Title: "The search for solid ground in text as data: A systematic review of validation practices and practical recommendations for validation"

Authors

1. Lukas Birkenmaier (corresponding author)[1], GESIS – Leibniz Institute for the Social Sciences, Mannheim, https://orcid.org/0009-0006-1538-0589 ,

2. Dr. Clemens M. Lechner, GESIS – Leibniz Institute for the Social Sciences, Mannheim, https://orcid.org/0000-0003-3053-8701

3. Prof. Dr. Claudia Wagner, GESIS – Leibniz Institute for the Social Sciences, Mannheim; RWTH Aachen University, Germany; Complexity Science Hub Vienna, Austria

Word Count: 11.073 (excluding references)

Word Count: 15.189 (including references)

[1] Lukas Birkenmaier, lukas.birkenmaier@gesis.org, GESIS – Leibniz Institute for the Social Sciences B6 4, 5, 68159 Mannheim, Germany

Abstract

Communication research frequently applies computational text analysis methods (CTAM) to detect and measure social science constructs. However, the validity of these measures can be difficult to assess. In addition, there are hardly any established standards and little guidance for researchers on how to best validate CTAM. But how do these challenges affect current validation practices in applied research? And what practical recommendations for better validation of text-based measures can we derive? To answer these questions, we conducted a systematic review of current validation practices and qualitative expert interviews. We focused on political communication, a subfield that has arguably played a pioneering role in embracing the application of CTAM in communication research. Our results show that researchers apply a great variety of validation steps, which, however, are rarely selected based on a unified understanding of validity. The qualitative interviews further reinforce this notion, as interviewees bemoan a lack of established guidelines and frameworks for validation. Based on our empirical findings, we therefore derive practical recommendations to guide researchers regarding when and how to validate CTAM. Moreover, we provide a preview of emerging validation frameworks that could prove beneficial for researchers working with text as data.

Keywords: Text as Data, Validity, Measurement, Social Science Constructs, Systematic Review, Qualitative Expert Interviews

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 analyse 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 text-based 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. In an effort 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 working with text as data undertake during validation. In the normative part, based on this empirical overview, we then derive practical recommendations and provide guidance on how researchers should approach validation. Additionally, we offer a glimpse into present and upcoming validation frameworks that may prove advantageous for researchers seeking guidance in different research contexts. Our intention is to furnish researchers with actionable advice while also inspiring future methodological advancements in the realm of validity and validation.

2. Theoretical Background

On the 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 1). 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).

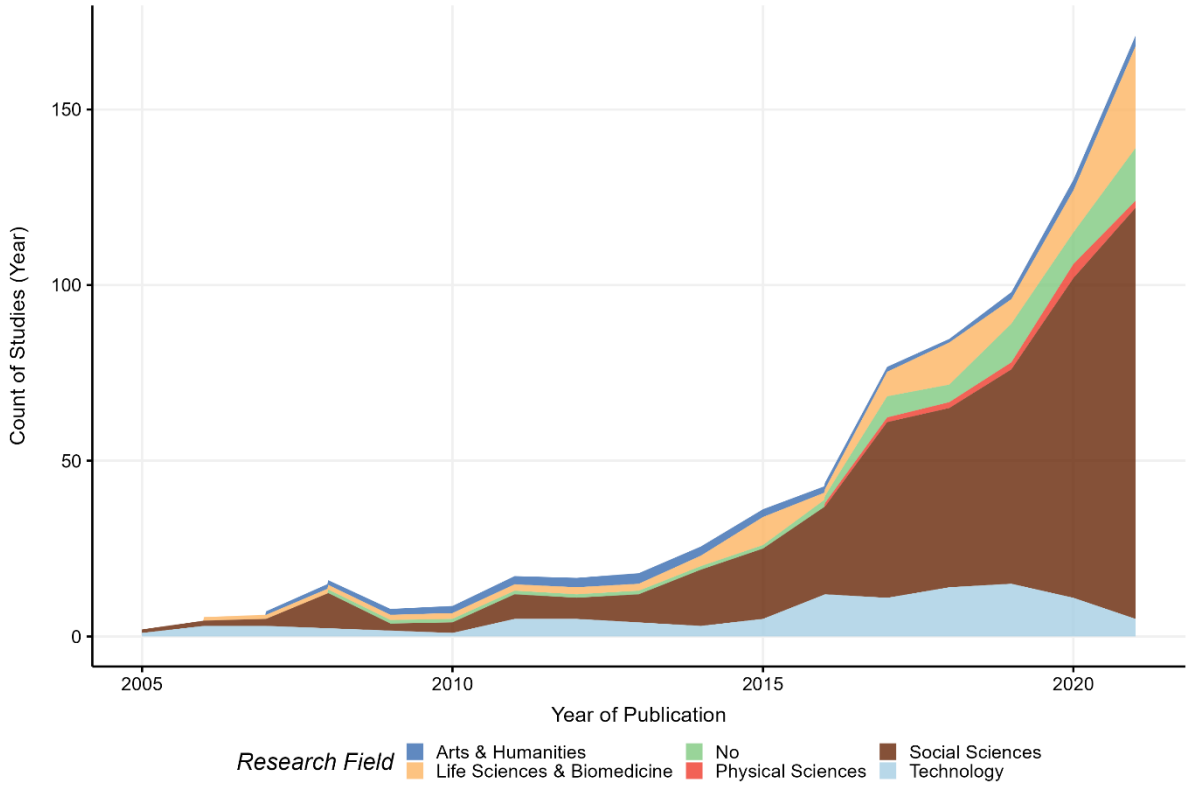

Figure 1: Annual frequency of newly published works utilizing text as a dataset within the Web of Science database.

