Preprint

ValiTex - a unified validation framework for computational text-based measures of social science constructs

***Lukas Birkenmaier[1], Claudia Wagner[1,2,3], Clemens Lechner[1]**

[1]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

[2]RWTH Aachen University, Germany

[3]Complexity Science Hub Vienna, Austria

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

ABSTRACT: Guidance on how to validate computational text-based measures of social science constructs is fragmented. Although scholars generally acknowledge the importance of validating their text-based measures, they often lack common terminology and a unified framework to do so. This paper introduces ValiTex, a new validation framework designed to assist scholars in validly measuring social science constructs based on textual data. The framework draws on a long-established validity concept in psychometrics but extends these concepts to cover the specific needs of computational text analysis. ValiTex consists of two components, a conceptual framework and a dynamic checklist. Whereas the conceptual framework provides a general structure along distinct phases on how to approach validation, the dynamic checklist defines specific validation steps and provides guidance on which steps might be considered recommendable (i.e., providing relevant and necessary validation evidence) or optional (i.e., useful for providing additional supporting validation evidence). We demonstrate the utility of the framework by applying it to a use case of detecting sexism from social media data.

*****Corresponding Author: lukas.birkenmaier@gesis.org

Introduction

Computational text-based measures of social science constructs are difficult to validate (Baden et al., 2020; Kantner & Overbeck, 2020). This is not only expressed in empirical studies illustrating concerningly high variability and inconsistent results (Gonçalves et al., 2013; van Atteveldt et al., 2021), but manifests itself in a lack of established standards on how to best validate text-based measures of social science constructs (Birkenmaier et al., 2023). Guidance on how to validate measures is particularly relevant because, in social science research, textual data and its grammatical patterns are not per se the unit of interest. Instead, the interest lies in using textual data and grammatical patterns to operationalize underlying theoretical constructs, which are derived using computational methods (Kantner & Overbeck, 2020 based on Krippendorff 2018, pp. 29–40). Naturally, however, there's a gap between the knowledge available from analysing text patterns and the knowledge researchers seek, that is, the true essence of the underlying constructs (Hughes, 2018). Therefore, in order to determine whether a measure is a valid operationalization of the underlying construct, it is crucial to thoroughly validate the relationship between text-based measures and theoretical constructs (Hughes, 2018). Whereas significant progress has been made in advancing validation theory and attendant practices across various domains of social science research (Adcock & Collier, 2001; Kane, 2013; Messick, 1989), there is a dearth of guidance on how to best validate computational text-based measures of social science constructs. Instead, methodological frameworks and best-practices for validation are often transferred from other social science disciplines without fully accounting for the specific nature of text-based measures (Adcock & Collier, 2001), or lack justification altogether. This, however, often results in confusion about what constitutes good and sufficient validation evidence. Furthermore, the lack of conceptual clarity makes critical evaluation and comparison of text-based research challenging, which is also expressed by the absence of a shared terminology for referencing different types of validation evidence (Birkenmaier et al., 2023).

This article aims to address the need for standardization and guidance by proposing a unified validation framework for computational text-based measures of social science constructs (ValiTex)[1]. The framework builds on well-established validity criteria and terminology employed in psychometrics and quantitative as well as qualitative social science research but is tailored to the distinct demands of computational text-based research. Through a case study that involves the measurement of sexism using social media data, we demonstrate how validation steps at every stage of the validation process can, in principle, be defined and evaluated using the framework proposed. To the best of our knowledge, *ValiTex* is the first attempt at providing a comprehensive and broadly applicable validation framework for computational text-based measures of social science constructs.

[1] For the accompanying webapp, please visit: https://lukasbirki.shinyapps.io/ValiTex/. You can also run the shiny app locally by installing the corresponding R-package from https://github.com/lukasbirki/ValiTex.

Background

A primer on the challenges for text as data

Text as data commonly entail the use of algorithms and automated software tools to assign labels (i.e., empirical measures) to text, varying in the extent of human supervision required (Grimmer & Stewart, 2013). However, using computational methods to measure social science constructs from texts is challenging (Drisko & Maschi, 2016; Krippendorff, 2018). Yet, many practical challenges that arise in the validation process are not unique to computational methods. On the contrary, extensive methodological research in the field of manual content analysis highlights the overall complex nature of text. Particularly, there is uncertainty about how social science constructs manifest in texts. There is consensus that textual data in itself has no objective qualities and, thus, cannot be characterized by a single meaning that could be "found" or "measured" directly (Drisko & Maschi, 2016; Krippendorff, 2018). Consider, for instance, a text sample of a politician's speeches in which features such as sentence structure, vocabulary usage, or reference to contextual information could be utilized to create a measure of communication style. Hence, any text-based measure alone would be just one among several potential operationalizations, lacking comprehensiveness (Krippendorff, 2018). Additionally, the task is complicated by the fact that most texts are gathered passively, meaning that their creation was driven by intentions and motivations unrelated to the research interest. This stands in contrast to actively collected forms of data, such as surveys, where data is intentionally collected to capture specific constructs.

More fundamentally, language can be interpreted in different ways depending on the readers perspective, knowledge, or cultural background. Examples of these characteristics are the presence of sarcasms or irony (Ravi & Ravi, 2015), ambiguity and polysemy (Boxman-Shabtai, 2020; Roberts, 1989), or context-specific references and interpretations (Krippendorff, 2018; Mayring, 2004). Although qualitative content analysis aims to reduce subjectivity by creating rigorous guidelines and rules for coding texts (Drisko & Maschi, 2016; Krippendorff, 2018), it is still very subjective, as it depend on the coders' subjective semantic interpretation and comprehension of language as they process and evaluate textual data (Baden et al., 2020).

Computational methods, in contrast, lack the ability to interpret textual data like humans. Instead, they rely on algorithms or neural networks that are programmed or trained to identify potentially significant patterns in the data to provide numerical summaries of the textual input, such as classifier confidence or sentiment scores (Baden et al., 2021). Computational methods, however, come with additional challenges. For example, the text analysis pipeline is usually characterized by high researchers' degrees of freedom. Because computational methods require a variety of analytical choices, starting from construct operationalization, method selection, data pre-processing and model parameter adjustment, researchers' choices often affect the outcomes of analyses significantly, resulting in alternative, sometimes contradictory findings (Denny & Spirling, 2018; Pipal et al., 2022). In addition, little is known about how biases and errors induced by computational methods influence in the measurement process. On the one hand, simpler methods such as dictionaries often build on overly simplistic assumptions on the structure of texts, ignore all sorts of relevant textual information, such as word order or sentence structure, and may be subjective and biased depending on how the dictionary was constructed. On the other hand, more complex methods, in particular large language models (Devlin et al., 2019), often conceal relevant information on how the output was generated. Furthermore, bias might arise from the data used to train and test the models, such as data leakage (Gibney, 2022; Kapoor & Narayanan, 2022), non-representative datasets in respect to the measured construct (Cai et al., 2022), or model-inherent human stereotypes (van Giffen et al., 2022).

Validation ensures the credibility of text-based measures

Given the inherent complexity of textual data and the limitations of computational methods, the primary justification for using text analysis methods is empirical (Grimmer et al., 2022). This means that text-based models are not justified based on a theoretical model of language but because they have either worked well in the past when applied to similar problems, or because they are the only methods available to deal with larger volumes of text. Because of these limitations, however, researchers need to *validate* their measures each time they apply it to a new empirical context.

Measurement validity is defined as whether a scientific test measures what it purports to measure (Cattell, 1946; Kelley, 1927). To demonstrate validity, researchers have to accumulate different validation evidence to justify that their measurement instrument is able to accurately reflect the theoretical construct of interest (King et al., 1994; Krippendorff, 2018). This is especially relevant when the constructs of interest are not directly observable ("latent"), as is often the case with social-science constructs such as individual attitudes, personality traits or communication frames (Flake et al., 2017). Because latent constructs need to be "translated" into empirical constructs, that is, into sets of interrelated variables or pieces of information that covary and can be observed (i.e., responses to survey scales or words in a text), validation ultimately verifies that the observed measures adequately capture the theoretical construct, rather than other relationships within the data (Grimmer et al., 2022).

