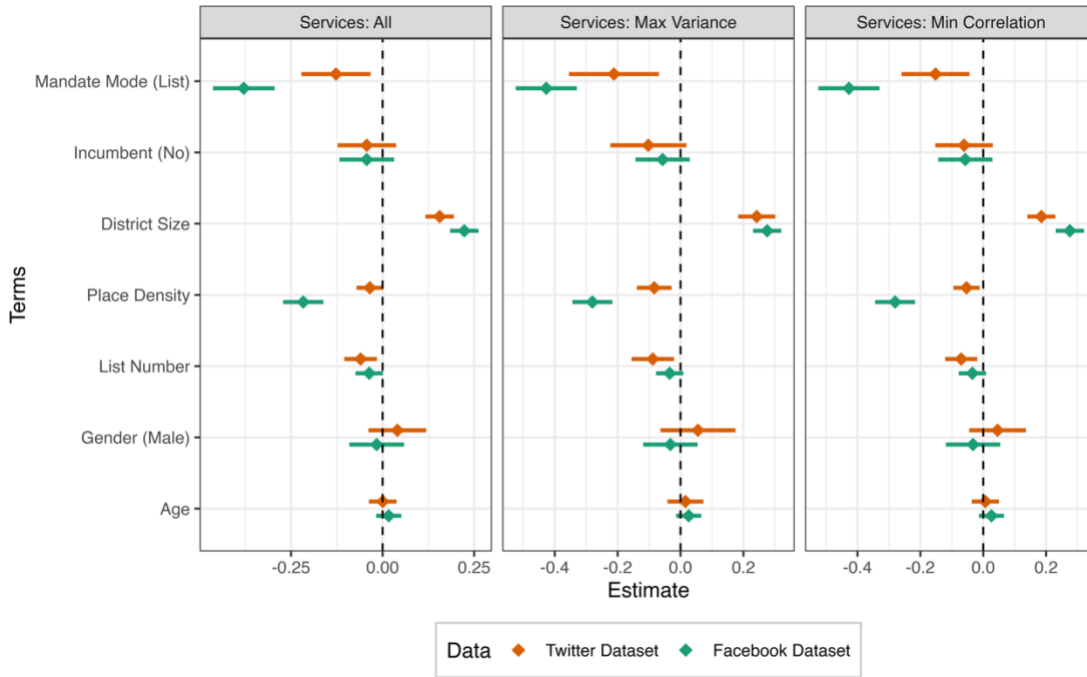
Figure 26: Regression Coefficients



Orange estimates refer to communication on Twitter, while green estimates represent communication on Facebook. For H1, we observe that nominally elected MPs are more likely to mention places closer to their constituency compared to party list candidates, supporting our expectations. Yet, the confidence intervals lie close to zero, particularly in the Twitter dataset, where they are comparatively wide ($p = .04$). For H2, the effect of mandate mode also aligns with our expectations: Nominally elected MPs are more likely to mention deprived places, i.e., places that lack access to public services. However, the size of the estimates differ between the two datasets, with stronger effects in the Facebook data. Overall, the results provide consistent, though not uniform, support for our hypotheses across platforms.

To assess the robustness of these findings, Figure 27 presents results based on alternative operationalizations of service deprivation.

Figure 27: Robustness Checks



Including all types of services, such as supermarkets, may introduce bias in the measurement due to low spatial variability and inflate correlation within the index. Two additional specifications were tested: (1) maximum variance, indicating a measure based on hospitals, train stations, and secondary schools, which captures greater spatial variation in service availability; and (2) minimum correlation, indicating a measure based on hospitals, train stations, and elementary schools, which reduces inter- service correlations while incorporating a more evenly distributed type of service. Across both measures, the results remain consistent, reinforcing the robustness of the observed effects.

5.5 Discussion and Conclusion

Our analysis aimed to shed light on the institutional motivation for geographical representation in an MMES. We developed a new method designed to detect places in political communication and applied it to a social media dataset of elected MPs. Our results show that the mandate mode has a significant and strong impact on geographic representation on social media. Our analysis extends previous research by matching, linking, and geolocating places using the Wikidata database. We showcase the advantage of our method by deriving the distance of places from constituency and deprivation in terms of access to everyday necessities. As expected, nominally elected

MPs are more likely to mention places located within or near their constituency, as well as places that lack access to essential services. This suggests that they follow, compared to list candidates, two modes of geographic representation: first, attending to locations within or close to their constituency to win the votes of their constituents; and second, signaling attentiveness and symbolic inclusion of underserved areas.

Our method will allow for a more thorough and detailed understanding of geographic representation. By linking places to a variety of characteristics, including economic, social, and other indicators, we can improve insights into the mechanisms driving geographic representation. Future research might explore the factors that affect how politicians perceive and communicate certain places. Is "Washington" or "Berlin" just a meeting place, a constituency, or a symbolic "swamp" of corruption? Can we detect stereotypical depictions of places that serve political agendas (Haffert 2021)? Do politicians' statements about geographic entities align with the actual characteristics of these locales, whether they are depicted as "crime hotspots" or "left-behind places"? Unpacking these mechanisms of geographic representation could enhance our understanding beyond the mere mentioning of places in political discourse.

However, our analysis also comes with limitations. On the one hand, our definition of places is limited to municipalities and does not include broader geographic references or street-level entities, such as neighborhoods, culturally or territorially defined regions, and cultural landmarks and institutions. Nevertheless, our method generally supports detailed handling and filtering of places based on their properties in the Wikidata database. For example, researchers can adapt the query in the matching process to also include "metropolitan regions" or "cultural landmarks". This will be particularly useful for researchers with domain knowledge who are interested in specific regions or policy contexts and wish to incorporate various forms of place into their analysis.

While we recognize that signaling behavior on social media plays a crucial role for geographic representation (Zittel et al. 2019), we still lack sufficient knowledge on the overall impact of constituency salience on voters' perception of politicians and their electoral reward. Therefore, additional research is necessary to investigate the consequences of signaling on the demand side of the politician-voter relationship, specifically how voters perceive and evaluate geographic representation from their elected officials.

5.6 Appendix

5.6.1 Appendix: Exemplary SPARQL Query

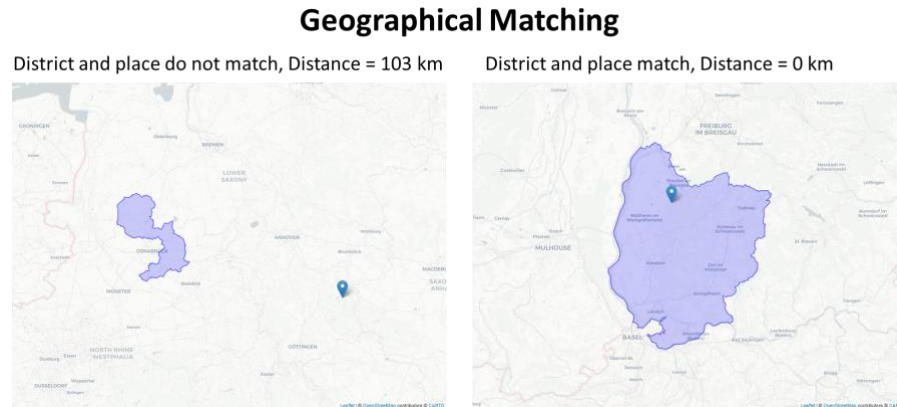
You can click on the link below to get to the API Query Interface.

<https://query.wikidata.org/#SELECT%20%3Fitem%20%28COUNT%28%3Fsitelink%29%20AS%20%3Fsites%29%20WHERE%20%7B%0A%20%20SERVICE%20wikibase%3Amwapi%20%7B%0A%20%20%20%20bd%3AServiceParam%20wikibase%3Aapi%20%27EntitySearch%27%20.%0A%20%20%20%20bd%3AServiceParam%20wikibase%3Aendpoint%20%27www.wikidata.org%27%20.%0A%20%20%20%20bd%3AServiceParam%20mwapi%3Asearch%20%27ahrtal%27%20.%20%23hier%20kannst%20du%20die%20suchbegriffe%20ab%20C%3A4ndern%20C%20die%20top%20%20%20%20%20bd%3AServiceParam%20mwapi%3Alanguage%20%27de%27%20.%0A%20%20%20%20%3Fitem%20wikibase%3AapiOutputItem%20mwapi%3Aitem.%0A%20%20%20%20%7D%0A%20%20%20%3Fsitelink%20schema%3Aabout%20%3Fitem.%0A%20%20%20%3Fitem%20wdt%3AP625%20%3FcoordinateLocation.%20%20%23%20Ensure%20item%20has%20a%20coordinate%20location%20%28indirectly%20filtering%20for%20places%29%0A%20%20MINUS%20%7B%3Fitem%20wdt%3AP31%20wdt%3AQ4167410.%7D%0A%7D%20GROUP%20BY%20%3Fitem%20ORDER%20BY%20DESC%28%3Fsites%29%20LIMIT%203>

```
SELECT ?item (COUNT(?sitelink) AS ?sites) WHERE { SERVICE wikibase:mwapi {  
  bd:serviceParam wikibase:api 'EntitySearch' . bd:serviceParam wikibase:endpoint  
'www.wikidata.org' bd:serviceParam mwapi:search 'ahrtal' . #example  
  bd:serviceParam mwapi:language 'de' .  
  ?item wikibase:apiOutputItem mwapi:item.  
}  
?sitelink schema:about ?item.  
?item wdt:P625 ?coordinateLocation. MINUS {?item wdt:P31 wd:Q4167410.}  
} GROUP BY ?item ORDER BY DESC(?sites) LIMIT 3
```