Notwithstanding the promising opportunities offered by computational methods, measurement validity and the operationalization of complex latent constructs raise significant concerns within empirical social science text analysis (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 et al. (2020), 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 (see Devlin et al., 2019), promise to solve some of these limitations. However, "black box" large language models 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).

On the Uncertainties of Validating Text-based Measures

Thorough validation, involving both theoretical reasoning and empirical evidence, stands as a fundamental prerequisite for the application of CTAM in empirical research. It ensures that what is meant to be measured corresponds accurately to the actual measurement. When reviewing the literature, however, researchers often encounter ambiguity surrounding the validation process of CTAM, in particular deciding *when* and *how* to validate.

Deciding When to Validate

Regarding the question on *when* to validate, there seems to be a growing consensus on the overall necessity of validation. Grimmer and Stewart (2013), for instance, emphasize that researchers who conduct computational text analysis should always validate empirical measures ("validate,

validate, validate" (p. 271)). This view, however, is not shared universally, or at least current practice may not live up to it. For example, scholars who apply off-the-shelf dictionaries (i.e., previously developed list of words) often assume that these methods have been validated before, and expect similar levels of reliability and validity without testing it on new data (Chan et al., 2021; van Atteveldt et al., 2021). Likewise, the rise of large language models raises new questions on the necessity of validation. For example, language models such as GPT-3, the model behind ChatGPT, allow for "zero-shot classification", that is, the generation of new labels which are not defined by the researchers. Hence, the question of whether to validate might not be as easy to answer, especially as large language models promise to outperform the quality of human annotations (Gilardi et al., 2023; F. Huang et al., 2023), which are generally considered the gold-standard for validation.

Deciding How To Validate

Even when researchers have the intention to validate, determining *how* to do so can pose significant challenges. These challenges stem from the fact that conceptual and practical guidance on how to validate text-based measures is scarce and fragmented, prompting scholars to seek guidance from other research fields.

On the one hand, researchers can refer to the methodological literature in social and behavioral science research. Various general validation frameworks exist in subdisciplines such as psychology (Association et al., 2014; 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.[2]

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.[3] 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.

Internal and External Validation

One key conceptual distinction can be made between internal and external validation (Birkenmaier et al., 2023; 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*

[2] 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).

[3] 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).

validity, statistical validity, content validity, semantic validity, or construct validity[4] (DiMaggio et al., 2013; Grimmer & Stewart, 2013). External 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.

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., 2020).

Generic and 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., 2023; 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

[4] 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)).

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.

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 questions on *when* and *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 strengthened 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 when and 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.

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 categorise 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 research. Specifically, we aim to answer the question of *when* and *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 contextualise 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.

3. Research Design

Systematic Review

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).[5] 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

[5] 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).

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 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 six articles that we knew should be relevant. We successfully located all these articles in our preliminary literature list (see Appendix 2 for reference). 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).

Coding Procedure

Identification of Validation Steps

Figure 2 visualises the stepwise coding procedure for the 96 eligible studies we retrieved. [6]

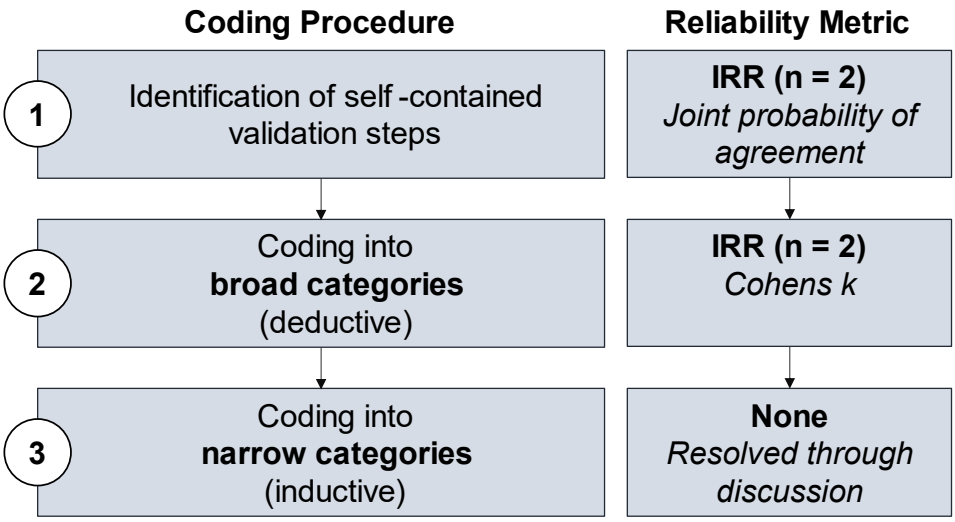

Figure 2: Coding Procedure, n (Number of Studies) = 96

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

[6] All data used for the review, including a list of all studies included in the review 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.

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 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%).