In the social sciences, a multitude of approaches have been developed to collect and evaluate validation evidence. Arguably the most advanced and influential toolbox for ensuring validity has been developed in psychometrics, the field of research that focuses on the measurement of psychological constructs (Adcock & Collier, 2001; Kane, 2013). Whereas earlier conceptualizations of validity up until the 1970s primarily distinguished between more or less distinct types of validity, such as criterion (i.e., predictive capability towards a suitable criterion measure) or content validity (i.e., considerations on the domain relevance) (Cronbach & Meehl, 1955; Cureton, 1951; Kelley, 1927), the present-day conception of validity involves gathering multiple sources of evidence to establish a comprehensive understanding of construct validity (E. K. Chan, 2014; Messick, 1989; Shepard, 1993; *The Standards for Educational and Psychological Testing*, 2014). Construct validity thereby represents a unified and multi-faceted understanding of validity, which can be seen as a continuous evaluation of the appropriateness of inferences from a method and its measures for a specific research context. To demonstrate construct validity, researchers have to systematically conduct and accumulate different types of validation evidence to strengthen the confidence in the validity of their measures (Cronbach & Meehl, 1955). Whereas this comprehensive understanding of validity has remained the dominant view in many social science domains, one can observe substantial differences in validation practices across subdisciplines of social science research (E. K. Chan, 2014). This is particularly noticeable for the validation of computational text-based measures of social science constructs, which lacks a single clear organizing framework altogether (Birkenmaier et al., 2023; Grimmer et al., 2022).

On the need for a comprehensive understanding for validating computational text-based measures of social science constructs

The strategies that researchers currently employ to validate their empirical measures are highly varied (Baden et al., 2021; Birkenmaier et al., 2023; Fang et al., 2022; Grimmer et al., 2022). To a certain extent, this follows from the diverse nature of text analysis measures, which necessitate different validation strategies. What is worrying, however, is the absence of a unifying framework to guide the variety of validation activities available. In absence of such a framework, there is a lack of conceptual clarity on how to best conduct and communicate validation for text-based measures within the social sciences (Baden et al., 2021). In their systematic review, Birkenmaier et al. (2023), for instance, show that only 9 per cent of validation steps reported in the papers they reviewed explicitly specified the type of validation evidence they referred to. Moreover, their result show that scholars did apply a great variety of validation steps but rarely justified their decisions based on a systematic comprehension of which types of validation evidence are applied. When validation steps and practices are applied in such a selective and inconsistent manner, however, the *uncertainty* over the validity of the measure – and hence about the credibility of any empirical results involving the measure – increases. In other words, without substantial validation evidence, researchers can never completely rely on the accuracy of their empirical findings, regardless of whether they turn out to be correct or inaccurate.[2] Furthermore, lacking methodological guidance also results in inconsistent reporting standards and confusion about how to provide replication materials, which make the evaluation of validation evidence even more challenging. Clearly, the lack of conceptual and practical guidance poses a problem for researchers, who can easily lose track of how they should validate their text-based measures in their substantive research projects, and how they should evaluate validation efforts by other scholars (Baden et al., 2021). Thus, we argue that a more comprehensive view of measurement validation is needed.

[2] There is a wide range variants of how computational measures might, in fact, turn out to be erroneous, such as replicating stereotypical associations or negative sentiment towards specific groups (for an overview, see Bender et al., 2021) or relying on insufficient proxies (Jankowski & Huber, 2022)

The ValiTex Framework

We propose a novel framework for validating computational text-based measures of social science constructs, called ValiTex, that equips researchers with the necessary terminology and tools to navigate the validation of text-based measures. Conceptually, ValiTex builds on decade-long history of methodological research in the social sciences, in particular psychometrics, which has developed the most established and comprehensive understanding of validation practices (Flake et al., 2017; Loevinger, 1957). To account for the unique characteristics of computational text-methods, we furthermore draw on earlier work of a systematic review of validation practices by Birkenmaier et al. (2023). In their review, Birkenmaier et al. provided an extensive empirical overview and classification of validation practices across 96 peer-reviewed articles (published in the field of political communication) that applied and validated text-based measures of social science constructs.[3] Their systematic overview not only charts the present state of validation in text-based research but is also accompanied by an initial taxonomy that describes and classifies different kinds of validation evidence. We incorporate the initial taxonomy proposed by Birkenmaier et al. for ValiTex, in particular the differentiation between overarching validity dimensions (i.e., types of related validation evidence) and empirical validation steps (i.e., single reported and clearly demarcated validation activities). In the upcoming sections, we will first offer a brief overview of the framework components, followed by an in-depth examination of the framework's application within the validation process.

[3] In addition, the review is further complemented by qualitative expert interviews to document further subtle validation evidence.

Components of the ValiTex Framework

ValiTex consists of two complementary components: the conceptual framework and a checklist. The conceptual framework, on the one hand, provides a general template across three distinct phases for how to approach validation for text-based measures of social science constructs. The checklist, on the other hand, then entails a detailed list of empirical validation steps for each phase, together with their description and an opinionated assessment of each validation step.

Conceptual Framework

Figure 1 displays the conceptual framework. Comprising three distinct phases, the framework is rooted in the well-established principles of measurement theory found within the psychometric literature which offers the most comprehensive and cohesive conception of validity for social science research (Flake et al., 2017; Loevinger, 1957). While the phases within the framework build

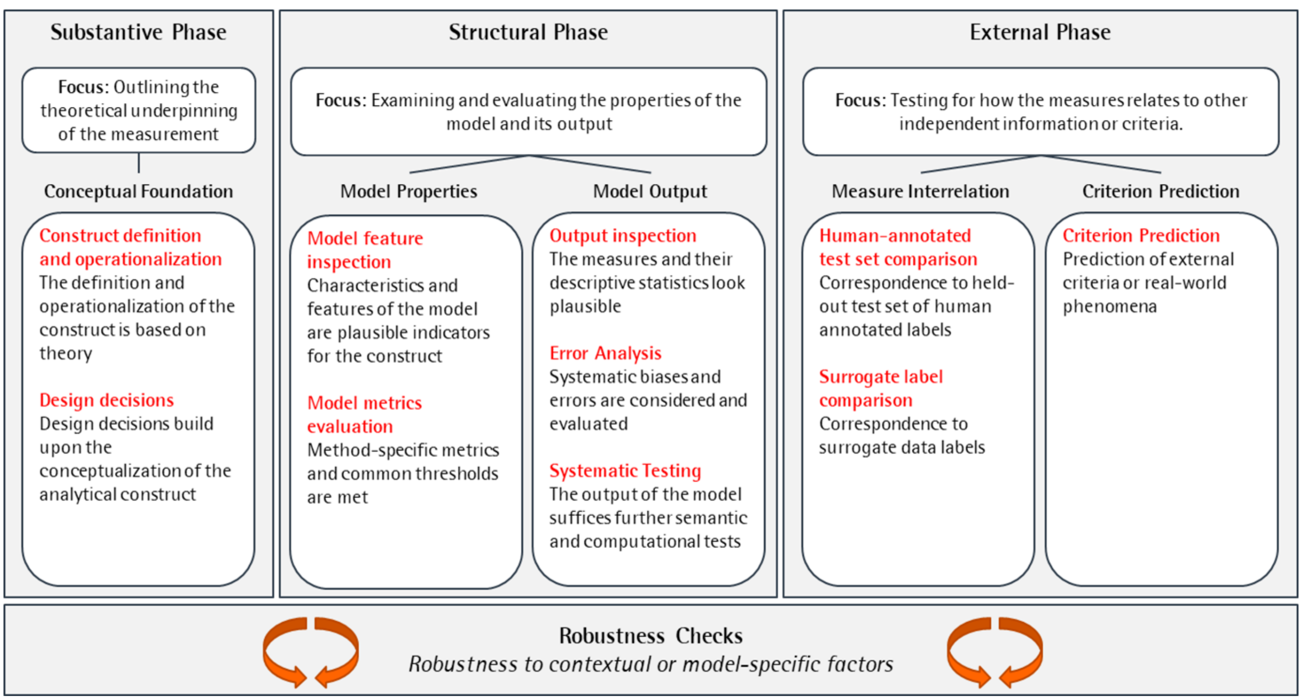

Figure 1: Conceptual Framework

up on each other, they should not be perceived as distinct chronological entities. Instead, they are designed to provide a general structure to group together overarching dimensions of validation steps that might be adapted by the researcher for specific research contexts.

In the substantive phase, researchers should always start with outlining the theoretical underpinning of the measurement. In the structural phase, researchers should examine and evaluate the properties of the text model and its output. In the external phase, researchers should test for how the empirical measures relate to other independent information or criteria. In addition, researchers should continuously apply robustness checks to demonstrate that their measurement is robust to changes in contextual or model-specific factors.

Checklist

For each phase in the conceptual framework, the **checklist** then defines and describes empirical validation steps for different use cases.