5.6.2 Appendix: Geographical Matching

Figure 28: Matching Graph



5.6.3 Appendix: Constituencies

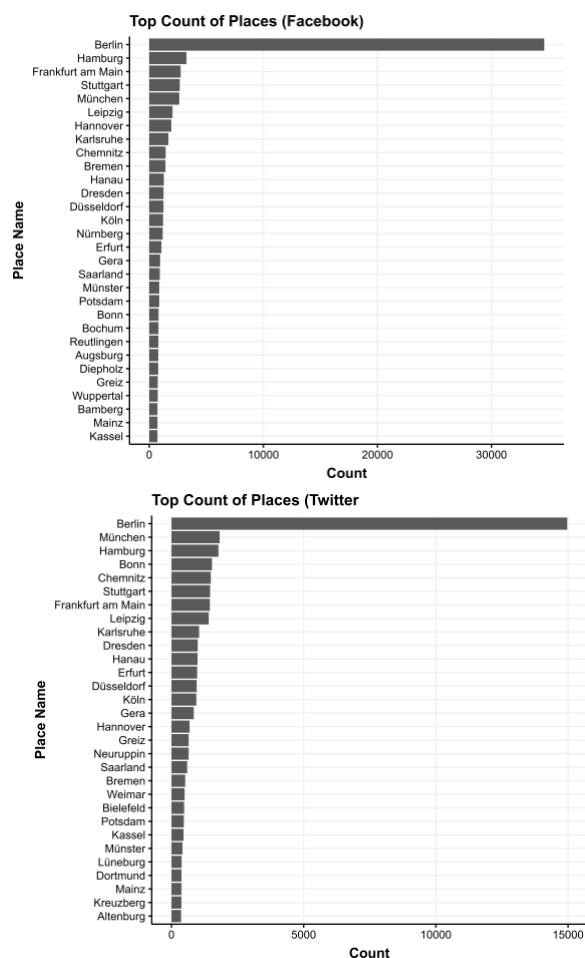
Wahlkreis	City
18	Hamburg
19	Hamburg
20	Hamburg
21	Hamburg
22	Hamburg
23	Hamburg
76	Berlin
77	Berlin
78	Berlin
79	Berlin
80	Berlin
81	Berlin
82	Berlin
83	Berlin
84	Berlin
85	Berlin
86	Berlin
115	Duisburg
116	Duisburg
118	Essen
119	Essen

120	Essen
182	Frankfurt
183	Frankfurt
217	München
218	München
219	München
220	München
244	Nürnberg
245	Nürnberg
258	Stuttgart
259	Stuttgart
93	Köln
94	Köln
95	Köln
101	Köln
106	Düsseldorf
107	Düsseldorf
140	Bochum
141	Bochum
142	Dortmund
143	Dortmund
41	Hannover
42	Hannover \

5.6.4 Appendix: Validation

Tba.

5.6.5 Appendix: Descriptive Overview Places



5.6.6 Appendix: Descriptive Overview Accounts

Table 15: Top 10 Accounts with no. Tweets

Last Name	First Name	Party	No Tweets
Brandner	Stephan Gü nther	AfD	6534
Jung	Christian	FDP	2529
Steineke	Sebastian	CDU	2030
Wanderwitz	Marco	CDU	966
Dö rner	Katja	GRÜ NE	929
Domscheit-Berg	Anke Margarete	LINKE	818
Bayram	Canan	GRÜ NE	789
Schipanski	Tankred	CDU	760
Rü ddel	Erwin Josef	CDU	751
Wegner	Kai	CDU	725

Table 16: Top 10 Accounts with no. Facebook Posts

Last Name	First Name	Party	No FB Posts
Brandner	Stephan Gü nther	AfD	4342
Jung	Christian	FDP	3903
Knoerig	Axel	CDU	2623

Donth	Michael	CDU	2044
Nissen	Ulrike	SPD	1839
Wiese	Dirk	SPD	1460
Oehme	Ulrich	AfD	1364
Schwabe	Frank	SPD	1243
Gastel	Matthias	GRÜNE	1079
Schraps	Johannes	SPD	1019

5.6.7 Appendix: Descriptive Statistics

Table 17: Descriptive Statistics

Twitter					
Variable	N	Mean	St. Dev.	Min	Max
District Match	364	0.429	0.266	0.000	1.000
District Distance	364	107	68	0	321
Service Distance	364	0.834	0.389	0.190	3.049
District Size	364	9,959	10,226	266	62,516
Place Density	364	1,762	1,230	131	13,205
Mandate (1=Direct Candidate)	364	0.327	0.470	0	1
Gender (1=Male)	364	0.687	0.464	0	1
Birthyear	364	1,970.242	9.382	1,944	1,992
List Number	338	7.459	10.097	1	63
Facebook					
Variable	N	Mean	St. Dev.	Min	Max
District Match	545	0.287	0.184	0.000	0.913
District Distance	545	125	57	12	325
Service Distance	545	1.346	0.540	0.422	3.948
District Size	545	10,941	9,568	266	62,516
Place Density	545	1,247	715	237	5,532
Mandate (1=Direct Candidate)	545	0.429	0.495	0	1
Gender (1=Male)	545	0.701	0.458	0	1
Birthyear	545	1,968.892	9.627	1,942	1,992
List Number	492	9.577	11.739	1	64

6. Conclusion

6.1.1 Main Findings and Contributions

6.2 Research Objectives

At the heart of this dissertation lies a key methodological challenge in social science research: how can researchers ensure valid CTAM measures of social constructs? To answer this question, this dissertation pursued three major research objectives.

Following the introduction in the first chapter of this dissertation, Part 1 of this dissertation set out with the first research objective of **understanding contemporary practices in the validation of CTAM measures of social constructs**. To this end, Chapter 3 reported on a systematic review of 96 CTAM studies in the field of political communication and insights gained from expert interviews. Results from the systematic review show that, in the majority of reviewed studies, researchers did engage in some form of validation of CTAM measures of social constructs, with only 14 per cent of studies reporting no validation. However, a closer look revealed a tendency to prioritize external validation - such as comparison with surrogate labels or human annotation - over internal validation. When internal validation was conducted, it often included a variety of techniques aimed at assessing both the statistical properties of the model and the interpretability or coherence of its output. The qualitative interviews with experts in the field of computational text analysis offered further insights into validation practices applied in the research community. Experts consistently reported that they carried out additional validation steps that were often not documented in their published work. These steps frequently involved conceptual considerations, such as refining construct definitions, making design decisions, as well as iterative practices like exploratory data analysis using descriptive statistics and the qualitative examination of model errors. Importantly, the empirical description of contemporary validation practices also laid the groundwork for a preliminary taxonomy of validation steps. This taxonomy has been used and explicitly referenced in recent work that measures constructs such as *calls to action* (Achmann-Denkler et al., 2024), *populism* (Zehnter, 2024), or *moral themes* (Malik et al., 2025).

Part 2 of the dissertation targeted the second research objective of **developing conceptual and practical guidance for CTAM validation**. Hence, Chapter 4 of this dissertation introduced the ValiText framework - a validation framework for CTAM measures of social constructs. The framework builds on the findings of the systematic review and a comprehensive assessment of methodological literature on measurement and validation in related fields of social science. First, it identifies the key types of validation evidence needed to ensure complementary approaches to validation in computational text-based research. Second, it offers practical guidance on which concrete validation steps researchers can consider when designing and evaluating their text-based measures. Third, it includes a documentation template in the form of a checklist that enables researchers to systematically record and communicate validation efforts.

Ultimately, Part 3 of the dissertation **demonstrated the usefulness of the framework for improving validation of CTAM measures of social constructs in applied research**. Toward that end, Chapters 5 and 6 presented empirical applications of the framework in the field of political communication. On the one hand, Chapter 5 presents a feasibility study that explores how politicians' public personality traits, i.e., agency and communion, can be measured using CTAM, drawing on *personality cues* as indicators of the respective traits. The validation included complementary types of validation evidence along substantive, structural, and external validation, including comparison with human annotations, systematic tests, and sensitivity analysis for partisan differences in German and US political contexts. The analysis further revealed that a wide range of substantive research questions, such as the impact of gender or partisan attitudes on the display of public personality traits, can be answered using valid CTAM.⁴⁴

On the other hand, Chapter 6 presents substantive research that assesses the impact of the mandate mode on politicians' *geographic representation*. The construct "geographic representation" was operationalized through the mentioning of geographic places, where the workflow of detecting these places automatically was evaluated using different validation steps, including comparisons with human annotations, descriptive analyses, and predictive validity checks by replicating

⁴⁴ As outlined in Chapter 5, we present extensive exploratory analyses touching upon a range of novel substantive research questions. However, because Chapter 5 presents a feasibility study based on a convenience sample and focused on validating CTAM measures of personality traits rather than generating new substantive insights, we refrain from making strong claims about our substantive findings and instead leave these questions for future research.

established findings on constituency salience from the literature. The analysis furthermore revealed that electoral system design—specifically mixed-member systems—affects politicians’ incentives to engage in geographic representation. Directly elected politicians, in contrast to those elected via party lists, were more likely to refer to geographically proximate locations and to areas characterized by service deprivation, indicating responsiveness to local needs and strategic use of regional signalling.

6.3 Contributions to the Research Community

In addition to the chapters addressing the three core research objectives, this dissertation also contributes to the wider social science research community by connected research as well as developing resources and datasets that can be used in subsequent research.

As such, the dissertation has resulted in a novel dataset of German politician public broadcast interviews (Birkenmaier, Sieber, et al., 2025), as well as publicly accessible resources and tutorials that document and outline the computational workflows applied. Specifically, these resources are the tutorial “ValiText - A Validation Framework for Text-Based Measures of Social Constructs” in the GESIS toolbox,⁴⁵ an educational resource from the Competence Center for Data Quality (KODAQs) (see Birkenmaier, 2024), and the tutorial “Linking Entities to the Wikidata Database” in the GESIS Methods Hub,⁴⁶ a collaborative online platform dedicated to advancing social science research by sharing computational methods and related tutorials for using, managing, and enriching digital behavioral data (see Birkenmaier, 2025). Moreover, all trained models, code, and datasets are thoroughly documented and reported in the respective chapters, with the exception of the social media material used in Chapters 4 and 5, which - owing to confidentiality and platform restrictions - can only be provided in aggregated form.

In addition to providing practical resources for validating CTAM measures of social constructs, this dissertation has also generated several publications addressing the issue of data quality. A persistent challenge in validating CTAM measures is precisely this problem of low data quality. During the examination of this topic, it became clear that a unified understanding of what constitutes data quality, particularly

⁴⁵ Online available at: <https://kodaqs-toolbox.gesis.org/>

⁴⁶ Online available at: <https://methodshub.gesis.org/>

for text data, but also for other forms of social science data, was lacking. To address this gap, we conducted a systematic review (Daikeler et al., 2023) and proposed a unified definition of data quality (Birkenmaier, Daikeler, et al., 2024), which subsequently informed several chapters of this dissertation.

One highly relevant application of data quality research in the field of CTAM is the use of synthetic training data for training, fine-tuning or prompting large language models. In Birkenmaier, Roth, and Sen (2024), I systematically examine the potential of theory-driven synthetic training data for advancing the measurement of social constructs. Specifically, we investigate how established knowledge from social science measurement instruments, such as survey scales or annotation codebooks, can be transferred into the theory-driven generation of synthetic data. Our findings show that this approach can substantially reduce the reliance on manually labeled data in political topic classification, while maintaining performance with only minimal losses.

6.4 Limitations

In addition to the limitations outlined in the individual chapters, this dissertation is subject to several overarching constraints that limit the generalizability of this dissertation.