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

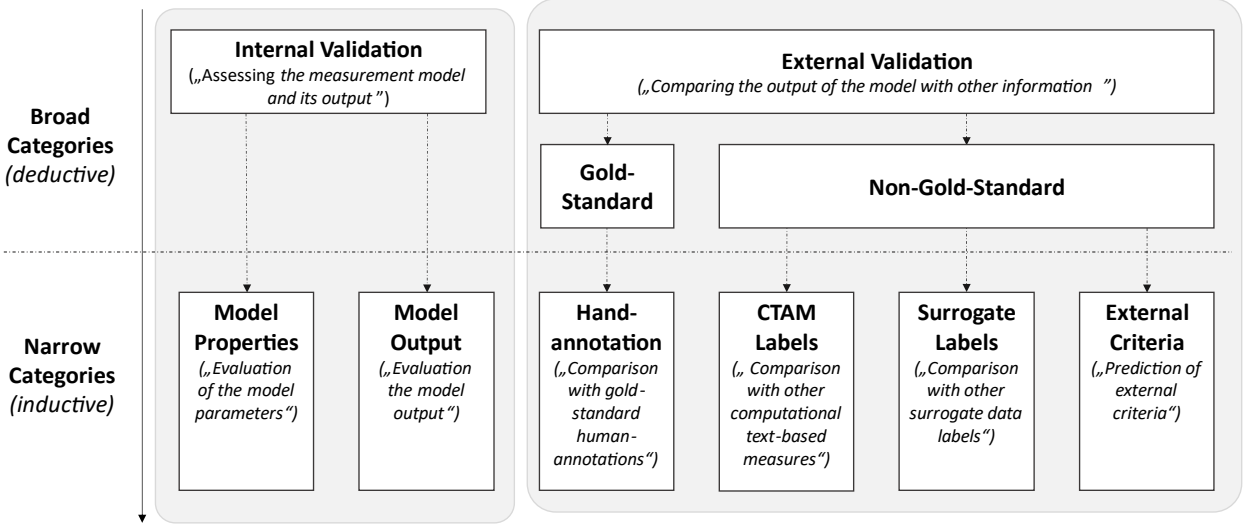

Figure 3:Coding Scheme

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 3, and the process of refining it is outlined in greater detail below.

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., 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).

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 selection criteria focused primarily on occupational status, gender, and regional background. In this regard, our aim was to include researchers with varied expertise and perspectives. 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., 2022) 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.

4. Results

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.

Basic Study Characteristics

Figure 4 summarises the key study characteristics.

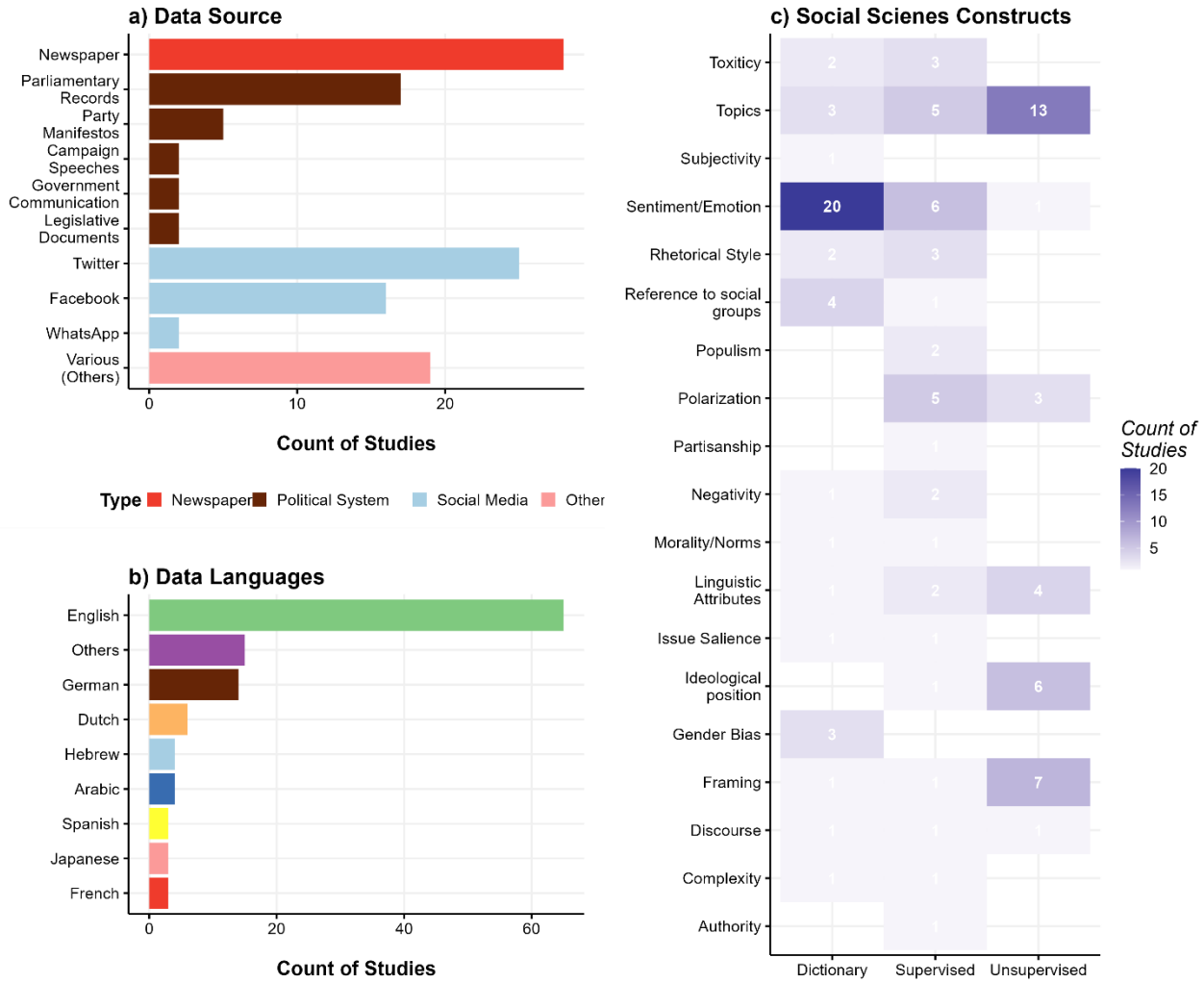

Figure 4: Overview Eligible Studies, n (number of studies) = 96

Figure 4a displays the origin of the most prominent data sources, clustered into major categories. Newspaper data constituted the most prominent data source analysed (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 4b 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 4c 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%).[7]

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

[7] 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).