Selecting the Appropriate Use Case

ValiTex primarily categorizes text models based on their underlying learning approaches, distinguishing between rule-based, supervised, semi-supervised, and unsupervised methods. However, ValiTex goes beyond these broad approaches, depending on whether the researcher has access to training / fine-tuning data that include gold-standard labels, and whether the researcher specifies the output categories. Table 1 provides an initial summary and description of these use cases, each accompanied by its own unique checklist. We deliberatively decided not to include a use case for closed source blackbox APIs (such as Perspective API, ChatGPT, among others), and we generally abstain from using them for measuring social science constructs. This is due to the lack of control researchers have over these models, which can undergo significant changes at any point without anyone noticing because the algorithms and models behind the 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 scientific measurement (see in particular Schoch et al., 2023)

#	Use Case	Learning Approach	Training / Finetuning data required	Known output categories	Description	Example
A	**Applying Dictionaries**	Rule-Based	No	Yes	Assign scores to text units using predefined word lists	A dictionary assigns polarity values ranging from -1 to 1 to each known text unit
B	**Classification using Traditional Supervised Machine Learning Model**	Supervised	Yes	Yes	Assign known output categories based on labelled training data	A Logistic regression model is trained on labelled customer reviews and predicts “positive” and “negative” reviews
C	**Classification using Finetuned Machine-Learning Model**	Semi-Supervised	Yes	Yes	Assign known output categories based on a fine-tuning process on a small subset of labelled data	A pretrained BERT model is fine-tuned on labelled social media posts and predicts “offensive” and “non-offensive” posts
D	**Zero-Shot/Few-Shot Classification (known output categories)**	Semi-Supervised	No	Yes	Assign known output categories without any finetuning on labelled data	GPT-3 predicts “political” or “non-political” speeches
E	**Zero-Shot/Few-Shot Classification (unknown output categories)**	Unsupervised	No	No	Assign unknown output categories without any finetuning on labelled data	GPT-3 suggests topics for texts (not prescribed by the researcher)
F	**Topic Modelling**	Unsupervised	No	No	Assign topics without any labeled data	An LDA topic model generates 13 topics coherent topics
G	**Text Scaling**	Unsupervised	No	Yes	Assign scale scores without any labelled data	A Wordfish model assigns scale values from -1 to 1 to politicians’ speeches

Table 1: Use Cases ValiTex that each come with an adapted checklist.

Documentation of Validation Steps

Once the appropriate use case is selected, the respective checklist then provides a detailed overview of validation steps available. For instance, ValiTex offers an initial evaluation regarding the relevance of each validation step. This evaluation is based on a comprehensive review of validation studies across various fields and current validation practices, as documented in Birkenmaier et al. (2023). It is further influenced by qualitative interviews conducted with researchers who utilize text-based measures, as reported in the same publication. Additionally, discussions among the authors and interactions with researchers at scientific conferences contribute to this classification. As a result, validation steps are categorized as either recommended (i.e., central for providing relevant and necessary validation evidence) or optional (i.e., helpful for offering additional supporting validation evidence). In practical terms, we advise researchers to perform all the recommended validation steps and provide a rationale in case they omit one or more of the recommended validation steps. This is because every validation step that is not performed increases the uncertainty about the validity of the measure and hence about the trustworthiness of any resulting research findings, whereas the validity-related evidence produced in each step acts to reduce that uncertainty. Therefore, it is in the best interest of both researchers and the consumers of research that the validation steps outlined by ValiTex be followed and reported. To guide researchers, the checklist also incorporates references to pertinent literature.[4]

Applying the ValiTex Framework

In the following chapter, we will outline the phases within ValiTex in greater detail, including a detailed discussion of validation steps within each phase.

Substantive Phase

At the beginning of the research process, the *substantive phase* involves outlining the theoretical underpinning of the measurement. Validation steps within the substantive phase should therefore demonstrate that the measurement is based on a strong conceptual foundation, including the operationalization of the construct and the design decisions around the measurement process. Naturally, some of the validation steps in the substantive phase might not qualify as empirical (Messick, 1989), which is why they are often disregarded in the presentation of validation evidence altogether (see Birkenmaier et al., 2023). However, from a conceptual point of view, any empirical

[4] Naturally, there is a trade off on the degree of practical guidance and heterogeneity of research contexts for text-based research. As ValiTex aims to provide a generally applicable and uniform framework, we emphasize the primary role of the researcher to adapt and judge the quality of specific validation steps. The information provided in the checklist might provide a basis for such judgments.

measurement requires a sufficient theoretical and methodological foundation, which should be demonstrated by the researcher using both theoretical reasoning and empirical evidence (Flake et al., 2017). Table 2 displays the validation steps in the substantive phase.

Table 2: Validation Steps in the Substantive Phase

ID	Validation Steps	Considerations	Performance Criteria	A	B	C	D	E	F	G
	Construct Definition & Operationalization									
I.1	**Documentation of the conceptual background**	Have I conducted a literature review or consulted with domain experts to gain a comprehensive understanding of conceptual background the construct?	Evidence of engagement with the construct	R	R	R	R	R	R	R
I.2	**Justification of the operationalization**	Have I sufficiently explained how the construct should manifest itself in the textual data?	Theoretical reasoning	R	R	R	R	R	R	R
I.3	**Manual Precoding**	Have I conducted a pilot study using manual coding to evaluate the inter-rater agreement and reliability on detecting the construct by hand?	Agreement between coders	O	O	O	O	O	O	O
Design Decisions										
I.4	**Justification of data collection decisions**	Have I selected a dataset that is representative and relevant to the research question and population of interest? Have I justified the data selection decisions (e.g., using keywords)? Have I assessed the quality and completeness of the dataset and checked for potential biases or inconsistencies?	Theoretical reasoning	R	R	R	R	R	R	R
I.5	**Justification of method choice**	Have I selected the appropriate type of method based on the operationalization of the construct and data characteristics? Have I justified the concrete selection of a particular model, and have I documented relevant features of the model?	Theoretical reasoning	R	R	R	R	R	R	R
I.6	**Justification of the level of analysis**	Have I selected the appropriate level of analysis? Have I considered potential problems when aggregating scores from lower to higher levels (e.g., sentence to paragraph level)?	Theoretical reasoning	R	R	R	R	R	R	R
I.7	**Justification of preprocessing decisions**	Have I justified the preprocessing decisions, such as removing stopwords, based on the presumed manifestation of the construct in the text?	Theoretical reasoning	R	R	R	R	R	R	R

Columns: A = Use Case "A. Applying Dictionaries", B = Use Case "B. Classification using Traditional Supervised Machine Learning Model", C = Use Case "C. Classification using Finetuned Machine-Learning Model", D = Use Case "D. Zero-Shot/Few-Shot Classification (known output categories)", E = Use Case "E. Zero-Shot/Few-Shot Classification (unknown output categories)", F = Use Case "F. Text Scaling", G = Use Case "G. Topic Modelling"
Values: R = Recommended, O = Optional, - = not applicable for this type of method;
An interactive version of the checklist is available online: https://lukasbirki.shinyapps.io/ValiTex/

Conceptual Foundation

Construct definition and operationalization. As a first step, researchers need to make sure that the definition and operationalization of the construct is based on theory. This task of defining a construct's scope and theoretical underpinnings is far from trivial and has received substantial attention in the psychometric literature (for a comprehensive review, see Rusticus (2014)). Most social science construct are defined by theoretical abstractions that are created to represent complex and multifaceted concepts that cannot be directly observed or measured, such as personality, polarized language, or ideology (Binning, 2023).

Usually, researchers should start with *documenting the conceptual background* (I.1) of the construct. This may involve referencing previous definitions or research on the construct's dimensionality and its manifestation in language, drawing on existing literature. Next, researchers should provide a *justification for the operationalization* (I.2), creating a connection between the construct definition and the textual data. As an example, if researchers aim to measure polarization in political discourse, they could make the argument that heightened polarization is likely to result in the use of more distinct vocabulary, as supported by previous studies (Lowe & Benoit, 2013; Peterson & Spirling, 2018). Similarly, the justification for the operationalization could also stem from the utilization of high-quality training data, potentially allowing a neural network to replicate distinct dimensions of the construct. To test their definition empirically, researchers can also conduct *a manual precoding* (I.3) exercise where two coders can manually label a subset of the data to determine if the construct is discernible by human coders.[5] On a more general note, this first validity dimension constitutes the first critical step to establish the link between the construct and the textual data, while making sure that the theoretical construct is indeed measurable using textual methods.