First, the rapidly evolving landscape of computational text analysis poses a fundamental challenge to the timeliness and completeness of any methodological contribution. Over the course of this dissertation, significant developments have occurred in research using CTAM - most notably the shift from *fine-tuning* large language models like RoBERTa to prompting off-the-shelf LLMs, which now dominate many contemporary applications (Trad & Chehab, 2024). At the same time, LLMs themselves have grown increasingly powerful, with open-source alternatives such as LLaMA, DeepSeek, and others offering high-quality performance and greater accessibility (DeepSeek-AI et al., 2024; P. Huang et al., 2025). While these advances open up exciting possibilities, they can also render certain techniques and examples discussed in this dissertation provisional. For instance, although a particular prompt design has been shown to yield the best results for identifying personality cues in political texts (see Chapter 5), the same design may become outdated with newer versions of the same model family or with other LLMs. Consequently, prompt-based methods in particular raise novel questions about reproducibility, consistency, and documentation—areas in which best practices are still evolving. The methodological

decisions, such as model and preprocessing choices, presented in this dissertation should therefore be interpreted in light of these ongoing technological developments.

Second, the validation framework developed in Chapter 3 offers only one conceptual perspective to the validation of CTAM measures of social constructs. Although the framework is grounded in conceptual reasoning and informed by methodological traditions from disciplines such as psychometrics and political science, it necessarily reflects a selective synthesis of existing knowledge. Other perspectives - such as those focusing more explicitly on the convergence with human labels - may yield different understandings of what constitutes valid measurement in computational text analysis. Therefore, while the proposed framework aims to provide conceptual clarity and practical orientation, it should not be regarded as exhaustive or universally applicable.

Third, the process of validation itself demands a high degree of flexibility and contextual sensitivity - qualities that resist rigid standardization. Determining when a computational measure is “valid enough” is not a straightforward matter. For instance, performance metrics like the F1 score offer a commonly used benchmark: values above 0.8 are often interpreted as strong, while those below 0.5 may be deemed inadequate. However, such cut-off points are not absolute. Their meaning depends heavily on the nature of the construct being measured, the complexity of the classification task, the quality of available data, and the goals of the analysis. Researchers must therefore be able to interpret performance metrics in light of their specific measurement design, potentially using relative benchmarks such as comparisons with prior studies or internal model selection criteria. This situational judgment, while necessary, introduces an element of subjectivity that cannot be entirely eliminated.

Finally, the dissertation is primarily grounded in the context of social science research, which may limit its scope and generalizability. While this focus provides a strong foundation for discussing the measurement and validation of *social science* constructs, it may overlook important trends and methodological advances from adjacent fields, particularly computational linguistics and natural language processing.⁴⁷ Traditionally, these fields often orient validation along maximizing performance metrics, but recently the field also engages extensively with critical issues for the measurement of social constructs such as annotation subjectivity, dataset bias,

⁴⁷ For more information on current debates in the field of computational linguistics, a good starting point is the Association for Computational Linguistics (ACL) available at: <https://aclanthology.org/>

and model evaluation (Plank, 2022; Röttger et al., 2022; Shankar et al., 2024; Zhou et al., 2023). Incorporating these perspectives more fully could enrich future work and help bridge the gap between subfields of empirical research on human behavior.

Taken together, these limitations underscore the importance of continued dialogue, reflexivity, and methodological pluralism in the development of computational text analysis as a tool for measuring social constructs.

6.5 Further Directions for Research

Given the limitations of this dissertation, several directions for further research can be identified.

Including Subjectivity & Uncertainty in Computational Text Analysis Research: Any attempt to measure social science constructs in text is inevitably associated with uncertainty. This is, most prominent, visible in the variability of human annotations on the same text: even when obvious labeling errors like fatigue or lack of concentration are accounted for, differences in interpretation remain both plausible and widespread (Plank, 2022). These differences are not just noise; rather, they can reflect the fact that multiple, sometimes plausible interpretations of the same text can coexist, particularly in cases with low inter-annotator agreement (Nie et al., 2020). CTAM, despite its sophistication and growing role as an alternative to manual coding, inherits this fundamental uncertainty. Yet, this uncertainty is often masked in computational text analysis - both in the labeled datasets used for training and evaluation, and in the outputs of the models themselves, where predictions are typically collapsed into the single most likely class.⁴⁸ Acknowledging uncertainty not as a nuisance but as an inherent characteristic of CTAM measures of social constructs requires a shift in research practice that moves beyond binary notions of capturing the presence or absence of social constructs. Instead, researchers should explicitly integrate uncertainty in their research, whether in annotation (e.g., by encouraging contrasting annotation paradigms and label sets (Röttger et al., 2022)) or in model training (e.g., by deliberately developing models that reflect diverse annotator perspectives or sociodemographic backgrounds (Fleisig et al., 2024; Ni et al., 2025)).

⁴⁸ The inherent ambiguousness in measurement of social constructs is not unique to computational text analysis: all methods that seek to capture latent constructs in the social sciences carry inherent uncertainty. Textual analysis, however, deserves special attention because text is usually collected passively, leaving fewer possibilities for applying error-mitigation strategies compared to e.g., survey-based measurement approaches (see Total Survey Error Framework (Groves et al., 2011) or similar).

Interpretability: Validating CTAM measures of social constructs requires methodological tools that allow researchers to critically reflect on both their research design and the specific CTAM they employ. In this context, explainable AI (XAI) methods can serve as a crucial bridge between CTAM and human semantic understanding (Das & Rad, 2020; Linardatos et al., 2020). Given the rapid developments of CTAM, validating CTAM measures of social constructs calls for stronger methodological innovation and the further integration of XAI methods in the social science research process. Hence, science research must engage with these emerging tools to critically assess research designs, uncover potential sources of misunderstanding, and strengthen substantive validation of CTAM applications (Swamy et al., 2025).

Integration of substantive research context: A unique advantage of social scientists who rely on CTAM measures of social constructs is their expertise in the contexts in which these constructs are studied. Measuring abstract constructs like “populist language” through CTAM approaches alone can only reveal part of the picture. To understand how populism takes root and evolves, one also needs to know the stories behind the numbers: the historical experiences of countries, their political institutions, and the cultural meanings attached to populist rhetoric. For instance, the same phrase may carry very different connotations in Spain, Hungary, or Germany, depending on national histories and public debates. This is where social scientists’ contextual expertise becomes indispensable. They can situate abstract constructs within broader theoretical traditions, connect them to historical and political developments, and identify the specific facets of populism that matter in a given case (see Zehnter, 2024). In this way, their research not only measures populism but also explains why it resonates differently across societies.

6.6 Concluding Remarks

Looking back, this dissertation set out to answer a deceptively simple question: how can researchers ensure valid CTAM measures of social constructs? Along the way, it has become clear that this question is not only about improving methods, but about strengthening the foundations of how we connect computational tools with social meaning. The work has shown that validation is neither a purely technical step nor a static checklist, but rather a continuous process of reflection, testing, and learning. By outlining conceptual guidance, shared terminology, and practical tools, this

dissertation hopes to make that process more transparent, systematic, and collaborative.

I therefore end this dissertation with a simple call to action: to treat validation not as an afterthought, but as part of the creative and critical core of research that makes research truly meaningful. When researchers take validity seriously, not as a burden, but as a shared responsibility, the quality and credibility of our findings improve, as does our understanding of the social world they aim to represent. If this dissertation contributes, even in a small way, to that ongoing effort, then it has achieved its goal.

7. Bibliography

- Achmann-Denkler, M., Fehle, J., Haim, M., & Wolff, C. (2024). *Detecting Calls to Action in Multimodal Content: Analysis of the 2021 German Federal Election Campaign on Instagram* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2409.02690>
- Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3), 529–546.
- Aichholzer, J., & Willmann, J. (2020). Desired personality traits in politicians: Similar to me but more of a leader. *Journal of Research in Personality*, 88, 103990.
- Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649. https://aclanthology.org/C18-1139/?utm_campaign=piqcy&utm_medium=email&utm_source=Revue%20newsletter
- Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., & Rauber, A. (2014). Visual methods for analyzing probabilistic classification data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 1703–1712.
- Alvero, A. J., Lee, J., Regla-Vargas, A., Kizilcec, R. F., Joachims, T., & Antonio, A. L. (2024). Large language models, social demography, and hegemony: Comparing authorship in human and synthetic text. *Journal of Big Data*, 11(1), 138. <https://doi.org/10.1186/s40537-024-00986-7>
- American Educational Research Association (Ed.). (2011). *Report and recommendations for the reauthorization of the institute of education sciences*. American Educational Research Association.
- Amsalem, E., Fogel-Dror, Yair, Shenhav, Shaul R., & and Sheaffer, T. (2020). Fine-Grained Analysis of Diversity Levels in the News. *Communication Methods and Measures*, 14(4), 266–284. <https://doi.org/10.1080/19312458.2020.1825659>
- Arnesen, S., & Peters, Y. (2018). The Legitimacy of Representation: How Descriptive, Formal, and Responsiveness Representation Affect the Acceptability of Political Decisions. *Comparative Political Studies*, 51(7), 868–899. <https://doi.org/10.1177/0010414017720702>
- Arnold, C., Biedebach, L., Küpfer, A., & Neunhoeffler, M. (2023). The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*.
- Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4), 555–596. <https://doi.org/10.1162/coli.07-034-R2>
- Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review*, 11(2), 150–166. <https://doi.org/10.1177/1088868306294907>

- 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>
- Bäck, H., & Debus, M. (2018). Representing the Region on the Floor: Socioeconomic Characteristics of Electoral Districts and Legislative Speechmaking. *Parliamentary Affairs*, 71(1), 73–102. <https://doi.org/10.1093/pa/gsx014>
- Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Shababo, G., & Van der Velden, M. (2022). *Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis*. Project Report. https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D6.2.pdf
- Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, 14(3), 165–183.
- 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*, 16(1), 1–18.
- Baden, C., & Tenenboim-Weinblatt, K. (2017). Convergent news? A longitudinal study of similarity and dissimilarity in the domestic and global coverage of the Israeli-Palestinian conflict. *Journal of Communication*, 67(1), 1–25.
- Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion*. <https://psycnet.apa.org/record/1967-02200-000>
- Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *Journal of Communication*, 71(5), 715–738.
- Bakker, M., Veldkamp, C. L. S., Assen, M. A. L. M. van, Cromptvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, 18(12), e3000937. <https://doi.org/10.1371/journal.pbio.3000937>
- Ballard, A. O., DeTamble, R., Dorsey, S., Heseltine, M., & Johnson, M. (2022). Dynamics of Polarizing Rhetoric in Congressional Tweets. *Legislative Studies Quarterly*, lsq.12374. <https://doi.org/10.1111/lcq.12374>
- Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6, 587–604. https://doi.org/10.1162/tacl_a_00041
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
- Berelson, B., & Lazarsfeld, P. F. (1948). *The analysis of communication content*. Universitetets Studentkontor.
- Bernhard, J., Teuffenbach, & Boomgaarden, H. G. (2023). Topic Model Validation Methods and their Impact on Model Selection and Evaluation. *Computational*