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.

When Did Researchers Validate?

Next, we turn to the assessment of validation practices. To do so, we first evaluate *when*

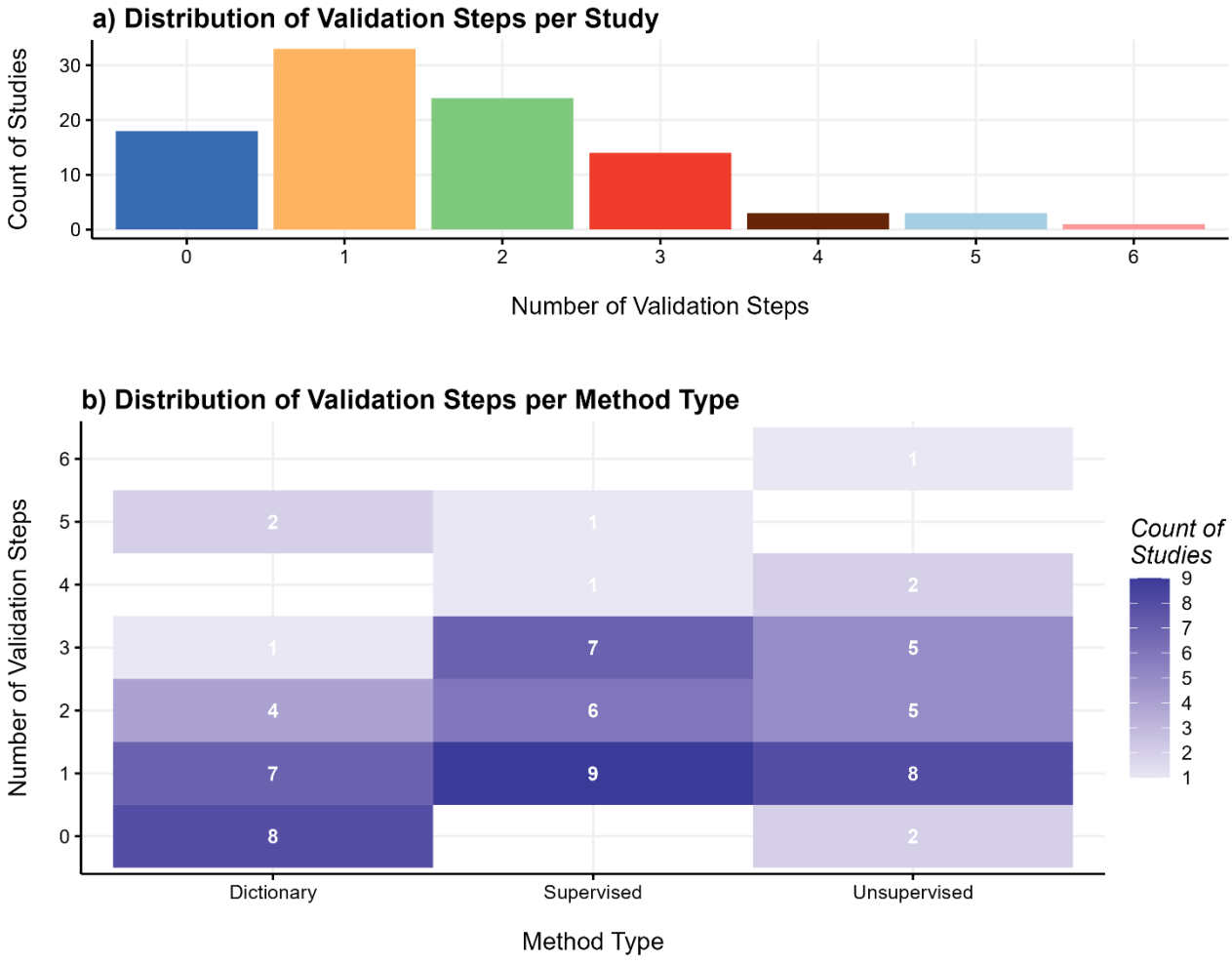

Figure 5: Number of Validation Steps (Uni- and bivariate), n (total number of validation steps) = 158

researchers reported to validate CTAM by plotting the total number of validation steps across studies.

Describing the number of validation steps per study might 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 5a depicts the distribution of the number of validation steps per CTAM applied.[8] 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 5b 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 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

[8] 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.

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.