Design Decisions. Based on the construct definition, researchers must then demonstrate that their design decisions are rooted in the conceptualization of the analytical construct. Crucial

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

design decisions involve, for example, selecting an appropriate text-based method, collecting the data, determining the level of analysis, and, ultimately, conducting data preprocessing. Broadly construed, design decisions should primarily be derived from the construct definition itself and its presumed manifestation in the textual data. This is to navigate through what Gelman and Loken (2016) call a "garden of forking paths" – having to take and justify multiple decisions in a setting of high researchers degree of freedom. Such justification may involve various strategies of theoretical reasoning, such as reference to the methodological literature or comprehensive discussions of unique aspects of the research design.

To start with, researchers should provide a *justification of data collection decisions* (I.4). This should include general considerations on the origin of the data, such as the level of knowledge about the author(s) of the text (i.e., whether it was written by a single author, multiple authors, a computer program, or an unknown unit), as well as the context of the data generation process (for example, the 280-character limit on Twitter data, or procedural rules for parliamentary speeches). Likewise, data collection decisions entail considerations about the data selection procedures, such as the querying of specific databases using a keyword string, as well as assessing the completeness of the dataset and checking for potential biases and inconsistencies.

Moreover, researchers should elucidate their *justification of the choice of method* (I.5). This entails justifying the overall type of method, such as rule-based, supervised, or unsupervised methods (see Grimmer & Stewart, 2013 for a general introduction), as well as describing the process used to select a specific model for analysis. Practically, researchers should also include documentation of the specific model and version, along with essential background information (e.g., details about the hardware and software setup or the timeframe during which the model was utilized). When researchers opt to utilize a publicly accessible API model, it becomes crucial to provide an in-depth explanation for the decision to favor this approach over training a model locally, given the comparatively lower adaptability associated with APIs. Consequently, researchers are required to furnish thorough documentation that outlines the proper usage of the API.

Furthermore, researchers need *justification of the level of analysis* (I.6) at which the construct should be measured, such as the sentence, paragraph, document, or corpus level. This step is important because a misalignment between the presumed manifestation of the construct and the level of measurement (as defined by the operationalization) will likely lead to inconsistent and theory-agnostic measures (Jankowski & Huber, 2022; McKenny et al., 2013).

Ultimately, researchers should outline their *justification of preprocessing decisions* (I.7), in particular discussing the motivations behind eliminating certain features from the original data (for a detailed discussion, see Denny & Spirling, 2018). Prominent preprocessing decisions might be the removal of stop words, punctuations, or phrases, lowercasing and stemming words to their most basic form, or the inclusion of n-grams.

Structural Phase

In the structural phase, researchers should then conduct validation steps to examine and evaluate the properties of the model and its output. The objective of the structural phase is to gain a deeper understanding of how the measurement process functions, as well as to identify any biases or errors introduced by the computational workflow. This may require an iterative procedure, where the initial model or features of the textual data are adapted to improve the measurement and remove systematic biases. Ultimately, after conducting validation steps in the structural phase, researchers should be able to demonstrate that their measurement model is consistent and comprehensible, and that the model is potentially ready to be tested on information unrelated to the training and optimization process (see external phase).

Table 2 displays the validation steps in the structural phase, which are grouped according to the model properties and the model output. However, it is important to note that model properties and output are closely related and provide two complementary perspectives on the measurement process, as the output directly relates to the model features and characteristics.

The validation procedures that occur during the structural phase differ considerably among the groups of methods and provide only a non-exhaustive list of possible validation steps.

Table 3: Validation Steps for the Structural Phase

ID	Validation Steps	Considerations	Performance Criteria	A	B	C	D	E	F	G
Model Feature Inspection										
II.1	Inspection of predictive features	Have I considered interpretable machine learning techniques such as LIME, ICE, or partial dependence? Have I considered the interpretability and relevance of the top-ranked features?	Subjective assessment	-	R	R	R	R	-	-
II.2	Inspection of scaling word weights	Have I interpreted the word weights in regard to the interpretation of my latent scale?	Subjective assessment	-	-	-	-	-	R	-
II.3	Inspection of top-ranked words in each topic	Have I inspected the top-ranked words in each topic and assessed their plausibility?	Subjective assessment	-	-	-	-	-	-	R
Model Metrics Evaluation										
II.4	Inspection of classifiers confidence strengths	Have I assessed the model's confidence level on individual predictions?	Subjective assessment	-	O	O	O	O	-	-
II.5	Evaluation of topic coherence metrics	Have I calculated coherence metrics, such as C_v, C_p, C_uci, or C_umass, and compared their consistency and correlation?	Metric comparison	-	-	-	-	-	-	R
II.6	Evaluation of number of (N-)tokens matched	Have I evaluated the number of matched tokens for each text? Have I set a minimum share of matched words / (n-)tokens for the text to be included in the analysis?	Threshold achievement	R	-	-	-	-	-	
Output Inspection										
II.7	Visual inspection of output	Have I visualized my output? Have I identified and visualized outliers and extreme values?	Subjective assessment	R	R	R	R	R	R	R
II.8	Visual inspection of measures over time	Have I plotted the temporal trends of my measures and assessed their stability and consistency over time?	Subjective assessment	O	O	O	O	O	O	O
II.9	Comparison of aggregated measures across known groups within the data	Have I compared the aggregated measures across known groups (e.g., across data characteristics or subsets of the data)?	Subjective assessment	R	R	R	R	R	R	

II.10	Comparison of data features for clusters of closely related measures	Have I compared important data features, such as the average length of text or how often certain words appear together, across texts with similar scores (e.g., same classes on a discrete scale or high/low values on a continuous scale)?	Subjective assessment	O	O	O	O	O	O	O
II.11	Reading top documents with the highest overall scores for each output category	Have I read the most outstanding documents for each type of output, such as for distinct groups or topics, or highest and lowest scores on a numerical scale?	Subjective assessment	R	R	R	R	R	R	R
Error Analysis										
II.12	Conducting quantitative error analysis using data grouping	Have I conducted an error analysis to compare the performance of my model across known subgroups?	Subjective assessment	R	R	R	R	R	R	O
II.13	Conducting qualitative error analysis of outstanding or deliberatively chosen observations	Have I conducted an error analysis to qualitatively evaluate the sources and types of errors associated with the measures?	Subjective assessment	R	R	R	R	R	R	O
Systematic Testing										
II.14	Conducting functional tests	Have I designed and conducted functional tests (i.e., manually prepared text samples) to evaluate the model's ability to detect specific patterns in a realistic or simulated scenario?	Metric assessment	O	O	O	O	O	O	O
II.15	Conducting adversarial or counterfactual tests	Have I designed and conducted adversarial or counterfactual tests to ensure that my model is sensitive to changes in the text data?	Metric assessment	O	O	O	O	O	O	O
II.16	Conducting computational text intrusion tasks	Have I designed and conducted computational intrusion tasks to whether the model is able to recognize texts unrelated to the construct of interest (e.g., by assigning low scores on the construct)?	Metric assessment	O	O	O	O	O	O	O
II.17	Conducting word intrusion tasks for human coders in top-words per topic	Have I designed and conducted word intrusion tasks to evaluate human coders' ability to identify intruder words among the top words in a topic?	Metric assessment	-	-	-	-	-	-	O
II.18	Back-coding of topics based on top-words per topic	Have I conducted back-coding by human coders to evaluate the interpretability and validity of the topics generated by the model?	Agreement between coders	-	-	-	-	-	-	O

Columns: A = Use Case “A. Applying Dictionaries”, B = Use Case “B. Classification using Traditional Supervised Machine Learning Model”, C = Use Case “C. Classification using Finetuned Machine-Learning Model”, D = Use Case “D. Zero-Shot/Few-Shot Classification (known output categories)”, E = Use Case “E. Zero-Shot/Few-Shot Classification (unknown output categories)”, F = Use Case “F. Text Scaling”, G = Use Case “G. Topic Modelling”
Values: R = Recommended, O = Optional, - = not applicable for this type of method;
An interactive version of the checklist is available online: https://lukasbirki.shinyapps.io/ValiTex/

Model Properties

To start with, researchers should conduct validation steps that relate directly to the model properties, namely model feature inspection and model metrics evaluation.

Model feature inspection. The goal of model feature inspection is to assess whether the characteristics and features of the model are plausible indicators for the construct. For instance, a model aiming to measure populism should prioritize meaningful words like "elite" or "establishment," rather than irrelevant ones unrelated to the construct of interest. Whenever possible, validation should therefore include the inspection of the learned weights of features and distributions (see Linardatos et al., 2020; Molnar, 2020). Concrete examples include *inspecting predictive features* (II.1) using methods such as LIME, ICE, or partial dependence (SU), *inspecting scaling word weights* (II.2) for text-scaling methods, or *inspecting top-ranked words for each topic* (II.3) for dictionaries.