- Bestvater, S. E., & Monroe, B. L. (2023). Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, 31(2), 235–256.
- Birkenmaier, L. (2024). *A Validation Framework for Text-Based Measures of Social Constructs* [Computer software]. Methods Hub. <https://doi.org/10.71627/VALITEXT>
- Birkenmaier, L. (2025). *Linking Entities to Wikidata Database* [Computer software]. <https://methodshub.gesis.org/library/tutorials/method-hub-linkage/>
- Birkenmaier, L., Daikeler, J., Fröhling, L., Gummer, T., Lechner, C., Lux, V., Schwalbach, J., Silber, H., Weiß, B., Weller, K., Wolf, C., Abel, D., Breuer, J., Dietze, S., Dimitrov, D., Döring, H., Hebel, A., Hochman, O., Jünger, S., ... Ziaja, S. (2024). Defining and Evaluating Data Quality for the Social Sciences: Position Paper. *GESIS Papers*, 2024/06, 32 S. <https://doi.org/10.21241/SSOAR.96764>
- Birkenmaier, L., Lechner, C. M., & Wagner, C. (2023a). The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. *Communication Methods and Measures*, 1–29. <https://doi.org/10.1080/19312458.2023.2285765>
- Birkenmaier, L., Lechner, C., & Wagner, C. (2023b). *ValiTex—A unified validation framework for computational text-based measures of social science constructs*. <https://doi.org/10.48550/ARXIV.2307.02863>
- Birkenmaier, L., Roth, M., & Sen, I. (2024). *From Measurement Instruments to Data: Leveraging Theory-Driven Synthetic Training Data for Classifying Social Constructs* (No. arXiv:2410.12622). arXiv. <https://doi.org/10.48550/arXiv.2410.12622>
- Birkenmaier, L., Sältzer, M., Stroppe, A.-K., & Wurthmann, C. (2025). *The distance divide—Geographic representation in a mixed-member system*.
- Birkenmaier, L., Sieber, L., & Bergstein, F. (2025, January 8). *PolInterviews—A Dataset of German Politician Public Broadcast Interviews*. arXiv.org. <https://arxiv.org/abs/2501.04484v2>
- Birkenmaier, L., Wagner, C., & Lechner, C. (2024). *ValiText—A unified validation framework for computational text-based measures of social constructs* (No. arXiv:2307.02863). arXiv. <https://doi.org/10.48550/arXiv.2307.02863>
- Bittner, A., & Peterson, D. A. (2018). Introduction: Personality, party leaders, and election campaigns. *Electoral Studies*, 54, 237–239.
- Bogner, A., Littig, B., & Menz, W. (2009). *Interviewing experts*. Springer. https://books.google.de/books?hl=de&lr=&id=m5SGDAAAQBAJ&oi=fnd&pg=PP1&dq=Interviewing+Experts&ots=RIxQWgXdwe&sig=CzHq1GwPX1GkbsSFm_ToxEBiOl8
- Borsboom, D., Mellenbergh, G., & Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111, 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Boukes, M., Velde, B., Araujo, T., & Vliegenthart, R. (2019). What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf

- Sentiment Analysis Tools. *Communication Methods and Measures*, 14, 1–22. <https://doi.org/10.1080/19312458.2019.1671966>
- Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Elsevier. <https://www.sciencedirect.com/science/article/pii/S1573441201050127>
- Boxman-Shabtai, L. (2020). Meaning Multiplicity Across Communication Subfields: Bridging the Gaps. *Journal of Communication*, 70(3), 401–423.
- 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>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. <https://doi.org/10.48550/arXiv.2005.14165>
- Bruinsma, B., & Gemenis, K. (2019). Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling. *Communication Methods and Measures*, 13(3), 212–227. <https://doi.org/10.1080/19312458.2019.1594741>
- Brunswik, E. (1952). *The conceptual framework of psychology*. Univ. of Chicago Press.
- Burlacu, A. (2021, July 26). *Going beyond simple error analysis of ML systems*. Alexandruburlacu.Github.Io. <https://alexandruburlacu.github.io/posts/2021-07-26-ml-error-analysis>
- Bustikova, L., Siroky, D. S., Alashri, S., & Alzahrani, S. (2020). Predicting Partisan Responsiveness: A Probabilistic Text Mining Time-Series Approach. *Political Analysis*, 28(1), 47–64. <https://doi.org/10.1017/pan.2019.18>
- Cai, W., Encarnacion, R., Chern, B., Corbett-Davies, S., Bogen, M., Bergman, S., & Goel, S. (2022). *Adaptive Sampling Strategies to Construct Equitable Training Datasets* (No. arXiv:2202.01327). arXiv. <http://arxiv.org/abs/2202.01327>
- Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio books.
- Caprara, G. V., Barbaranelli, C., Consiglio, C., Picconi, L., & Zimbardo, P. G. (2003). Personalities of politicians and voters: Unique and synergistic relationships. *Journal of Personality and Social Psychology*, 84(4), 849. <https://doi.org/10.1037/0022-3514.84.4.849>
- Caprara, G. V., & Zimbardo, P. G. (2004). Personalizing politics: A congruency model of political preference. *American Psychologist*, 59(7), 581. <https://doi.org/10.1037/0003-066X.59.7.581>
- Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind. *Political Psychology*, 29(6), 807–840. <https://doi.org/10.1111/j.1467-9221.2008.00668.x>
- Castanho Silva, B., Jungkunz, S., Helbling, M., & Littvay, L. (2020). An Empirical Comparison of Seven Populist Attitudes Scales. *Political Research Quarterly*, 73(2), 409–424. <https://doi.org/10.1177/1065912919833176>

- Ceccarelli, L. (1998). Polysemy: Multiple meanings in rhetorical criticism. *Quarterly Journal of Speech*, 84(4), 395–415.
- Chaffee, S. H., & Metzger, M. J. (2001). The End of Mass Communication? *Mass Communication and Society*, 4(4), 365–379. https://doi.org/10.1207/S15327825MCS0404_3
- Chan, C., & Sältzer, M. (2020). oolong: An R package for validating automated content analysis tools. *Journal of Open Source Software*. <https://doi.org/10.21105/joss.02461>
- Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, 22. <https://proceedings.neurips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html>
- Chen, J., Qiu, L., & Ho, M.-H. R. (2020). A meta-analysis of linguistic markers of extraversion: Positive emotion and social process words. *Journal of Research in Personality*, 89, 104035. <https://doi.org/10.1016/j.jrp.2020.104035>
- Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, 11(4), 464–473. <https://doi.org/10.1177/1948550619875149>
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412.
- Cohen, J. (1960). *Chance, skill, and luck: The psychology of guessing and gambling*.
- Colliander, J., Marder, B., Falkman, L. L., Madestam, J., Modig, E., & Sagfossen, S. (2017). The social media balancing act: Testing the use of a balanced self-presentation strategy for politicians using twitter. *Computers in Human Behavior*, 74, 277–285.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
- Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Houghton Mifflin Boston.
- Costa Jr., P. T., & McCrae, R. R. (1995). Domains and Facets: Hierarchical Personality Assessment Using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, 64(1), 21–50. https://doi.org/10.1207/s15327752jpa6401_2
- Craig, R. T. (1999). Communication Theory as a Field. *Communication Theory*, 9(2), 119–161. <https://doi.org/10.1111/j.1468-2885.1999.tb00355.x>
- Cramer, K. J. (2016). *The politics of resentment: Rural consciousness in Wisconsin and the rise of Scott Walker*. University of Chicago Press.
- Cremaschi, S., Rettl, P., Cappelluti, M., & De Vries, C. E. (2024). Geographies of discontent: Public service deprivation and the rise of the far right in Italy. *American Journal of Political Science*, ajps.12936. <https://doi.org/10.1111/ajps.12936>

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281.
- Daikeler, J., Fröhling, L., Weller, K., Sen, I., Schwalbach, J., Birkenmaier, L., & Silber, H. (2023). Assessing Data Quality in the Age of Digital Social Research: A Systematic Review. *Preprint*. <https://osf.io/gfbjk/download>
- Das, A., & Rad, P. (2020). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey* (No. arXiv:2006.11371). arXiv. <http://arxiv.org/abs/2006.11371>
- de Leeuw, S. E., Azrout, R., Rekker, R. S., & Van Spanje, J. H. (2020). After all this time? The impact of media and authoritarian history on political news coverage in twelve western countries. *Journal of Communication*, 70(5), 744–767.
- de Vries, R. E., & van Prooijen, J.-W. (2019). Voters rating politicians' personality: Evaluative biases and assumed similarity on honesty-humility and openness to experience. *Personality and Individual Differences*, 144, 100–104.
- Decadri, S., & Boussalis, C. (2020). Populism, party membership, and language complexity in the Italian chamber of deputies. *Journal of Elections, Public Opinion and Parties*. <https://www.tandfonline.com/doi/full/10.1080/17457289.2019.1593182>
- DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., ... Pan, Z. (2024). *DeepSeek-V3 Technical Report* (No. arXiv:2412.19437). arXiv. <https://doi.org/10.48550/arXiv.2412.19437>
- Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(2), 168–189. <https://doi.org/10.1017/pan.2017.44>
- Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). *Assessing Language Model Deployment with Risk Cards* (No. arXiv:2303.18190). arXiv. <https://doi.org/10.48550/arXiv.2303.18190>
- 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>
- Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., & Johnson, B. K. (2021). An agenda for open science in communication. *Journal of Communication*, 71(1), 1–26.
- DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2), 205395171560290. <https://doi.org/10.1177/2053951715602908>
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606. <https://doi.org/10.1016/j.poetic.2013.08.004>