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 6 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

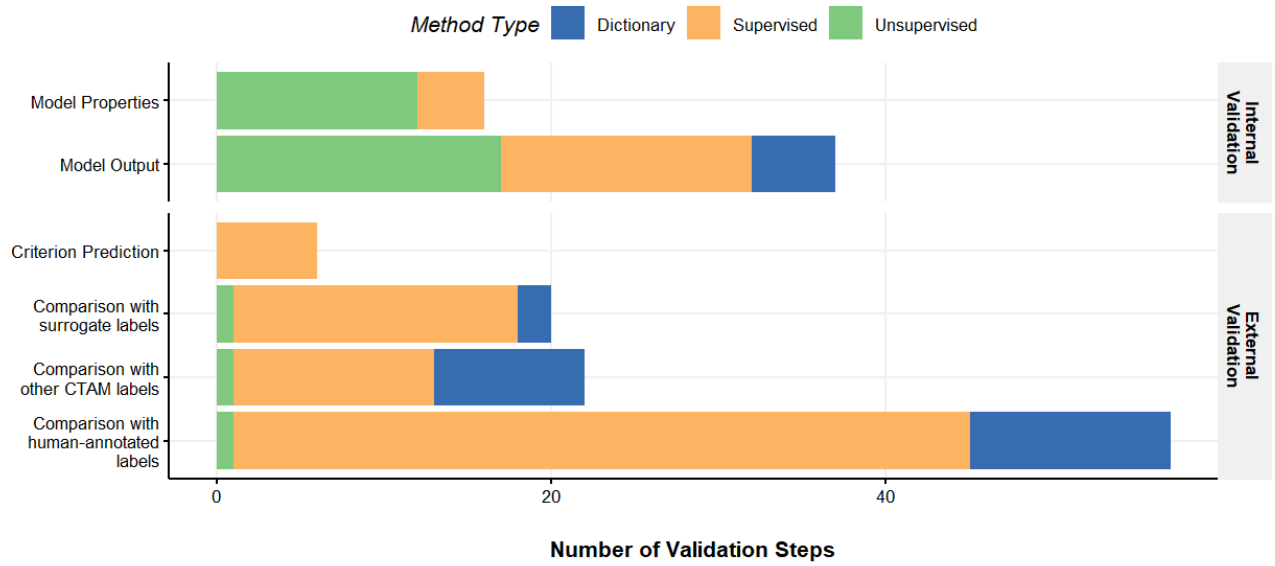

Figure 6. Categories of Validation Steps, n (total number of validation steps) = 158

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 5 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 7 provides a more detailed overview of the same data that, however, displays the concrete validation steps which will be subsequently discussed in greater detail.

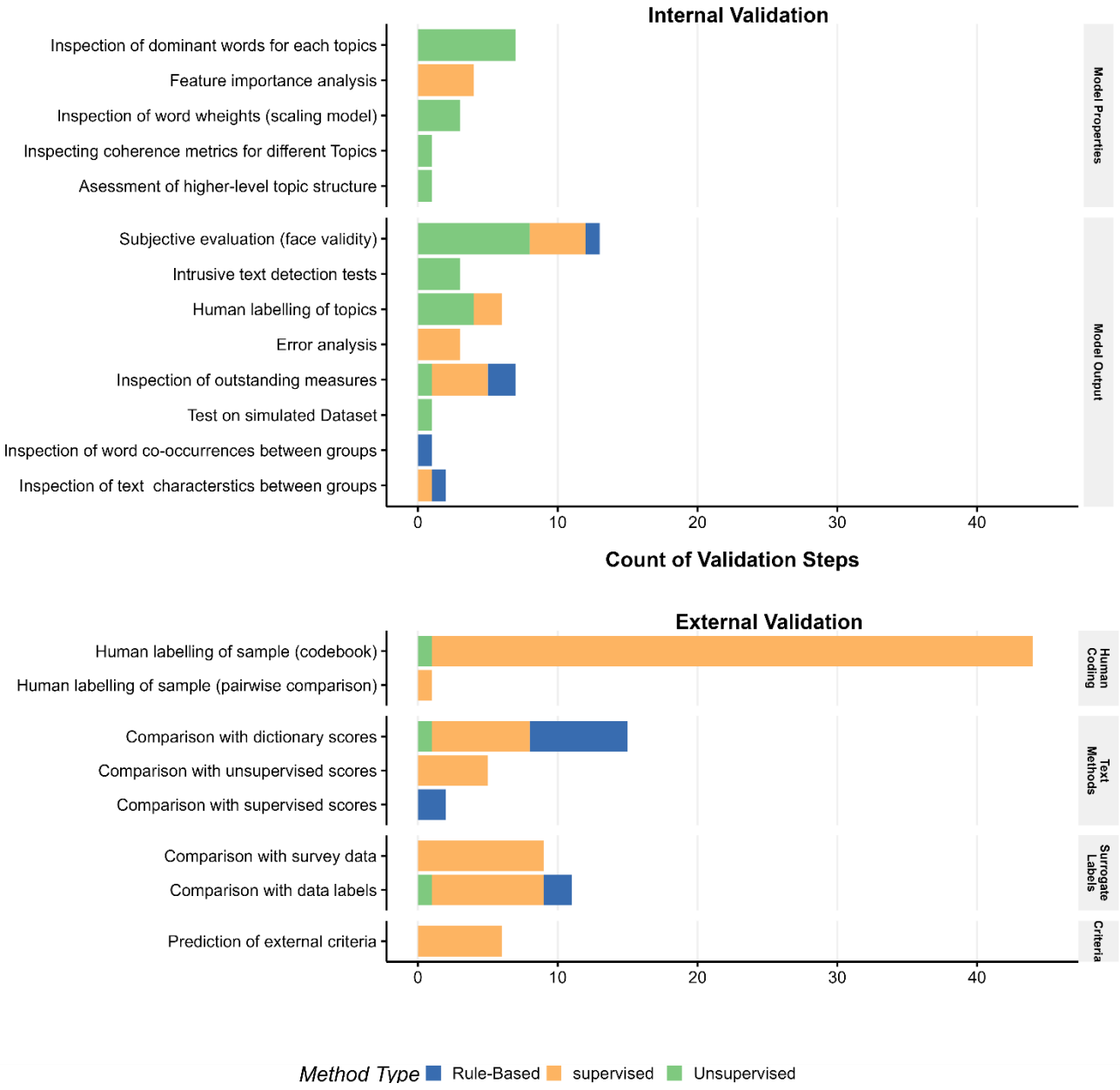

Figure 7: Overview of specific Validation Steps, n (total number of validation steps) = 158

Internal Validation

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

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.