Model metrics evaluation. Moreover, researchers should ensure that method-specific metrics and common thresholds are met. Whereas quantitative metrics of model performance can never function as a sufficient criterion to justify validity (in particular because these metrics tend to be indifferent to the actual content of the texts (Rüdiger et al., 2022)), they still present necessary conditions for valid measures.

Whereas these metrics might differ significantly across methods applied, researchers should *inspect and compare topic coherence metrics* (II.4) for topic models (UST), *inspect the number of (N-) tokens matched* (II.5) *by* a dictionary (DI), or *inspecting confidence strength* (II.6) of supervised methods (SU).

Model Output

For the model output, researchers should conduct validation steps that relate to output inspection, error analysis, and systematic testing.

Output inspection. Researchers should conduct validation steps to ensure that the measures and their descriptive statistics look plausible. For text analysis, this form of validation steps is often referred to as "face validity" as it solely includes argument-based validation steps that require no formal analysis involved (Goet, 2019; Temporão et al., 2018).[6]

For example, this should include the *visual inspection of the output* (II.7), in particular outliers and extreme values to get a feeling on the distribution of scores and to identify outliers which might hint on inconsistencies. If possible, researchers might also *visually inspect measures over time* (II.8) to assess their stability and consistency. Furthermore, researchers should also inspect the grouped output visually. On the one hand side, this could include the *comparison of aggregated measures across known groups* (II.9), such as comparing mean ideology scores across politicians. On the other hand, researchers should also *compare data features for clusters of closely related measures* (II.10). Examples of this might include to compare features such as text length or most frequently used words across groups of texts with similar output measures (e.g., (e.g., same classes on a discrete scale or high/low values on a continuous scale).

Likewise, it is recommended to *read the top documents with the highest overall scores for each output category* (II.11), such as for distinct groups or topics, or highest and lowest scores on a numerical scale.

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

Error analysis. Additionally, researchers should apply error analysis to ensure that systematic biases and errors are considered and evaluated. We here use error analysis as an umbrella term for a set of exploratory analyses that attempt to analyze where errors in the text analysis pipeline emerge and how they might affect the outcomes (Wu et al., 2019). Accordingly, error analysis usually relies on some form of labelled data, where the labels might be produced specifically for the error analysis, or be part of the validation or (potentially) test set.[7]

For the application of error analysis, ValiTex distinguish between two major strategies. On the one hand, researchers can *conduct quantitative error analysis using data grouping* (II.12) to identify especially problematic categories (per-label performance) (Alsallakh et al., 2014). On the other hand, researchers *conduct qualitative error analysis of outstanding or deliberatively chosen observations* (II.13) to manually identify problematic text characteristics (for an application, see Wadhwa et al., 2018).

Systematic testing. Additionally, researchers should ensure that the output of the model suffices further semantic and computational tests. These systematic tests may encompass various strategies and experimental approaches, all aimed at demonstrating that the model can generate meaningful measures for the construct being measured.

For example, researchers can *conduct functional tests* (II.14), which consist of intentionally designed test cases (Röttger et al., 2021). Functional tests aim to provide more focused diagnostics by deliberately creating ambiguous texts that could be challenging for the model to interpret. Consequently, a high number of incorrect predictions on these test cases can reveal limitations and systematic biases that may affect the model's validity (Gardner et al., 2020). Likewise, researchers

[7] When labels from the independent test set are used to adapt the model properties following an error analysis, this carries the danger of imposing relevant information from the test set into the training process (Raschka, 2020). Therefore, it should generally be avoided to use the test set more than once to avoid introducing bias when estimating the model performance. In practice, however, error analysis is often conducted after evaluating the performance on the test set.

can also create *adversarial or counterfactual tests* (II.15), i.e., strategically synthesized versions of the original texts to put the model to test. For example, adversarial examples could include changing the polarity of sentences, or adding negations (Ilyas et al., 2019). Thus, increased vulnerability to these adversarial changes could highlight important shortcomings and, thus, threats to validity. Another approach researchers can take is *conducting computational text intrusion tasks* (II.16). These tests rely on the assumption that a model should correctly label "intrusive" texts, i.e., texts from a different sample with distinct characteristics (Huang et al., 2020).

Likewise, more method-specific tasks are available, such as *conducting word intrusion tasks to top-words for specific topics* (II.17) to evaluate whether a topic has human-identifiable semantic coherence (Chang et al., 2009). Another test for topic models could be to ask human coders to *back-code topics based on top-words per topic* (II.18) to test for semantic coherency of topics.

External Phase

In the external phase, researchers should ultimately test for how the measures relates to independent information or criteria. During the external phase, information outside the scope of the textual data in which the measure was constructed thus serves as an external benchmark (hence “external” phase). Such external data can either be independent measures of the same construct (measure interrelation) or external criteria which one expects the text-based measure to be able to predict (criterion prediction). Because the primary justification for using computational text analysis methods is empirical (i.e., scalability and efficiency, see Grimmer et al., 2022), external validation ultimately demonstrates a robust relationship with variables unrelated to the respective measurement model. However, it is important to note that external validation alone is not sufficient because even a model with strong biases (lacking substantive and structural evidence) may exhibit meaningful but misleading associations with external criteria. Hence, focusing excessively on output optimization might result in overlooking errors without a complete understanding of their nature. Furthermore, to establish external validation as a reliable source of evidence for validity, researchers must clearly state their assumptions about how their measures should correlate with other metrics and external benchmarks. Ideally, this should involve preregistering hypotheses and defining the expected magnitude and direction of correlations that would support the measure's validity.

Table 3 displays the validation steps in the external phase.

Table 4: Validation Steps in the External Phase

ID	Validation Steps	Considerations	Performance Criteria	A	B	C	D	E	F	G
Measure Interrelation										
III.1	Comparison of measures with human-annotated test set ("gold-standard data")	Have I labelled a subset of the data using a codebook or pairwise comparison method to serve as the gold standard for evaluation?	Correspondence to human-annotated test set	R	R	R	R	R	R	R
III.2	Comparison of measures with surrogate labels	Have I collected or generated surrogate labels (e.g., expert surveys, contextual labels) as another benchmark for evaluation?	Correspondence to surrogate labels	O	O	O	O	O	O	O
Criterion Prediction										
III.3	Prediction of external criteria or real-world phenomena	Have I formulated expected relationship of my measures with external criteria? Have I confirmed these relationships empirically?	Correspondence to external criteria	O	O	O	O	O	O	O

Columns: A = Use Case "A. Applying Dictionaries", B = Use Case "B. Classification using Traditional Supervised Machine Learning Model", C = Use Case "C. Classification using Finetuned Machine-Learning Model", D = Use Case "D. Zero-Shot/Few-Shot Classification (known output categories)", E = Use Case "E. Zero-Shot/Few-Shot Classification (unknown output categories)", F = Use Case "F. Text Scaling", G = Use Case "G. Topic Modelling"
Values: R = Recommended, O = Optional, - = not applicable for this type of method;
An interactive version of the checklist is available online: https://lukasbirki.shinyapps.io/ValiTex/

Measure Interrelation

To assess measure interrelations, researchers should conduct validation steps that relate to human-annotated test set comparison and optionally surrogate label comparison.

Human-annotated test set comparison. On the one hand side, researchers should *compare measures with a human-annotated test set* (III.1), often referred to as benchmark or "gold-standard" data (Lowe & Benoit, 2013).[8] Generally, human annotations are considered the most reliable form of validation for text-based research, regardless of the measurement design. However, the quality of manual annotations can also be compromised, potentially due to misinterpretations, the presence of complex categories, and insufficient coder training (Grimmer et al., 2022). While a detailed discussion on how to best conduct human-annotation is beyond the scope of this paper, the dynamic checklist offers additional references to methodological standards and best practices, including coding procedures (such as developing a codebook or using pairwise comparison) and reporting various classification metrics (such as accuracy, precision, recall and F1-score).

Surrogate label comparison. Alternatively, researchers might further *compare measures with surrogate labels* (III.2) that constitute similar, but closely related characteristics of the data, such as the party label as a proxy measure for ideology. However, it needs to be justified that the surrogate labels are good proxies.