- Dobbrick, T., Jakob, J., Chan, C.-H., & Wessler, H. (2021). Enhancing Theory-Informed Dictionary Approaches with “Glass-box” Machine Learning: The Case of Integrative Complexity in Social Media Comments. *Communication Methods and Measures*, 1–18. <https://doi.org/10.1080/19312458.2021.1999913>
- Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). *Show Your Work: Improved Reporting of Experimental Results* (No. arXiv:1909.03004). arXiv. <http://arxiv.org/abs/1909.03004>
- Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (No. arXiv:1702.08608). arXiv. <https://doi.org/10.48550/arXiv.1702.08608>
- Dovring, K. (1954). Quantitative Semantics in 18th Century Sweden. *The Public Opinion Quarterly*, 18(4), 389–394.
- Drisko, J. W., & Maschi, T. (2016). *Content analysis*. Pocket Guide to Social Work Re.
- Drost, E. A. (n.d.). *Validity and Reliability in Social Science Research*.
- 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>
- 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*, 0(0), 1–23. <https://doi.org/10.1080/19312458.2023.2230560>
- El-Masri, M., Ramsay, A., Ahmed, H. M., & Ahmad, T. (2021). Positive sentiments as coping mechanisms and path to resilience: The case of Qatar blockade. *Information, Communication & Society*, 24(13), 1835–1853. <https://doi.org/10.1080/1369118X.2020.1748086>
- Eriksson, K. (2018). Republicans Value Agency, Democrats Value Communion. *Social Psychology Quarterly*, 81(2), 173–184. <https://doi.org/10.1177/0190272518771060>
- Fang, Q., Nguyen, D., & Oberski, D. L. (2022). Evaluating the Construct Validity of Text Embeddings with Application to Survey Questions. *arXiv:2202.09166 [Cs, Stat]*. <http://arxiv.org/abs/2202.09166>
- Federico, C. M., & Malka, A. (2018). The Contingent, Contextual Nature of the Relationship Between Needs for Security and Certainty and Political Preferences: Evidence and Implications. *Political Psychology*, 39(S1), 3–48. <https://doi.org/10.1111/pops.12477>
- Fenno, R. F. (2003). *Home style: House members in their districts*. Longman.
- Fernandes, J. M., Goplerud, M., & Won, M. (2019). Legislative Bellwethers: The Role of Committee Membership in Parliamentary Debate. *Legislative Studies Quarterly*, 44(2), 307–343. <https://doi.org/10.1111/lsq.12226>
- 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>

- Fleisig, E., Abebe, R., & Klein, D. (2024). *When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks* (No. arXiv:2305.06626). arXiv. <https://doi.org/10.48550/arXiv.2305.06626>
- Frenkel-Brunswik, E. (1949). Intolerance of ambiguity as an emotional and perceptual personality variable. *Journal of Personality*, 18(1). <https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=00223506&asa=N&AN=8930758&h=U%2BmDTvGB16tZA%2Bog1e1ksKp67T4i3T7NwEtNyciVN9HXIUureESaRyFV8Dbi8Yzrhz6feOQgxFf3ZlXV3gGIhA%3D%3D&crl=c>
- Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102(4), 652–670. <https://doi.org/10.1037/0033-295X.102.4.652>
- Funk, C. L. (1997). Implications of Political Expertise in Candidate Trait Evaluations. *Political Research Quarterly*, 50(3), 675–697. <https://doi.org/10.1177/106591299705000309>
- Garbe, L., Selvik, L.-M., & Lemaire, P. (2021). How African countries respond to fake news and hate speech. *Information, Communication & Society*, 1–18. <https://doi.org/10.1080/1369118X.2021.1994623>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). *Datasheets for Datasets* (No. arXiv:1803.09010). arXiv. <http://arxiv.org/abs/1803.09010>
- Gelman, A., & Loken, E. (2016). The statistical crisis in science. *The Best Writing on Mathematics* (Pitici M, Ed), 305–318.
- Gelman, J., & Wilson, S. L. (2022). Measuring Congressional Partisanship and Its Consequences. *Legislative Studies Quarterly*, 47(1), 225–256. <https://doi.org/10.1111/lsq.12331>
- George, A. L., & George, J. L. (1998). *Presidential personality and performance*. Boulder: WestviewPress.
- Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M., & Ha, S. E. (2010). Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review*, 104(1), 111–133.
- Geva, M., Goldberg, Y., & Berant, J. (2019). *Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets* (No. arXiv:1908.07898; Version 1). arXiv. <http://arxiv.org/abs/1908.07898>
- Gibney, E. (2022). Could machine learning fuel a reproducibility crisis in science? *Nature*, 608(7922), 250–251. <https://doi.org/10.1038/d41586-022-02035-w>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Gilardi, F., Gessler, T., Kubli, M., & Muller, S. (2022). Social Media and Political Agenda Setting. *Political Communication*, 39(1), 39–60. <https://doi.org/10.1080/10584609.2021.1910390>
- Goertz, G. (2008). *Concepts, theories, and numbers: A checklist for constructing, evaluating, and using concepts or quantitative measures*.

- Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, 27(4), 518–539.
- Goffman, E. (2016). The presentation of self in everyday life. In *Social Theory Re-Wired* (pp. 482–493). Routledge.
- Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”. *AI Magazine*, 38(3), Article 3. <https://doi.org/10.1609/aimag.v38i3.2741>
- Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, 10(10), 1645–1654. <https://doi.org/10.1111/2041-210X.13268>
- Greene, D., & Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, 25(1), 77–94. <https://doi.org/10.1017/pan.2016.7>
- 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.
- 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.
- Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology*. John Wiley & Sons.
- Gschwend, T., & Zittel, T. (2018). Who brings home the pork? Parties and the role of localness in committee assignments in mixed-member proportional systems. *Party Politics*, 24(5), 488–500. <https://doi.org/10.1177/1354068816678884>
- Gulati, G. J. (2004). Members of Congress and Presentation of Self on the World Wide Web. *Harvard International Journal of Press/Politics*, 9(1), 22–40. <https://doi.org/10.1177/1081180X03259758>
- Guo, Y., Guo, M., Su, J., Yang, Z., Zhu, M., Li, H., Qiu, M., & Liu, S. S. (2024). *Bias in Large Language Models: Origin, Evaluation, and Mitigation* (No. arXiv:2411.10915). arXiv. <https://doi.org/10.48550/arXiv.2411.10915>
- Haffert, L. (2021). Unequal Geographic Representation in a Mixed-Member Electoral System: Evidence from the German Bundestag. *German Politics*, 1–24. <https://doi.org/10.1080/09644008.2021.1982901>
- Hanania, R. (2017). The personalities of politicians: A big five survey of American legislators. *Personality and Individual Differences*, 108, 164–167.
- Harmening, M., Baumert, J., Block, S., Gross, M., Nyhuis, D., & Velimsky, J. (2025). Geographic representation in local politics: Evidence from parliamentary questions in German city councils. *European Journal of Political Research*, 64(4), 1999–2017. <https://doi.org/10.1111/1475-6765.70019>
- Hasson, Y., Tamir, M., Brahms, K. S., Cohrs, J. C., & Halperin, E. (2018). Are Liberals and Conservatives Equally Motivated to Feel Empathy Toward Others? *Personality and Social Psychology Bulletin*, 44(10), 1449–1459. <https://doi.org/10.1177/0146167218769867>
- Hayes, D. (2009). Has Television Personalized Voting Behavior? *Political Behavior*, 31(2), 231–260. <https://doi.org/10.1007/s11109-008-9070-0>

- Hays, D. (1960). *Automatic content analysis*. Rand Corporation.
- Hedrih, V. (2019). *Adapting Psychological Tests and Measurement Instruments for Cross-Cultural Research: An Introduction*. Routledge. <https://doi.org/10.4324/9780429264788>
- Heger, A. K., Marquis, L. B., Vorvoreanu, M., Wallach, H., & Wortman Vaughan, J. (2022). Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–29. <https://doi.org/10.1145/3555760>
- Helfferrich, C. (2022). Leitfaden-und experteninterviews. In *Handbuch Methoden der empirischen Sozialforschung* (pp. 875–892). Springer.
- Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., & Morris, C. (2014). Text to ideology or text to party status? *From Text to Political Positions: Text Analysis across Disciplines*, 55, 93–15.
- Ho, J. C., & Chan, C. (2023). Evaluating Transferability in Multilingual Text Analyses. *Computational Communication Research*, 5(2), Article 2. <https://doi.org/10.5117/CCR2023.2.2.HO>
- Hogan, R. (1982). A socioanalytic theory of personality. *Nebraska Symposium on Motivation*.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q. de, Gesmundo, A., Attariyan, M., & Gelly, S. (2019). *Parameter-Efficient Transfer Learning for NLP* (No. arXiv:1902.00751). arXiv. <https://doi.org/10.48550/arXiv.1902.00751>
- House, F. (2022). *Freedom in the world 2022* [Report]. Freedom House. <https://apo.org.au/node/316708>
- 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.
- Huang, F., Kwak, H., & An, J. (2023). Is chatgpt better than human annotators? Potential and limitations of chatgpt in explaining implicit hate speech. *arXiv Preprint arXiv:2302.07736*.
- Huang, L., Perry, P. O., & Spirling, A. (2020). A General Model of Author “Style” with Application to the UK House of Commons, 1935–2018. *Political Analysis*, 28(3), 412–434. <https://doi.org/10.1017/pan.2019.49>
- Huang, P., Lin, Z., Imbot, S., Fu, W., & Tu, E. (2025). *Analysis of LLM Bias (Chinese Propaganda & Anti-US Sentiment) in DeepSeek-R1 vs. ChatGPT o3-mini-high* (No. arXiv:2506.01814; Version 1). arXiv. <https://doi.org/10.48550/arXiv.2506.01814>
- Hughes, D. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In *The Wiley handbook of psychometric testing: A multidisciplinary approach to survey, scale and test development*. John Wiley & Sons Ltd.
- Humphreys, M., De la Sierra, R. S., & Van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1), 1–20.

- Hyatt, C., Campbell, W. K., Lynam, D. R., & Miller, J. D. (2018). Dr. Jekyll or Mr. Hyde? President Donald Trump's personality profile as perceived from different political viewpoints. *Collabra: Psychology*, 4(1), 29.
- Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, 32.
- Immelman, A. (1993). The assessment of political personality: A psychodiagnostically relevant conceptualization and methodology. *Political Psychology*, 725–741.
- Jankowski, M., & Huber, R. A. (2022). When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models. *Political Analysis*, 1–15.
- Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. *Proceedings of the Second Workshop on NLP and Computational Social Science*, 7–16. <https://aclanthology.org/W17-2902/>
- Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2022). Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, 75, 102420. <https://doi.org/10.1016/j.electstud.2021.102420>
- Joly, J., Soroka, S., & Loewen, P. (2019). Nice guys finish last: Personality and political success. *Acta Politica*, 54(4), 667–683. <https://doi.org/10.1057/s41269-018-0095-z>
- Kajs, L. T., & McCollum, D. L. (2009). Examining tolerance for ambiguity in the domain of educational leadership. *Academy of Educational Leadership Journal*, 13(2), 1–16.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.
- Kantner, C., & Overbeck, M. (2020). Exploring soft concepts with hard corpus-analytic methods. *Reflektierte Algorithmische Textanalyse. Interdisziplinäre (s) Arbeiten in Der CRETA-Werkstatt*, 169–189.
- Kapoor, S., & Narayanan, A. (2022). *Leakage and the Reproducibility Crisis in ML-based Science* (No. arXiv:2207.07048). arXiv. <https://doi.org/10.48550/arXiv.2207.07048>
- Kelley, T. L. (1927). *Interpretation of educational measurements* (p. 353). World Book Co.
- Kim, I. S., Londregan, J., & Ratkovic, M. (2018). Estimating Spatial Preferences from Votes and Text. *Political Analysis*, 26(2), 210–229. <https://doi.org/10.1017/pan.2018.7>
- Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*, 71(6), 922–946. <https://doi.org/10.1093/joc/jqab034>
- Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 65(23), 2276–2284.