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).

External Validation

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).

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 (Gelman & Wilson, 2022).

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).

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).

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 practices.

Generally, interviewees confirmed that they often carried out additional 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 1, along with specific quotes from the interviews.

Table 1: 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.

5. Discussion

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, scaled, and often automated. 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 therefore 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 carry out additional validation steps that they do 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.

Practical Recommendations for Improving Validation Practices

To tackle the issue of conceptual ambiguity regarding validity, we present practical recommendations based on our empirical findings and the previous literature on validity. In a nutshell, our recommendations are that researchers should (1) provide an explicit definition of the construct to be measured, (2) always validate the measure they use, (3) combine 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. Afterward, 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.

Justify your Construct Definition and Operationalization

First and foremost, we strongly urge authors to always outline and justify their construct definition and way of operationalization. Without a clear definition of these conceptual considerations, it is impossible to provide evidence for validity because the foundation for assessment remains uncertain (Clark & Watson, 2019). In other domains of social science research, reporting the construct definition and ways of operationalization is already well-established practice. Survey scales, for example, typically require validation that 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, 2011; 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. For example, it might be comparatively easy to validate whether a text-based measure is able to correctly identify party names in a text corpus. On the other hand, it might require more validation when applying CTAM to measure changes in political disenchantment language within the comment section of political videos from German and French TikTok users. Validating all aspects surrounding this fictitious example would require extensive validation for (1) the conceptualization of political disenchantment (2) the identification of political videos, (3) the consideration of multilingual differences between French and German (4) and the assurance of measurement stability over time. 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 research design, such as data and method selection, or preprocessing decisions (Denny & Spirling, 2018).

Always Validate CTAM that Measure Social Science Constructs

Next, we encourage researchers to always validate CTAM that measure social science constructs. Admittedly, results of our review indicate that most researchers in our sample did validate CTAM, 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 validation is necessary will become even more pressing with the rise of large language models (F. Huang et al., 2023; Kuzman et al., 2023; Reiss, 2023). Powerful pre-trained language models are widely shared on platforms such as *Hugging Face*, and interfaces such as the *ChatGPT* API allow for a user-friendly interaction with and utilization of large language models. Whereas we encourage scholars to take advantage of these impressive language models, we want to stress two major points.

On the one hand, we contend 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.

On the other hand, we stress that, at least for now, any text-based measurement needs careful validation, including state-of-the-art large language models. Because even when the text model and its parameters are within the control of the researcher and the results are, in principle, reproducible, there is still insufficient understanding of various types of biases and limitations inherent to CTAM (as discussed by Bender et al., 2021).[9] Speaking more generally, our recommendation to always validate CTAM that measure social science constructs 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 a "continuous" or "ongoing process." This involves the need for researchers to provide diverse validation evidence to establish the validity of a measurement instrument for a particular research objective (Flake et al., 2017). However, determining when an instrument is adequately validated for a given context requires constant reevaluation, as there exists no clear-cut threshold independent of the research context.

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 (see Figure 3). This fundamental distinction between internal and external validation aligns with core

[9] Bender et al. (2021), for instance, describe large language models as *stochastic parrot*, that is, "systems for haphazardly stitching together sequences of linguistic forms they have observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning" (p. 617).

principles in validation theory, despite potential variations in its reference across literature (Flake et al., 2017; Loevinger, 1957).[10]

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.

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.

[10] 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."

Always Conduct External Validation Using Human Annotations

Irrespective of the research context, we highly recommend to conduct external validation with human-annotated labels.[11] 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 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.

Maximize Transparency and Replicability

Lastly, we argue that scholars should maximize transparency and replicability of their validation. In our review, 42 per cent of studies did not report access to reproducibility materials, either code or data (see Figure 8), which is worrying despite clear recommendations for good scientific practice within communication research (Bakker et al., 2021; Dienlin et al., 2021; Humphreys et al., 2013). Thus, 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

[11] The only exception to this recommendation might be when the nature of the analysis is profoundly exploratory in nature, such as findings general patterns or word-cooccurrences in unstructured data.

studies and submitting registered reports, as well as conducting replication studies to verify computational reproducibility (Dienlin et al., 2021, Schoch et al., 2023).