Criterion Prediction

Criterion Prediction. Likewise, researchers might also include validation steps that include *prediction of external criteria or real-world phenomena* (III.3) (called "external criteria" in the psychometric tradition). Validation steps in this category include tests on whether the measures confirm presumed relationships regarding real-world phenomena, such as the prediction of voting behavior based on textual measures of political ideology (Lauderdale & Herzog, 2016; Rheault & Cochrane, 2020) – much like work performance or school achievement have served as external criteria for tests of cognitive ability or personality in the psychometric tradition.

[8] In their book, Grimmer et al. 2022, more rigorously, distinguish between validation with and without gold-standard data, arguing that a strong relationship between empirical measures and gold-standard data might provide sufficient validation evidence after all. We take a slightly different perspective, arguing that gold-standard labels for text-based measures of social science construct is hardly attainable and, thus, comparison with a human-annotated test set should always be complemented by other validation steps in the substantive and structural phase.

Robustness Checks

Next to the three phases of the validation process outlined above, one fundamental principle of the ValiTex framework also includes the continuous test of robustness checks to assess the impact of researchers' degree of freedom. On a general note, one could see robustness checks as additional means to test whether design decisions in the substantive phase might have a sustainable effect on the measure's outcome. Table 5 provides an overview over relevant aspects which might be tested. For instance, researchers should r*erun the analysis using different preprocessing steps* (IV.1). As Denny and Spirling (2018) demonstrate, the choices made during data pre-processing have a significant impact. Given that many pre-processing decisions rely on established practices and general rules developed over time, demonstrating the robustness of the output in relation to these decisions is vital.

In addition, robustness checks should also include *rerunning the analysis using different hyperparameter settings* (IV.2). Reporting and reflecting on hyperparameters, that is those settings that help specify a respective text model, is a crucial part of any text model (Arnold et al., 2023).

Moreover, researchers should also rer*un their analysis using alternative text-based measures* (IV.3). For instance, when applying a dictionary, one can often find related dictionaries which might be applied to compare changes to on the outcomes of the measurement (van Atteveldt et al., 2021).

Table 5: Robustness Checks

ID	Validation Steps	Considerations	Performance Criteria	A	B	C	D	E	F	G
IV.1	Rerunning the analysis using different preprocessing steps	Have I rerun the analysis using different preprocessing settings (e.g., stop word removal, stemming, lemmatization)?	Change to previous measurement outcome	R	R	R	R	R	R	R
IV.2	Rerunning the analysis using different hyperparameter settings	Have I rerun the analysis using different hyperparameter settings?	Change to previous measurement outcome	R	R	R	R	R	R	R
IV.3	Rerunning the analysis using alternative text-based methods	Have I rerun the analysis with alternative text-based methods?	Change to previous measurement outcome	O	O	O	O	O	O	O
IV.4	Rerunning the analysis with different levels of aggregation	Have I replicated the same study using different levels of aggregation (e.g., token, word, sentence, paragraph, document level)?	Change to previous measurement outcome	O	O	O	O	O	O	O
IV.5	Rerunning the analysis with a different, but related dataset	Have I replicated the same study using a different, but related dataset?	Change to previous measurement outcome	O	O	O	O	O	O	O
IV.6	Rerunning the analysis using different subsets of the data	Have I rerun the analysis using different subsets of the data?	Change to previous measurement outcome	O	O	O	O	O	O	O
IV.7	Rerunning the analysis using different thresholds	Have I rerun the analysis using different thresholds (e.g., min. number of tokens matched, max. document frequency)?	Change to previous measurement outcome	O	O	O	O	O	O	O
IV.8	Rerunning the analysis using different number of topics	Have I rerun the analysis using different number of topics?	Change to previous measurement outcome	-	-	-	-	-	-	R

Columns: A = Use Case “A. Applying Dictionaries”, B = Use Case “B. Classification using Traditional Supervised Machine Learning Model”, C = Use Case “C. Classification using Finetuned Machine-Learning Model”, D = Use Case “D. Zero-Shot/Few-Shot Classification (known output categories)”, E = Use Case “E. Zero-Shot/Few-Shot Classification (unknown output categories)”, F = Use Case “F. Text Scaling”, G = Use Case “G. Topic Modelling”
Values: R = Recommended, O = Optional, - = not applicable for this type of method;
An interactive version of the checklist is available online: https://lukasbirki.shinyapps.io/ValiTex/

Ultimately, it is crucial to evaluate various other aspects of the analysis. For instance, this could involve rerunning the analysis with *different levels of aggregation* (IV.4) (e.g., from paragraph to document level), *a different, but related dataset* (IV.5) *(e.g.,* party manifestos from another election), *different subsets of the data* (IV.6) (e.g., only responses from a specific country), *different thresholds* (IV.6) (e.g., setting the minimum share of tokens matched by a dictionary for a document), or *different number of topics* (IV.7) for a topic model.

Illustrative Example: Measuring Sexism in Social Media Comments

After presenting the ValiTex framework, we now demonstrate its practical application through an illustrative example of measuring sexism in social media comments. This example highlights how the terminology and taxonomy of validation step proffered by the ValiTex framework and its attendant checklists, contribute to a more systematic, rigorous, and easily understandable approach to validation. When researchers apply the framework and checklist, they can effectively navigate the process of validating their measures for a specific study. Moreover, ValiTex assists research consumers in assessing the presented validity evidence. It is essential to acknowledge that each validation step reported in a study contributes to reducing uncertainty about the measure's validity, thus enhancing the trustworthiness of the study's substantive findings. Toward that end, we revisit and document validation steps in a study conducted by Samory et al. (2019) that measured sexism in social media comments. Specifically, this study used supervised machine learning to measure sexism for different datasets. We chose the study by Samory et al. (2019) because of their comprehensive exploration and contemplation of the intricacies tied to measuring sexism. We proceed by scrutinizing the validation steps employed in their study, aligning them with the conceptual framework and the checklist for “C. Classification using Finetuned Machine-Learning Model” outlined in ValiTex. The filled-out checklist containing the different validation steps can be found in the Appendix 1.

Substantive Phase

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

Afterwards, they provide a *justification of the operationalization* (I.2)**,** that is, the link between their construct and the textual data. Because they primarily apply supervised machine-learning methods, the authors' underlying justification lies in the provision of high-quality training data that enables the text model to autonomously learn and adapt relevant patterns in the data. To annotate sexism within their training data, the authors develop a detailed codebook based on four subdimensions of sexism identified in the previous literature (e.g., "behavioral expectations" and "endorsement of inequality"). In addition, they extend the codebook to not only differentiate between sexist content, but also between varying degrees of sexist phrasing.

To test their operationalization using a codebook empirically, they rely on *human precoding* (I.3) to label the sexist items from the survey scales into relevant subcategories, finding considerable agreement.

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

Furthermore, for the *justification of method choice* (I.5), the authors rely primarily on a supervised approach. Although implicitly, their argumentation is that only supervised models can replicate human coding's that distinguish between different subdimensions of sexism. However, they do not only rely on one specific type of method but rather select a variety of models, such as a Logit model, a CNN and a BERT finetuned model with increasing complexity to systematically compare their performance, allowing for a thorough evaluation of measurement performance across different models. Likewise, they also decide to include dictionary baseline models (see the section on robustness checks *rerunning the analysis using alternative text-based methods* (IV.3)) provide a rationale for selecting different models.

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

Structural Phase

During the structural phase, Samory et al. (2021) proceed with a combination of validation steps to examine and evaluate the properties of the model and its output. To evaluate the model properties, they *inspect of predictive features* (II.1). To do so, they evaluate the most predictive words for each sexism category (unigrams) and compare them across their data sets and methods applied. Thus, they observe that some models, which are trained on slightly adapted adversarial examples (see Wallace et al., 2019; Zhang et al., 2019), exhibit more general features, which indicates increased model robustness and, partially, performance.[9]

To evaluate the model output, they furthermore conduct extensive *error analysis using data grouping* (II.12) of misclassified examples to identify systematic errors on their most promising BERT model. They specifically investigate the influence of various factors to assess where the model misclassifies messages using a logistic regression model. Relevant factors which they

[9] Since they hand-annotated all their data using the codebook, they simultaneously report on the model overall performances on the hand-annotated test sets. This highlights the interconnected nature of validation steps in both the structural and external phases, necessitating an iterative and reciprocal approach between the two phases.

consider are the type of model used (i.e., whether it was trained on original or adversarial examples), or the origin of the training data. Furthermore, they also evaluate the impact of the initial agreement among the coders on the probability of misclassifying errors.