- King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry*. Princeton university press.
- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications.
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 1–26.
- Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). *ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification* (No. arXiv:2303.03953). arXiv. <https://doi.org/10.48550/arXiv.2303.03953>
- Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and Best Practices in Content Analysis. *Journalism & Mass Communication Quarterly*, 92(4), 791–811. <https://doi.org/10.1177/1077699015607338>
- Lauderdale, B. E., & Herzog, A. (2016). Measuring Political Positions from Legislative Speech. *Political Analysis*, 24(3), 374–394. <https://doi.org/10.1093/pan/mpw017>
- Laustsen, L., & Bor, A. (2017). The relative weight of character traits in political candidate evaluations: Warmth is more important than competence, leadership and integrity. *Electoral Studies*, 49, 96–107.
- Lee, J., & Lim, Y.-S. (2016). Gendered campaign tweets: The cases of Hillary Clinton and Donald Trump. *PUBLIC RELATIONS REVIEW*, 42(5), 849–855. <https://doi.org/10.1016/j.pubrev.2016.07.004>
- Lehmann, P., Burst, T., Matthieß, T., Regel, S., Volkens, A., Weßels, B., Zehnter, L., & Wissenschaftszentrum Berlin Für Sozialforschung (WZB). (2022). *Manifesto Project Dataset* (Version 2022a) [Data set]. Manifesto Project. <https://doi.org/10.25522/MANIFESTO.MPDS.2022A>
- Lehmann, P., Franzmann, S., Al-Gaddooa, D., Burst, T., Ivanusch, C., Regel, S., Riethmüller, F., Volkens, A., Weßels, B., Zehnter, L., Wissenschaftszentrum Berlin Für Sozialforschung (WZB), & Institut Für Demokratieforschung Göttingen (IfDem). (2024). *Manifesto Project Dataset* (Version 2024a) [Data set]. Manifesto Project. <https://doi.org/10.25522/MANIFESTO.MPDS.2024A>
- Lenzner, T., Hadler, P., & Neuert, C. (2023). An experimental test of the effectiveness of cognitive interviewing in pretesting questionnaires. *Quality & Quantity*, 57(4), 3199–3217. <https://doi.org/10.1007/s11135-022-01489-4>
- Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, 57(1), 34–52. <https://doi.org/10.1080/08838151.2012.761702>
- Lewis-Beck, M., Bryman, A. E., & Liao, T. F. (2003). *The Sage encyclopedia of social science research methods*. Sage Publications.
- 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>
- Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, 31(3), 366–379.
- Licht, H., Abou-Chadi, T., Barberá, P., & Hua, W. (2024). Measuring and Understanding Parties' Anti-elite Strategies. *The Journal of Politics*, 000–000. <https://doi.org/10.1086/730711>
- Licht, H., & Lind, F. (2023). Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research*, 5(2), Article 2. <https://doi.org/10.5117/CCR2023.2.2.LICH>
- Lilienfeld, S. O., Waldman, I. D., Landfield, K., Watts, A. L., Rubenzer, S., & Faschingbauer, T. R. (2012). Fearless dominance and the US presidency: Implications of psychopathic personality traits for successful and unsuccessful political leadership. *Journal of Personality and Social Psychology*, 103(3), 489.
- Lin, R., & Ng, H. T. (2023). Mind the Biases: Quantifying Cognitive Biases in Language Model Prompting. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 5269–5281). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.324>
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1), 18. <https://doi.org/10.3390/e23010018>
- Lind, F., Schoonvelde, M., Baden, C., Dolinsky, A., Pipal, C., Shababo, G., & van der Velden, M. (2023). *A framework for validating multilingual computational text analysis*.
- Lipton, Z. C. (2017). *The Mythos of Model Interpretability* (No. arXiv:1606.03490). arXiv. <http://arxiv.org/abs/1606.03490>
- Liu, H., Thekinen, J., Mollaoglu, S., Tang, D., Yang, J., Cheng, Y., Liu, H., & Tang, J. (2021). *Toward Annotator Group Bias in Crowdsourcing* (No. arXiv:2110.08038; Version 1). arXiv. <https://doi.org/10.48550/arXiv.2110.08038>
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635–694.
- 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.
- Macanovic, A. (2022). Text mining for social science – The state and the future of computational text analysis in sociology. *Social Science Research*, 108, 102784. <https://doi.org/10.1016/j.ssresearch.2022.102784>
- Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2022). Machine Translation Vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections. *Communication Methods and Measures*, 16(1), 19–38. <https://doi.org/10.1080/19312458.2021.1955845>
- Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., & Häussler, T. (2021). Applying LDA topic modeling in

- communication research: Toward a valid and reliable methodology. In *Computational methods for communication science* (pp. 13–38). Routledge.
- Maier, J., Oschatz, Corinna, Stier, Sebastian, & Zettler, I. (2023). A Short Scale to Measure Self-Reported Aversive Personality in Political Elites. *Journal of Personality Assessment*, 105(5), 625–635. <https://doi.org/10.1080/00223891.2022.2130341>
- Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30, 457–500. <https://doi.org/10.1613/jair.2349>
- Malik, M., Youk, S., Hopp, F. R., Curry, O. S., Cheong, M., Alfano, M., & Weber, R. (2025). The Extended Morality as Cooperation Dictionary (eMACD): A Crowd-Sourced Approach via the Moral Narrative Analyzer Platform. *Communication Methods and Measures*. <https://www.tandfonline.com/doi/abs/10.1080/19312458.2025.2500329>
- Mayring, P. (2004). Qualitative content analysis. *A Companion to Qualitative Research*, 1(2), 159–176.
- McAllister, I. (2007). The Personalization of Politics. In R. J. Dalton & H. Klingemann (Eds), *The Oxford Handbook of Political Behavior* (p. o). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199270125.003.0030>
- McCue, C. P., & Gopoian, J. D. (2000). Dispositional Empathy and the Political Gender Gap. *Women & Politics*, 21(2), 1–20. https://doi.org/10.1300/J014v21n02_01
- McDonald, J. D. (2008). Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire*, 1(1), 1–19.
- McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods*, 16, 152–184. <https://doi.org/10.1177/1094428112459910>
- Meeks, L. (2016). Aligning and Trespassing: Candidates’ Party-Based Issue and Trait Ownership on Twitter. *Journalism & Mass Communication Quarterly*, 93(4), 1050–1072. <https://doi.org/10.1177/1077699015609284>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
- Messick, S. (1989). Validity. In *Educational measurement*, 3rd ed (pp. 13–103). American Council on Education.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). arXiv. <https://doi.org/10.48550/arXiv.1301.3781>
- Miller, A. H., Wattenberg, M. P., & Malanchuk, O. (1986). Schematic assessments of presidential candidates. *American Political Science Review*, 80(2), 521–540.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

- Mondak, J. J. (1995). Competence, Integrity, and the Electoral Success of Congressional Incumbents. *The Journal of Politics*, 57(4), 1043–1069. <https://doi.org/10.2307/2960401>
- Mueller, S. D., & Saeltzer, M. (2020). Twitter made me do it! Twitter's tonal platform incentive and its effect on online campaigning. *Information, Communication & Society*, 1–26. <https://doi.org/10.1080/1369118X.2020.1850841>
- Nai, A. (2019). The electoral success of angels and demons: Big five, dark triad, and performance at the ballot box. *Journal of Social and Political Psychology*, 7(2), 830–862.
- Nai, A., Maier, J., & Vranić, J. (2021). Personality Goes a Long Way (for Some). An Experimental Investigation Into Candidate Personality Traits, Voters' Profile, and Perceived Likeability. *Frontiers in Political Science*, 3. <https://doi.org/10.3389/fpos.2021.636745>
- Nai, A., Tresch, A., & Maier, J. (2022). Tailored negativity. Campaign consultants, candidate personality, and attack politics. *Swiss Political Science Review*, 28(2), 338–360. <https://doi.org/10.1111/spsr.12525>
- Neuendorf, K. A. (2017). *The content analysis guidebook*. sage.
- Ni, J., Fan, Y., Zouhar, V., Rooein, D., Hoyle, A., Sachan, M., Leippold, M., Hovy, D., & Ash, E. (2025). *Can Reasoning Help Large Language Models Capture Human Annotator Disagreement?* (No. arXiv:2506.19467). arXiv. <https://doi.org/10.48550/arXiv.2506.19467>
- Nie, Y., Zhou, X., & Bansal, M. (2020). *What Can We Learn from Collective Human Opinions on Natural Language Inference Data?* (No. arXiv:2010.03532). arXiv. <https://doi.org/10.48550/arXiv.2010.03532>
- Norris, P. (2004). *Electoral engineering: Voting rules and political behavior*. Cambridge university press.
- Nyholt, N. (2024). Left behind: Voters' reactions to local school and hospital closures. *European Journal of Political Research*, 63(3), 884–905. <https://doi.org/10.1111/1475-6765.12622>
- Nyhuis, D., Block, S., Harmening, M., & Henneke, M. (2025). Geographic representation in parliamentary questions: The benefits of Wikipedia for automated analyses of legislative text. *The Journal of Legislative Studies*, 1–13. <https://doi.org/10.1080/13572334.2025.2518633>
- Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., & Oliveira, E. L. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 15.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1–135.
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934–952. <https://doi.org/10.1037/pspp0000020>