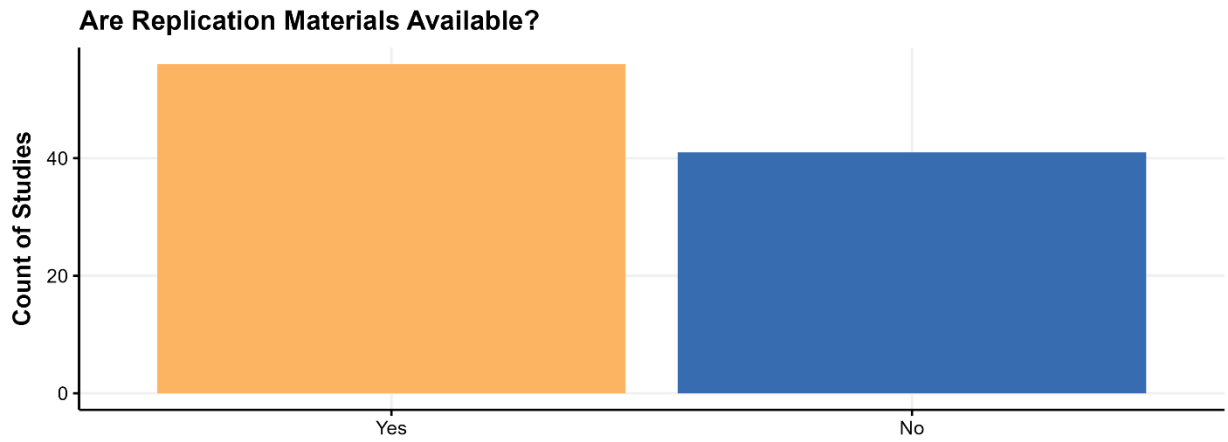

Figure 8: Replication Materials

Furthermore, researchers should document all the validation steps they implemented, refraining from leaving out any validation evidence.

Transparency also calls for a 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). [12]

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

An Outlook on more Unified Validation Frameworks

We are confident that the five recommendations outlined thus far will already assist researchers in improving their validation practices in the context of CTAM. What lends weight to our recommendations is that we based them directly on our empirical engagement with CTAM validation, supplemented by a synthesis of prior methodological research on validity within other social science fields. However, it is important to recognize that our recommendations, even though they provide strong conceptual guidelines on how to approach validation, cannot provide researchers with all-encompassing practical guidance for all scenarios involving CTAM validation. Therefore, we will 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.

One area in which recently there has been progress toward more principled validation approaches is the validation of multilingual text analysis. 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. The proposed to follow a process logic, including data validation, input validation, process validation, and output validation. Similarly, Ho and Chan (2023) proposed 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.

In what appears to be the most comprehensive effort to date, Birkenmaier et al. (2023) proposed a unified validation framework for the validation of text-based measures of social science constructs. Their framework, *ValiTex*, is specifically designed for the validation of (monolingual) computational text-based measures of social-science constructs. Conceptually, *ValiTex* is organized along three distinct phases that originate from the psychometric literature on measurement theory and consists of two components, a conceptual model, and a checklist. The conceptual model establishes a general structure for how to approach validation, including three distinct phases that are grounded on decade-long research on validation in the social sciences. The checklist, on the other hand, then provides a detailed overview of validation steps for each type of validation evidence and CTAM to be applied, together with an opiniated opinion on the overall relevance for each validation step.

Similarly, researchers might also make use of method-specific frameworks tailored towards specific workflows or methods. Often, these method-specific contributions propose valuable workflows for validation, 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 (Chan & Sältzer, 2020; 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.

Despite notable variations in the terminology and the focus of the validation frameworks presented, 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.

6. 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. (2023) 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.

7. 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 reserach. 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.

8. References

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

Association, A. E. R., Association, A. P., & Education, N. C. on M. in. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

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.

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.

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/lsq.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

Birkenmaier, L., Lechner, C., & Wagner, C. (2023). *ValiTex—A unified validation framework for computational text-based measures of social science constructs*. https://doi.org/10.48550/ARXIV.2307.02863

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

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

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

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

Ceccarelli, L. (1998). Polysemy: Multiple meanings in rhetorical criticism. *Quarterly Journal of Speech*, *84*(4), 395–415.

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

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

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.*

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

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.

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

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* (arXiv:1909.03004). arXiv. http://arxiv.org/abs/1909.03004

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105–123.

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

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

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., Daumeé III, H., & Crawford, K. (2018). Datasheets for Datasets. *ArXiv:1803.09010 [Cs]*. http://arxiv.org/abs/1803.09010

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

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

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.

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.

Hedrih, V. (2019). *Adapting Psychological Tests and Measurement Instruments for Cross-Cultural Research: An Introduction*. Routledge. https://doi.org/10.4324/9780429264788

Helfferich, C. (2022). Leitfaden-und experteninterviews. In *Handbuch Methoden der empirischen Sozialforschung* (pp. 875–892). Springer.

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

House, F. (2022). *Freedom in the world 2022* [Report]. Freedom House. https://apo.org.au/node/316708

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

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.

Jankowski, M., & Huber, R. A. (2022). When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models. *Political Analysis*, 1–15.

Kapoor, S., & Narayanan, A. (2022). *Leakage and the Reproducibility Crisis in ML-based Science* (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

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry*. Princeton university press.

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

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). *ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification* (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

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) [dataset]. Manifesto Project. https://doi.org/10.25522/MANIFESTO.MPDS.2022A

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

Lipton, Z. C. (2017). *The Mythos of Model Interpretability* (arXiv:1606.03490). arXiv. http://arxiv.org/abs/1606.03490

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.

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.

Mayring, P. (2004). Qualitative content analysis. *A Companion to Qualitative Research*, *1*(2), 159–176.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

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

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

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.

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

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, 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. (2022). Robust speech recognition via large-scale weak supervision. *ArXiv Preprint ArXiv:2212.04356*.