External Phase

In the external phase, Samory et al. (2019) primarily rely on *comparing their measures human-annotated test set* (III.1). To calculate classification performance, they apply k-fold cross-validation and report F1 scores, which are calculated from the precision (i.e., the accuracy of positive predictions) and recall (i.e., the completeness of positive predictions) of their measures. The evaluation of F1 score is widely regarded as the most viable metric, as alternative metrics such as accuracy (i.e., the overall ratio of positive predictions) can be misleading when dealing with imbalanced data (Spelmen & Porkodi, 2018). Performance metrics show that the fine-tuned BERT model that incorporates the adversarial examples into the training process achieves on average the best performance across all datasets (F1 > 0.8).

Robustness Checks

Throughout the three validation phases, Samory et al. (2019) conduct a series of robustness checks. Thus, some of these robustness checks might not represent distinct validation steps per se but are closely connected to the overall validation process as outlined in this chapter. For instance, the authors *rerun their analysis using alternative text-based methods* (IV.3). In particular the inclusion of a simple baseline model (toxicity and gender word dictionary) demonstrates that their supervised models achieve higher overall performance on the hand-annotated text set, providing evidence that their supervised models are more successful in replication human-annotated measures of sexism. However, F1 scores of the baseline models are only reported on average across all datasets, making a comparison of performance for specific datasets difficult.

Likewise, the authors also evaluate their measurement by *rerunning their analysis with different, but related datasets* (IV.5). They rely on Twitter datasets collected by different researchers and sampling strategies, as well as data from psychological survey scales that are designed to measure sexism. Moreover, in the classification process, they *rerun their analysis using different subsets of the data* (IV.6), trying out different combinations of the datasets while evaluating the overall performance on the hand-annotated test set.

As our discussion draws to a close, this example highlights how ValiTex can simplify the process of explaining validation. It achieves this by offering step-by-step guidance through a logical framework consisting of three distinct phases and a consistent taxonomy of empirical validation steps. Through the analysis of the validation process presented by Samory et al., we gain a more comprehensive understanding of their validation approach.

During the substantive phase, it became evident that the authors deeply engaged with the conceptual background related to measuring sexism in text. In the structural phase, they documented significant validation steps such as *inspection predictive features* (II.1) and *error analysis using data grouping* (II.12). However, they lacked to report validation steps connected to output inspection, such as providing a *visual inspection of the output* (II.7) or *reading top documents with the highest overall scores for each output category* (II.11). Moving on to the external phase, the authors primarily relied on comparing their measures with a meticulously coded human-annotated test set. Additionally, for the purpose of robustness checks, they employed a variety of models and datasets. This decision was a direct outcome of their comparative research design, which involved the integration of multiple methods and data sources.

Critical Reflection

We developed the ValiTex framework with the intention of tackling the manifold challenges associated with validating text-based measures of social science constructs. More specifically, we designed our framework to provide a comprehensive and uniform perspective on validity, and to equip scholars with the vocabulary and methodological tools to validate text-based measures. We foresee that ValiTex will help researchers to validate text-based measures in a more rigorous and systematic way, and to decrease the uncertainty surrounding the findings obtained from text-based measures.

The envisioned role of our framework encompasses at least two main aspects. Firstly, it can function as a "methodological tool," offering structured guidance for researchers to navigate validation. By providing a conceptual framework and listing validation steps available for researchers, ValiTex aims to provide clear guidelines for thorough validation. Secondly, ValiTex can also serve as a documentation scheme, streamlining communication between researchers engaged in validation and research consumers. More generally, our aim is for the framework to enhance the quality of conversations surrounding the validation of text-based measures, ultimately leading to a more refined conceptual understanding.

To advance the overall quality of validation practices beyond the framework proposed, however, we want to emphasize some general conditions for successful validation in greater detail.

Transparency. First and foremost, we want to stress the role of transparency in the validation process. Given that computational text-based measures are susceptible to errors and biases inherent to the models themselves (Abid et al., 2021; Liang et al., 2021), we argue that researchers should openly acknowledge and embrace the potential limitations and biases of their measurement methods. By doing so, they can make their decisions regarding the measurement process transparent, enabling a deeper comprehension and recognition of the challenges and uncertainties that come with text-based measures and their validation for both researchers and readers.

In practical terms, transparency should entail rigorously documenting the validation process and providing access to the data and code utilized. Fortunately, many scholars in the text-as-data domain are already at the forefront of this development. In the realm of large language models, for example, there is a considerable drive towards documentation standards, both for the textual data (Bender & Friedman, 2018; Gebru et al., 2021; Heger et al., 2022) and the respective measurement model and its parameters (Derczynski et al., 2023; Dodge et al., 2019; Rogers et al., 2021). Moreover, to address the inherent research freedom associated with text analysis, researchers should anchor their work within robust open science standards, particularly embracing preregistration and registered reports, while also establishing the necessary infrastructure to facilitate replication studies (Schoch et al., 2023).

Human-in-the-loop. Second, we are convinced that, at least for now, any computational text-based method should still rely on human semantic understanding as an absolute benchmark (van Atteveldt et al., 2021; Weber et al., 2018). In ValiTex, each phase incorporates recommended validation steps that directly rely on human judgment. Human judgement is crucial because too often, computational methods are prone to rely on spurious relations or noise in the data, thereby lacking a deeper ontological sense of error which should prevent trusting empirical measures blindly (see Jankowski & Huber, 2022). Thus, we argue that the role of human judgement is not only crucial for more simple textual methods, such as dictionaries or topic models, but also for validating measures derived from large language models, as human annotations are especially relevant in providing the models with high-quality labels to evaluate performance. However, human annotations do not automatically lead to accurate labels for (training and) testing models. On the contrary, without adequate guidance and training, human annotation bias might results in low-quality data labels (see Geva et al., 2019; Liu et al., 2021), which might be further aggravated by coders' limited focus, fatigue, and evolving interpretations of the underlying conceptual categories (Neuendorf, 2017). Nevertheless, until today, human annotations remain the most dependable form of annotation that is not easily replaced in the validation process.[10]

Adaptability and Flexibility. Although our proposed framework offers a general validation workflow that will be broadly applicable, there may be a need to customize validation strategies and have differing viewpoints on what constitutes satisfactory validation evidence in specific studies. This follows from the intricate and multifaceted nature of textual data, which makes certain validation steps inapplicable in some specific research context. For example, researchers frequently encounter language-specific obstacles that are unique to their specific dataset, such as linguistic idiosyncrasies or dialect variations (Larina, 2015). Similarly, the way meaning is encoded can vary significantly between different languages in multilingual settings (for a detailed discussion of multilingual challenges in validation, see Lind et al., 2023). Because no single framework can anticipate all possible scenarios, researchers must be able to adapt recommended practices to sufficiently validate their methods.

This also applies to determining appropriate cut-off values and metrics, which unfortunately lack universal interpretability. As a case in point, consider the two commonly used metrics Krippendorff's alpha (Krippendorff, 2018) for inter-coder agreement and F1 score as a measure of a model's classification accuracy. Whereas general rules of thumb exist for both metrics, with values above 0.8 generally considered good, and values below 0.67 (alpha) and 0.5 (F1) considered poor, both metrics depend heavily on the nature of the construct and contextual factors within the

[10] Recently, scholars have started to supplement human annotation by automated annotation (Huang et al., 2023; Wang et al., 2021). However, it is yet unclear whether automated methods are able to reach comparable semantic understanding similar to human.

measurement design.[11] Thus, researchers must be able to interpret and adapt their specific cut-off values, such as by comparing it to related work or picking the best-performing model out of several competing models.[12]

In sum, we do not claim that our proposed framework provides a "holy grail" that sweeps away all concerns in the process of validating text-based measures. Nevertheless, we emphasize that our framework offers a much-needed structure for discussions on validity, benefiting both researchers conducting validation and research consumers seeking solid documentation. Likewise, we hope that our taxonomy of recommended and optional validation steps provides a starting point for a more normative discussion on validation. Because, as evidenced by the decade-long history of validity in the social sciences, methodological discussions concerning validation commonly echo dominant perspectives and best practices within research communities. Therefore, we strongly advocate for community efforts and critical discussion to agree on generally acknowledged standards for the validation of text-based measures of social science constructs.

[11] For instance, it might be comparatively easier to identify general sentiments than complex constructs, such as chauvinism.

[12] This, however, caries the danger of deliberatively picking the best model, regardless of theoretical considerations and, potentially, variability in the empirical results derived from empirical measures (for a detailed discussion, see C. Chan et al., 2021).