- Park, J. Y., & Montgomery, J. M. (2023). *Validating the text-to-measure pipeline: A procedure-based approach to creating measures of latent concepts with supervised machine learning* [Working Paper].
- Parthasarathy, R., Rao, V., & Palaniswamy, N. (2019). Deliberative democracy in an unequal world: A text-as-data study of south India's village assemblies. *American Political Science Review*, 113(3), 623–640.
- Partsch, M. V., Bluemke, M., & Lechner, C. M. (2022). Revisiting the hierarchical structure of the 24 VIA character strengths: Three global dimensions may suffice to capture their essence. *European Journal of Personality*, 36(5), 825–845. <https://doi.org/10.1177/08902070211017760>
- Patzelt, W. J. (2013). German MPs and their roles. In *Members of Parliament in Western Europe* (pp. 55–78). Routledge. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315036847-4/german-mps-roles-werner-patzelt>
- Pauwels, T. (2017). CHAPTER 6: MEASURING POPULISM: A REVIEW OF CURRENT APPROACHES. In R. C. Heinisch, C. Holtz-Bacha, & O. Mazzoleni (Eds), *Political Populism* (pp. 123–136). Nomos Verlagsgesellschaft mbH & Co. KG. <https://doi.org/10.5771/9783845271491-123>
- Pebesma, E., & Bivand, R. (2023). *Spatial data science: With applications in R*. Chapman and Hall/CRC. <https://www.taylorfrancis.com/books/mono/10.1201/9780429459016/spatial-data-science-edzer-pebesma-roger-bivand>
- Pedersen, H. H., & Rahat, G. (2021). Political personalization and personalized politics within and beyond the behavioural arena. *Party Politics*, 27(2), 211–219. <https://doi.org/10.1177/1354068819855712>
- Perloff, R. M. (2013). *Political Communication: Politics, Press, and Public in America*. Routledge. <https://doi.org/10.4324/9780203811832>
- Petersen, M. B., & Laustsen, L. (2020). Dominant leaders and the political psychology of followership. *Current Opinion in Psychology*, 33, 136–141.
- Peterson, A. J., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, 26(01), 120. <https://doi.org/10.1017/pan.2017.39>
- Pillai, R., Williams, E. A., Lowe, K. B., & Jung, D. I. (2003). Personality, transformational leadership, trust, and the 2000 U.S. presidential vote. *The Leadership Quarterly*, 14(2), 161–192. [https://doi.org/10.1016/S1048-9843\(03\)00008-0](https://doi.org/10.1016/S1048-9843(03)00008-0)
- Plank, B. (2022). *The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation* (No. arXiv:2211.02570). arXiv. <https://doi.org/10.48550/arXiv.2211.02570>
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for Creating Better Concept Definitions in the Organizational, Behavioral, and Social Sciences. *Organizational Research Methods*, 19(2), 159–203. <https://doi.org/10.1177/1094428115624965>
- Pradel, F. (2021). Biased Representation of Politicians in Google and Wikipedia Search? The Joint Effect of Party Identity, Gender Identity and Elections.

- Political Communication*, 38(4), 447–478.
<https://doi.org/10.1080/10584609.2020.1793846>
- Pratas, M., Fernandes, J. M., & Won, M. (2025). Inequalities in place-based representation: Looking inside electoral districts. *European Journal of Political Research*, 64(3), 1165–1188. <https://doi.org/10.1111/1475-6765.12735>
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespín, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209–228.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (n.d.). *Robust Speech Recognition via Large-Scale Weak Supervision*.
- Rahat, G., & Kenig, O. (2018). *From Party Politics to Personalized Politics?: Party Change and Political Personalization in Democracies*. Oxford University Press.
- Ramey, A. J., Klingler, J. D., & Hollibaugh, G. E. (2019). Measuring elite personality using speech. *Political Science Research and Methods*, 7(1), 163–184.
- Rammstedt, B., Beierlein, C., Brähler, E., Eid, M., Hartig, J., Kersting, M., Liebig, S., Lukas, J., Mayer, A.-K., & Menold, N. (2015). Quality standards for the development, application, and evaluation of measurement instruments in social science survey research. *RatSWD Working Paper Series*, 245.
- Rauchfleisch, A., & Kaiser, J. (2020). The False positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10), e0241045. <https://doi.org/10.1371/journal.pone.0241045>
- Rauh, C. (2018). Validating a sentiment dictionary for German political language—A workbench note. *Journal of Information Technology & Politics*, 15(4), 319–343. <https://doi.org/10.1080/19331681.2018.1485608>
- Rauh, C., & Schwalbach, J. (2020). *The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies* [Data set]. Harvard Dataverse. <https://doi.org/10.7910/DVN/L4OAKN>
- Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14–46.
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154.
- Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark* (No. arXiv:2304.11085). arXiv. <https://doi.org/10.48550/arXiv.2304.11085>
- Repke, L., Birkenmaier, L., & Lechner, C. (2024). *Validity in Survey Research-From Research Design to Measurement Instruments*. <https://www.ssoar.info/ssoar/handle/document/95646>
- Rheault, L., & Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, 28(1), 112–133. <https://doi.org/10.1017/pan.2019.26>
- Ribeiro, F., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1), 1–29.

- Roberts, C. W. (1989). Other than counting words: A linguistic approach to content analysis. *Social Forces*, 68(1), 147–177.
- Robinson, O. C. (2014). Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative Research in Psychology*, 11(1), 25–41.
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. <https://doi.org/10.1145/2684822.2685324>
- Rodríguez-Pose, A. (2018). The revenge of the places that don't matter (and what to do about it). *Cambridge Journal of Regions, Economy and Society*, 11(1), 189–209. <https://doi.org/10.1093/cjres/rsx024>
- Rogers, A., Baldwin, T., & Leins, K. (2021). `Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4821–4833. <https://doi.org/10.18653/v1/2021.findings-emnlp.414>
- Rossiter, E. L. (2022). Measuring Agenda Setting in Interactive Political Communication. *American Journal of Political Science*, 66(2), 337–351. <https://doi.org/10.1111/ajps.12653>
- Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2022). *Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks* (No. arXiv:2112.07475). arXiv. <https://doi.org/10.48550/arXiv.2112.07475>
- Rubenzler, S. J., Faschingbauer, T. R., & Ones, D. S. (2000). Assessing the U.S. Presidents Using the Revised NEO Personality Inventory. *Assessment*, 7(4), 403–419. <https://doi.org/10.1177/107319110000700408>
- Ruch, G. M. (1924). *The improvement of the written examination*. Scott, Foresman.
- Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, 12(2–3), 140–157. <https://doi.org/10.1080/19312458.2018.1455817>
- Rusticus, S. (2014). Content Validity. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 1261–1262). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_553
- Sältzer, M., Stier, S., Bäuerle, J., Blumenberg, M., Mechkova, V., Pemstein, D., Seim, B., & Wilson, S. (2023). *Twitter accounts of candidates in the German federal election 2021 (GLES) Twitter-Accounts der Kandidierenden zur Bundestagswahl 2021 (GLES)* (Version 2.1.0) [Data set]. GESIS. <https://doi.org/10.4232/1.14233>
- Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021). “Call me sexist, but...”: Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 573–584. <https://doi.org/10.1609/icwsm.v15i1.18085>
- Sang, E. F. T. K., & Meulder, F. D. (2003). *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition* (No. arXiv:cs/0306050). arXiv. <https://doi.org/10.48550/arXiv.cs/0306050>

- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *International Conference on Machine Learning*, 29971–30004. <https://proceedings.mlr.press/v202/santurkar23a.html>
- Saucier, G., Thalmayer, A. G., Payne, D. L., Carlson, R., Sanogo, L., Ole-Kotikash, L., Church, A. T., Katigbak, M. S., Somer, O., Szarota, P., Szirmák, Z., & Zhou, X. (2014). A Basic Bivariate Structure of Personality Attributes Evident Across Nine Languages. *Journal of Personality*, 82(1), 1–14. <https://doi.org/10.1111/jopy.12028>
- Schmahl, K. G., Viering, T. J., Makrodimitis, S., Jahfari, A. N., Tax, D., & Loog, M. (2020). Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 94–103. <https://aclanthology.org/2020.nlpccs-1.11/>
- Schmidgall, S., Harris, C., Essien, I., Olshvang, D., Rahman, T., Kim, J. W., Ziaei, R., Eshraghian, J., Abadir, P., & Chellappa, R. (2024). *Addressing cognitive bias in medical language models* (No. arXiv:2402.08113). arXiv. <https://doi.org/10.48550/arXiv.2402.08113>
- Schoch, D., Chan, C., Wagner, C., & Bleier, A. (2023). *Computational Reproducibility in Computational Social Science* (No. arXiv:2307.01918). arXiv. <https://doi.org/10.48550/arXiv.2307.01918>
- Schoen, H., & Schumann, S. (2005). Missing Links? Zur Position von Persönlichkeitsfaktoren in Ansätzen zur Erklärung von Wahlverhalten. In J. W. Falter, O. W. Gabriel, & B. Weßels (Eds), *Wahlen und Wähler* (pp. 388–413). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-80516-4_16
- Schoonvelde, M., Schumacher, G., & Bakker, B. N. (2019). Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology. *Journal of Social and Political Psychology*, 7(1), Article 1. <https://doi.org/10.5964/jspp.v7i1.964>
- Schub, R. (2022). Informing the Leader: Bureaucracies and International Crises. *American Political Science Review*, 1–17. <https://doi.org/10.1017/S0003055422000168>
- Schumacher, G., & Zettler, I. (2019). House of Cards or West Wing? Self-reported HEXACO Traits of Danish Politicians. *Personality and Individual Differences*, 141, 173–181. <https://doi.org/10.1016/j.paid.2018.12.028>
- Schürmann, L., & Stier, S. (2022). Who Represents the Constituency? Online Political Communication by Members of Parliament in the German Mixed-Member Electoral System. *Legislative Studies Quarterly*, lsq.12379. <https://doi.org/10.1111/lsq.12379>
- Scimago. (2022, October 3). *SJR - About Us*. <https://www.scimagojr.com/aboutus.php>
- Sciar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting* (No. arXiv:2310.11324). arXiv. <http://arxiv.org/abs/2310.11324>

- Scott, C., & Medeiros, M. (2020). Personality and political careers: What personality types are likely to run for office and get elected? *Personality and Individual Differences*, 152, 109600. <https://doi.org/10.1016/j.paid.2019.109600>
- Sebeok, T. A., & Zeps, V. J. (1958). An Analysis of Structured Content, with Application of Electronic Computer Research, in Psycholinguistics. *Language and Speech*, 1(3), 181–193. <https://doi.org/10.1177/002383095800100304>
- Sen, I., Samory, M., Wagner, C., & Augenstein, I. (2022). Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4716–4726). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.347>
- Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A., & Arawjo, I. (2024). Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–14. <https://doi.org/10.1145/3654777.3676450>
- Shermer, M. (2008). Patternicity: Finding meaningful patterns in meaningless noise. *Scientific American*, 299(5), 48.
- Shugart, M. S. (Ed.). (2005). *Mixed-member electoral systems: The best of both worlds?* (Repr). Oxford University Press.
- Smrek, M. (2023). Mavericks or Loyalists? Popular Ballot Jumpers and Party Discipline in the Flexible-List PR Context. *Political Research Quarterly*, 76(1), 323–336. <https://doi.org/10.1177/10659129221087961>
- Song, H., Eberl, J.-M., & Eisele, O. (2020). Less fragmented than we thought? Toward clarification of a subdisciplinary linkage in communication science, 2010–2019. *Journal of Communication*, 70(3), 310–334.
- Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020a). 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>
- Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020b). 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>
- Speed, J. G. (1893). Do newspapers now give the news. *Forum*, 15, 705–711.
- Sprong, S., Jetten, J., Wang, Z., Peters, K., Mols, F., Verkuyten, M., Bastian, B., Ariyanto, A., Autin, F., Ayub, N., Badea, C., Besta, T., Butera, F., Costa-Lopes, R., Cui, L., Fantini, C., Finchilescu, G., Gaertner, L., Gollwitzer, M., ... Wohl, M. J. A. (2019). “Our Country Needs a Strong Leader Right Now”: Economic Inequality Enhances the Wish for a Strong Leader. *Psychological Science*, 30(11), 1625–1637. <https://doi.org/10.1177/0956797619875472>