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

Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark* (arXiv:2304.11085). arXiv. https://doi.org/10.48550/arXiv.2304.11085

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.

Robinson, O. C. (2014). Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative Research in Psychology*, *11*(1), 25–41.

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

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

Schoch, D., Chan, C., Wagner, C., & Bleier, A. (2023). *Computational Reproducibility in Computational Social Science* (arXiv:2307.01918). arXiv. https://doi.org/10.48550/arXiv.2307.01918

Schub, R. (2022). Informing the Leader: Bureaucracies and International Crises. *American Political Science Review*, 1–17. https://doi.org/10.1017/S0003055422000168

Scimago. (2022, October 3). *SJR - About Us*. https://www.scimagojr.com/aboutus.php

Shermer, M. (2008). Patternicity: Finding meaningful patterns in meaningless noise. *Scientific American*, *299*(5), 48.

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. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Stemler, S. (2000). An overview of content analysis. *Practical Assessment, Research, and Evaluation*, *7*(1), 17.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). 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

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

van Atteveldt, W., & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

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

Waisbord, S. (2019). *Communication: A post-discipline*. John Wiley & Sons.

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

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.

9. Appendices

Appendix 1: 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

Appendix 2: Benchmark Literature

- Goet, Niels D. 'Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015'. Political Analysis 27, no. 4 (2019): 518–39.
- Gilardi, Fabrizio, Theresa Gessler, Mael Kubli, and Stefan Muller. 'Social Media and Political Agenda Setting'. *Political Communication* 39, no. 1 (2 January 2022): 39–60.
- Di Cocco, J., & Monechi, B. (2022). How Populist are Parties? Measuring Degrees of Populism in Party Manifestos Using Supervised Machine Learning. *Political Analysis*, 30(3), 311-327. doi:10.1017/pan.2021.29
- 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. Jo*urnal 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.

-

Appendix 3: 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*)

Appendix 4: 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

Appendix 5: Overview of Studies per Journal

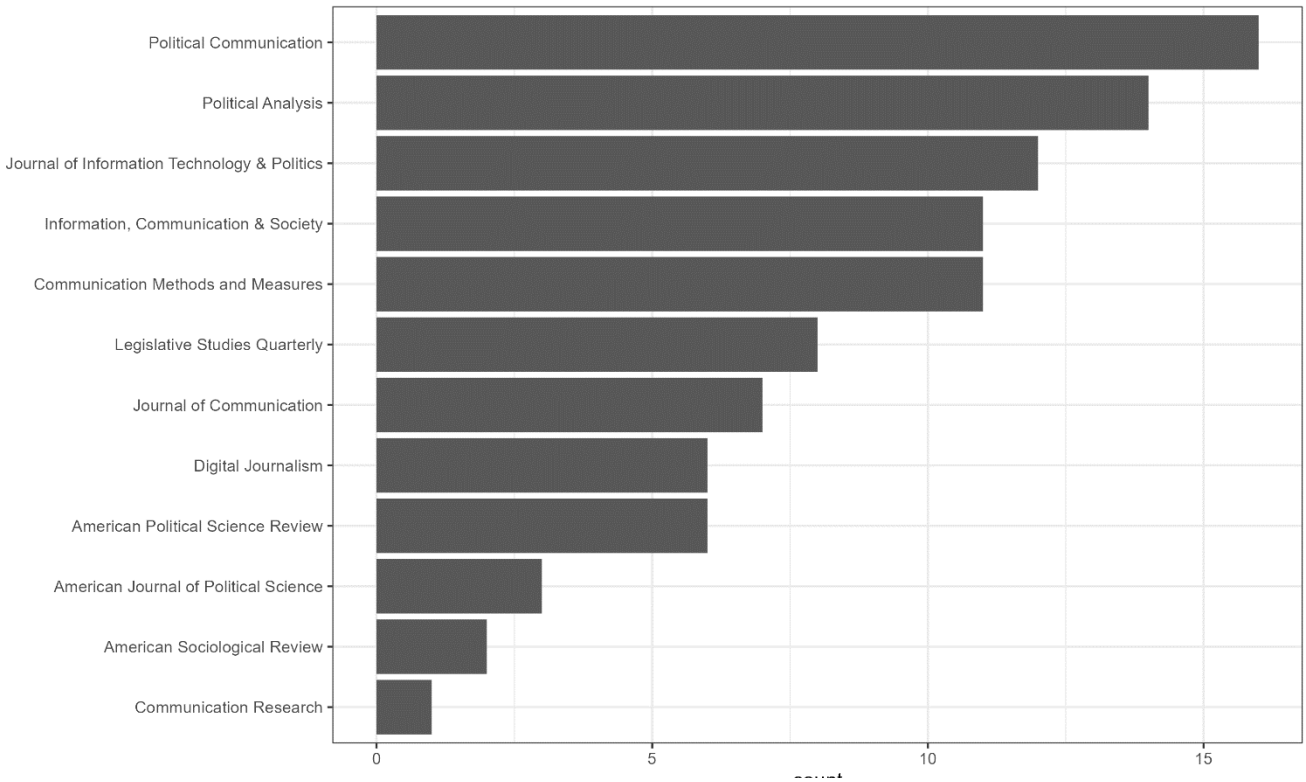

Figure 9: Overview of Studies per Journal

Appendix 6: 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? (*Costs, missing guidelines etc*.)
	Do you have any ideas on **how to improve** the validation of computer-assisted text methods?

Conclusion	Is there anything else you would like to tell us?