Conclusion

Validating computational text-based measures of social science constructs can be tedious. Whereas convincing validation evidence can increase the trustworthiness of empirical measures, the uncertainty surrounding the true value of the underlying constructs can never be eliminated. However, this should not be interpreted as text-based measures being inferior to other forms of social science data. On the contrary, we contend that conducting promising text-based research is very much possible, particularly when validation evidence is presented in a structured and compelling manner. In this regard, we hope that ValiTex can assist researchers in this endeavour and advance the overall implementation of validation practices in social science research. Because increased confidence about the validity of text-based measures will not only aid researchers in presenting their research, but also enable various stakeholders – from public institutions to individual researchers– to evaluate and rely on empirical evidence, thus enhancing the overall credibility and impact of text-based research.

Acknowledgments:

We want to thank Thomas Knopf, Matthias Blümke, Benjamin Guiddenau, Leon Fröhling, Arthur Spirling, Verena Kunz, and Martijn Schoonvelde for their insightful feedback.

Disclosure statement:

No potential conflict of interest was reported by the author(s).

Funding:

This research is partly supported by the Leibniz-Foundation Grant “Digital Dehumanization: Measurements, Exposure and Prevalence”

Appendix

Appendix 1

Filled Out Validation Steps for the study by Samori et al. (2019)

This section needs to be updated. If you are interested in a preliminary version, please contact Lukas.birkenmaier@gesis.org

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

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

Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., & Rauber, A. (2014). Visual methods for analyzing probabilistic classification data. *IEEE transactions on visualization and computer graphics*, *20*(12), 1703–1712.

Arnold, C., Biedebach, L., Küpfer, A., & Neunhoeffer, M. (2023). The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*.

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.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Binning, J. (2023). Construct. In *Brittanica (Online Encyclopedia)*. https://www.britannica.com/science/construct

Birkenmaier, L., Lechner, C. M., & Wagner, C. (2023). *The search for solid ground in text as data: A systematic review of validation approaches* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/kgn6f

Boxman-Shabtai, L. (2020). Meaning Multiplicity Across Communication Subfields: Bridging the Gaps. *Journal of Communication*, *70*(3), 401–423.

Cai, W., Encarnacion, R., Chern, B., Corbett-Davies, S., Bogen, M., Bergman, S., & Goel, S. (2022). *Adaptive Sampling Strategies to Construct Equitable Training Datasets* (arXiv:2202.01327). arXiv. http://arxiv.org/abs/2202.01327

Cattell, R. B. (1946). *Description and measurement of personality*. World Book Company.

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, E. K. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. *Validity and validation in social, behavioral, and health sciences*, 9–24.

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

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

Cureton, E. (1951). Validity. In *Educational Measurement*. American Council on Education.

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). *Assessing Language Model Deployment with Risk Cards* (arXiv:2303.18190). arXiv. https://doi.org/10.48550/arXiv.2303.18190

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

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

Drisko, J. W., & Maschi, T. (2016). *Content analysis*. Pocket Guide to Social Work Re.

Fang, Q., Nguyen, D., & Oberski, D. L. (2022). Evaluating the Construct Validity of Text Embeddings with Application to Survey Questions. *arXiv:2202.09166 [cs, stat]*. http://arxiv.org/abs/2202.09166

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

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., & Gottumukkala, A. (2020). Evaluating Models' Local Decision Boundaries via Contrast Sets. *arXiv preprint arXiv:2004.02709*.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). *Datasheets for Datasets* (arXiv:1803.09010). arXiv. http://arxiv.org/abs/1803.09010

Gelman, A., & Loken, E. (2016). The statistical crisis in science. *The best writing on mathematics (Pitici M, ed)*, 305–318.

Geva, M., Goldberg, Y., & Berant, J. (2019). *Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets* (arXiv:1908.07898; Version 1). arXiv. http://arxiv.org/abs/1908.07898

Gibney, E. (2022). Could machine learning fuel a reproducibility crisis in science? *Nature*, *608*(7922), 250–251. https://doi.org/10.1038/d41586-022-02035-w

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

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and Combining Sentiment Analysis Methods. *Proceedings of the first ACM conference on Online social networks - COSN '13*, 27–38. https://doi.org/10.1145/2512938.2512951

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.

Heger, A. K., Marquis, L. B., Vorvoreanu, M., Wallach, H., & Wortman Vaughan, J. (2022). Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–29. https://doi.org/10.1145/3555760

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

Hughes, D. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In *The Wiley handbook of psychometric testing: A multidisciplinary approach to survey, scale and test development*. John Wiley & Sons Ltd.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, *32*.

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kantner, C., & Overbeck, M. (2020). Exploring soft concepts with hard corpus-analytic methods. *Reflektierte algorithmische Textanalyse. Interdisziplinäre (s) Arbeiten in der CRETA-Werkstatt*, 169–189.

Kapoor, S., & Narayanan, A. (2022). *Leakage and the Reproducibility Crisis in ML-based Science* (arXiv:2207.07048). arXiv. https://doi.org/10.48550/arXiv.2207.07048

Kelley, T. L. (1927). *Interpretation of educational measurements* (S. 353). World Book Co.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

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

Larina, T. (2015). Culture-specific communicative styles as a framework for interpreting linguistic and cultural idiosyncrasies. *International Review of Pragmatics*, *7*(2), 195–215.

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

Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models. *International Conference on Machine Learning*, 6565–6576.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lind, F., Schoonvelde, M., Baden, C., Dolinsky, A., Pipal, C., Shababo, G., & van der Velden, M. (2023). *A framework for validating multilingual computational text analysis*.

Liu, H., Thekinen, J., Mollaoglu, S., Tang, D., Yang, J., Cheng, Y., Liu, H., & Tang, J. (2021). *Toward Annotator Group Bias in Crowdsourcing* (arXiv:2110.08038; Version 1). arXiv. https://doi.org/10.48550/arXiv.2110.08038

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports*, *3*(3), 635–694.

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

Mayring, P. (2004). Qualitative content analysis. *A companion to qualitative research*, *1*(2), 159–176.

McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods*, *16*, 152–184. https://doi.org/10.1177/1094428112459910

Messick, S. (1989). Validity. In *Educational measurement, 3rd ed* (S. 13–103). American Council on Education.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Neuendorf, K. A. (2017). *The content analysis guidebook*. sage.

Peterson, A. J., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, *26*(01), 120. https://doi.org/10.1017/pan.2017.39

Pipal, C., Song, H., & Boomgaarden, H. G. (2022). If You Have Choices, Why Not Choose (and Share) All of Them? A Multiverse Approach to Understanding News Engagement on Social Media. *Digital Journalism*, 1–21. https://doi.org/10.1080/21670811.2022.2036623

Raschka, S. (2020). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning* (arXiv:1811.12808). arXiv. http://arxiv.org/abs/1811.12808

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

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-based systems*, *89*, 14–46.

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

Roberts, C. W. (1989). Other than counting words: A linguistic approach to content analysis. *Social Forces*, *68*(1), 147–177.

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

Rüdiger, M., Antons, D., Joshi, A. M., & Salge, T.-O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLOS ONE*, *17*(4), e0266325. https://doi.org/10.1371/journal.pone.0266325

Rusticus, S. (2014). Content Validity. In A. C. Michalos (Hrsg.), *Encyclopedia of Quality of Life and Well-Being Research* (S. 1261–1262). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_553

Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021). "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 573–584. https://doi.org/10.1609/icwsm.v15i1.18085

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

Shepard, L. A. (1993). Chapter 9: Evaluating test validity. *Review of research in education*, *19*(1), 405–450.

Spelmen, V. S., & Porkodi, R. (2018). A review on handling imbalanced data. *2018 international conference on current trends towards converging technologies (ICCTCT)*, 1–11.

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

The Standards for Educational and Psychological Testing. (2014). Https://Www.Apa.Org. https://www.apa.org/science/programs/testing/standards 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 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

Wadhwa, S., Chandu, K. R., & Nyberg, E. (2018). Comparative analysis of neural QA models on SQuAD. *arXiv preprint arXiv:1806.06972*.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal Adversarial Triggers for Attacking and Analyzing NLP. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2153–2162. https://doi.org/10.18653/v1/D19-1221

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018). Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions. *Communication Methods and Measures*, *12*(2–3), 119–139. https://doi.org/10.1080/19312458.2018.1447656

Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. (2019). Errudite: Scalable, Reproducible, and Testable Error Analysis. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 747–763. https://doi.org/10.18653/v1/P19-1073

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2019). *Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey* (arXiv:1901.06796). arXiv. https://doi.org/10.48550/arXiv.1901.06796