- Stanyer, J. (2008). Elected representatives, online self-presentation and the personal vote: Party, personality and webstyles in the United States and United Kingdom. *Information, Communication & Society*, 11(3), 414–432. <https://doi.org/10.1080/13691180802025681>
- Stemler, S. (2000). An overview of content analysis. *Practical Assessment, Research, and Evaluation*, 7(1), 17.
- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684), 677–680. <https://doi.org/10.1126/science.103.2684.677>
- Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018a). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication*, 35(1), 50–74. <https://doi.org/10.1080/10584609.2017.1334728>
- Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018b). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication*, 35(1), 50–74. <https://doi.org/10.1080/10584609.2017.1334728>
- Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2020). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. In *Studying politics across media* (pp. 50–74). Routledge. <https://www.taylorfrancis.com/chapters/edit/10.4324/9780429202483-4/election-campaigning-social-media-politicians-audiences-mediation-political-communication-facebook-twitter-sebastian-stier-arnim-bleier-haiko-lietz-markus-strohmaier>
- Stone, P., Dunphy, D., Smith, M., & Ogilvie, D. (1966). The General Inquirer: A Computer Approach to Content Analysis. In *American Educational Research Journal—AMER EDUC RES J* (Vol. 4). <https://doi.org/10.2307/1161774>
- Stroppe, A.-K. (2023). Left behind in a public services wasteland? On the accessibility of public services and political trust. *Political Geography*, 105, 102905.
- Swamy, V., Frej, J., & Käser, T. (2025). The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations. *Journal of Artificial Intelligence Research*, 84. <https://www.jair.org/index.php/jair/article/view/17970>
- Tackett, J. L., Smack, A. J., Herzhoff, K., Reardon, K. W., Daoud, S., & Granic, I. (2017). Measuring child personality when child personality was not measured: Application of a thin-slice approach. *Personality and Mental Health*, 11(1), 4–13. <https://doi.org/10.1002/pmh.1351>
- Talat, Z., Névél, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., & Radev, D. (2022). You reap what you sow: On the challenges of bias evaluation under multilingual settings. *Proceedings of BigScience Episode# 5—Workshop on Challenges & Perspectives in Creating Large Language Models*, 26–41. <https://aclanthology.org/2022.bigscience-1.3/>
- Temporão, M., Vande Kerckhove, C., van der Linden, C., Dufresne, Y., & Hendrickx, J. M. (2018). Ideological Scaling of Social Media Users: A Dynamic Lexicon Approach. *Political Analysis*, 26(4), 457–473. <https://doi.org/10.1017/pan.2018.30>

- Tenenboim-Weinblatt, K., & Lee, C. (2020). Speaking across communication subfields. *Journal of Communication*, 70(3), 303–309.
- Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). OCTIS: Comparing and Optimizing Topic models is Simple! *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263–270. <https://doi.org/10.18653/v1/2021.eacl-demos.31>
- Theocharis, Y., & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication. *Political Communication*, 38(1–2), 1–22. <https://doi.org/10.1080/10584609.2020.1833121>
- Thielmann, I., & Hilbig, B. E. (2019). Nomological consistency: A comprehensive test of the equivalence of different trait indicators for the same constructs. *Journal of Personality*, 87(3), 715–730. <https://doi.org/10.1111/jopy.12428>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (No. arXiv:2302.13971). arXiv. <https://doi.org/10.48550/arXiv.2302.13971>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (No. arXiv:2307.09288). arXiv. <https://doi.org/10.48550/arXiv.2307.09288>
- Trad, F., & Chehab, A. (2024). Prompt Engineering or Fine-Tuning? A Case Study on Phishing Detection with Large Language Models. *Machine Learning and Knowledge Extraction*, 6(1), 367–384. <https://doi.org/10.3390/make6010018>
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72, 1385–1470. <https://doi.org/10.1613/jair.1.12752>
- Van Aelst, P., Sheafer, T., & Stanyer, J. (2012). The personalization of mediated political communication: A review of concepts, operationalizations and key findings. *Journalism*, 13(2), 203–220. <https://doi.org/10.1177/1464884911427802>
- 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.
- Van Der Velden, M., Schumacher, G., & Vis, B. (2018). Living in the Past or Living in the Future? Analyzing Parties' Platform Change In Between Elections, The Netherlands 1997–2014. *Political Communication*, 35(3), 393–412. <https://doi.org/10.1080/10584609.2017.1384771>
- van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation

- methods. *Journal of Business Research*, 144, 93–106.
<https://doi.org/10.1016/j.jbusres.2022.01.076>
- Van Zoonen, L. (2005). *Entertaining the citizen: When politics and popular culture converge*. Rowman & Littlefield.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Łukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.
<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Visser, B. A., Book, A. S., & Volk, A. A. (2017). Is Hillary dishonest and Donald narcissistic? A HEXACO analysis of the presidential candidates' public personas. *Personality and Individual Differences*, 106, 281–286.
- Wadhwa, S., Chandu, K. R., & Nyberg, E. (2018). Comparative analysis of neural QA models on SQuAD. *arXiv Preprint arXiv:1806.06972*.
- Waisbord, S. (2019). *Communication: A post-discipline*. John Wiley & Sons.
- Walker, S. G., Schafer, M., & Young, M. D. (1998). Systematic Procedures for Operational Code Analysis: Measuring and Modeling Jimmy Carter's Operational Code. *International Studies Quarterly*, 42(1), 175–189.
<https://doi.org/10.1111/0020-8833.00074>
- Wankmüller, S. (2022). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis. *Sociological Methods & Research*, 00491241221134527. <https://doi.org/10.1177/00491241221134527>
- Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., & Poli, I. (2024). *Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference* (No. arXiv:2412.13663). arXiv.
<https://doi.org/10.48550/arXiv.2412.13663>
- Watts, A. L., Lilienfeld, S. O., Smith, S. F., Miller, J. D., Campbell, W. K., Waldman, I. D., Rubenzer, S. J., & Faschingbauer, T. J. (2013). The Double-Edged Sword of Grandiose Narcissism: Implications for Successful and Unsuccessful Leadership Among U.S. Presidents. *Psychological Science*, 24(12), 2379–2389.
<https://doi.org/10.1177/0956797613491970>
- Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018). Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions. *Communication Methods and Measures*, 12(2–3), 119–139.
<https://doi.org/10.1080/19312458.2018.1447656>
- Widmann, T., & Wich, M. (2022). Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text. *Political Analysis*, 1–16.
<https://doi.org/10.1017/pan.2022.15>
- Winter, D. G., Hermann, M. G., Weintraub, W., & Walker, S. G. (1991). The Personalities of Bush and Gorbachev Measured at a Distance: Procedures, Portraits, and Policy. *Political Psychology*, 12(2), 215–245.
<https://doi.org/10.2307/3791463>

- Woolley, J. T., & Peters, G. (1999). The American presidency project. Santa Barbara, CA. Available from World Wide Web: [Http://Www. Presidency. Ucsb. Edu/Ws](http://www.presidency.ucsb.edu/Ws). <https://www.jfkthelastspeech.org/wp-content/uploads/2023/11/JFK-Speech-at-San-Diego-State.pdf>
- Wright, B. D. (1997). A History of Social Science Measurement. *Educational Measurement: Issues and Practice*, 16(4), 33–45. <https://doi.org/10.1111/j.1745-3992.1997.tb00606.x>
- Wright, J. D., & Tomlinson, M. F. (2018). Personality profiles of Hillary Clinton and Donald Trump: Fooled by your own politics. *Personality and Individual Differences*, 128, 21–24.
- Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. (2019). Errudite: Scalable, Reproducible, and Testable Error Analysis. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 747–763. <https://doi.org/10.18653/v1/P19-1073>
- Wurthmann, L. C., & Sältzer, M. (2025). Issue congruence between candidates' Twitter communication and constituencies in an MMES: Migration as an exemplary case. *Parliamentary Affairs*, gsafo33. <https://doi.org/10.1093/pa/gsafo33>
- Wuttke, A., Schimpf, C., & Schoen, H. (2020). When the Whole Is Greater than the Sum of Its Parts: On the Conceptualization and Measurement of Populist Attitudes and Other Multidimensional Constructs. *American Political Science Review*, 114(2), 356–374. <https://doi.org/10.1017/S0003055419000807>
- Xiao, Y., & Xiao, L. (2020). Effects of anonymity on comment persuasiveness in wikipedia articles for deletion discussions. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 104–115. <https://aclanthology.org/2020.nlpcss-1.12/>
- Yang, T., & Fang, K. (2021). How dark corners collude: A study on an online Chinese alt-right community. *Information, Communication & Society*, 0(0), 1–18. <https://doi.org/10.1080/1369118X.2021.1954230>
- Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, 38(1–2), 98–139. <https://doi.org/10.1080/10584609.2020.1785067>
- 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>
- Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds), *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. 3914–3923). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1404>
- Ying, L., Montgomery, J. M., & Stewart, B. M. (2022). Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. *Political Analysis*, 30(4), 570–589.

- Zehnter, L. (2024). Measuring Populism Through Text Analysis. *Wer Ist Das Volk? Analyzing and Understanding Populist Political Communication*, 43.
- Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The Nomological Net of the HEXACO Model of Personality: A Large-Scale Meta-Analytic Investigation. *Perspectives on Psychological Science*, 15(3), 723–760. <https://doi.org/10.1177/1745691619895036>
- Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *International Conference on Machine Learning*, 12697–12706. <http://proceedings.mlr.press/v139/zhao21c.html>
- Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., & Han, J. (2023). *Don't Make Your LLM an Evaluation Benchmark Cheater* (No. arXiv:2311.01964). arXiv. <https://doi.org/10.48550/arXiv.2311.01964>
- Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237–291.
- Zittel, T., Nyhuis, D., & Baumann, M. (2019). Geographic Representation in Party-Dominated Legislatures: A Quantitative Text Analysis of Parliamentary Questions in the German Bundestag. *Legislative Studies Quarterly*, 44(4), 681–711. <https://doi.org/10.1111/lsq.12238>