**ValiText Checklist**

**Use Case C: Prompt-based classification using LLMs**

This checklist accompanies the [ValiTex](https://arxiv.org/abs/2307.02863)t framework for validating text-based measures of social constructs by Birkenmaier et al. (2024). Each row within the table corresponds to one validation step (i.e., specific tests that can be executed to produce validation evidence). As outlined in the corresponding paper, researchers should initially follow the order of the phases, starting with the substantive validation steps and ending with external validation steps while continuously considering robustness checks. However, researchers might adapt this process to their individual use case.

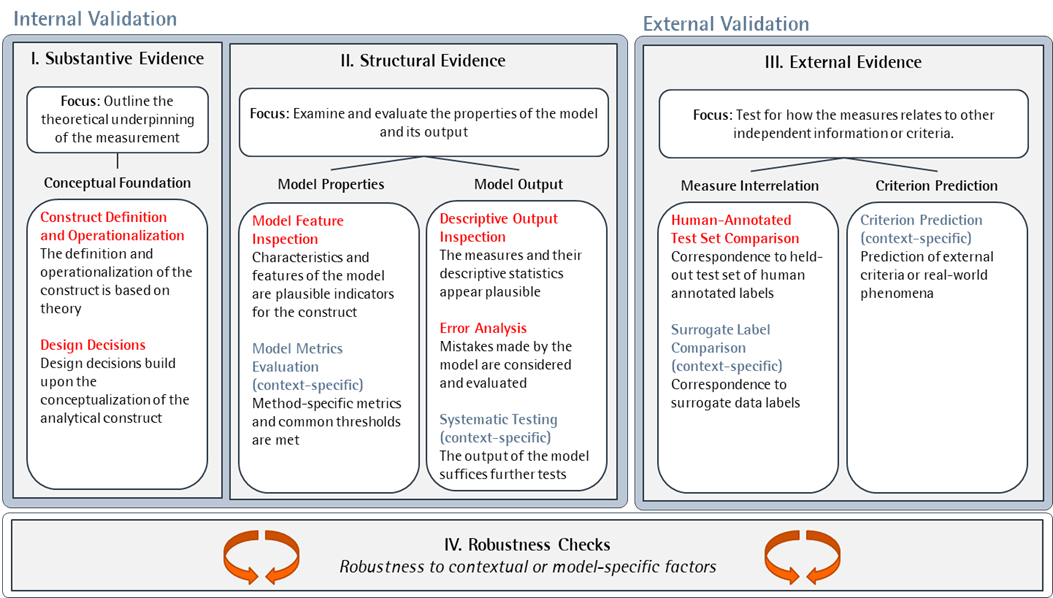


Figure 1: Framework

# Substantive Evidence

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Validation Step** | **Documentation** | **Considerations** | **Performance Criteria** | **Source / References** |
| **Construct Definition and Operationalization** | | | | | |
| I.1 | Documentation of the conceptual background |  | Have I conducted a literature review or consulted with domain experts to gain a sufficient understanding of conceptual background of the construct? | Summarizing existing literature on the conceptual background of the construct | Krippendorf (2018) |
| I.2 | Justification of the operationalization |  | Have I sufficiently explained how the construct should manifest itself in the textual data? Have I documented my operationalization in a codebook? | Providing definition and conceptualization of the construct | Krippendorf (2018) |
| I.3 | Manual Precoding |  | Have I reached sufficient interrater agreement for a subsample of the textual data? Have I ensured that the construct can be detected in the textual data? Have I outlined my rules of coding uncertainty across coders? | Reaching sufficient interrater agreement (e.g., Krippendorff’s alpha α) | Krippendorf (2018), Plank (2022) |
| **Design Decisions** | | | | | |
| C.1 | Including examples into the prompt |  | Have I supplied clear examples of the construct in the prompt? | Outlining the rationale behind the provided example | (Gilardi et al., 2023) |
| I.4 | Justification of data collection decisions |  | Have I selected a dataset that is representative and relevant to the research question and population of interest? Have I justified the data selection decisions (e.g., using keywords)? Have I assessed the quality and completeness of the dataset and checked for potential biases or inconsistencies? | Outlining the rationale behind data selection / collection decisions; Documenting potential limitations and data quality issues | Krippendorf (2018) |
| I.5 | Justification of method choice |  | Have I selected the appropriate type of method based on the operationalization of the construct and data characteristics? Have I justified the concrete selection of a particular model? | Outlining the rationale behind method selection; Documenting potential limitations in comparison to alternative methods | Grimmer et al. (2022) |
| I.6 | Justification of the level of analysis |  | Have I selected the appropriate level of analysis? Have I considered potential problems when aggregating measures from lower to higher levels (e.g., sentence to paragraph level)? | Outline the rationale behind the selected level of analysis (e.g., token, sentence, or paragraph level). | Jankowski & Huber (2022) |
| I.7 | Justification of preprocessing decisions |  | Have I justified relevant changes to the text prior to the analysis, such as removing certain words or phrases? | Outlining the rationale behind preprocessing decisions | Grimmer et al. (2022) |

# Structural Evidence

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Validation Step | Documentation | Considerations | Performance Criteria | Source / References |
| Model Feature Inspection | | | | | |
| I.1 | Inspection of predictive model features |  | Have I inspected the predictive features for my model? Have I assured they are conceptually aligned with the construct being measured? | Qualitative evaluation of top-ranked model features using feature-importance methods like e.g., LIME or ICE | Molnar (2020), Küpfer & Meyer (2023) |
| Descriptive Output Inspection | | | | | |
| II.2 | Visual inspection of output |  | Have I visualized my output descriptively? Have I identified and visualized outliers and extreme values? | Plotting descriptive statistics; discussing the plausibility of the observed distribution | Goet (2019) |
| II.3 | Comparison of aggregated measures across known groups |  | Have I aggregated the output scores across known groups (e.g., mean share of sexist sentences across social media user demographics)? | Plotting aggregated measures across groups; discussing the plausibility of the observed distribution | Goet (2019) |
| II.4 | Qualitatively assess top documents with the highest overall scores for each output category |  | Have I assessed the most outstanding documents for each type of output, such as labels with the highest confidence, or highest and lowest scores on a numerical scale? | Qualitative evaluation to ensure that the top-ranked texts align with the construct | Goet (2019) |
| Error Analysis | | | | | |
| II.5 | Error analysis using data grouping |  | Have I conducted error analysis to compare the performance of my model across known subgroups? | Comparing performance metrics (i.e., F1) across subgroups | Wu et al. (2019) |
| II.6 | Error analysis of outstanding or deliberatively chosen observations |  | Have I conducted error analysis to qualitatively evaluate the sources and types of errors associated with the measures? | Exploring the underlying causes of misclassifications by qualitatively screening misclassified examples | (Wu et al., 2019) |
| Systematic Testing (context-specific) | | | | | |
| V.1 | Counterfactual tests |  | Have I tested that my model is sensitive to meaningful changes in the text data? | Evaluating performance metrics (i.e., F1) for new dataset of counterfactual examples | (Garg et al., 2019) |
| V.2 | Adversarial tests |  | Have I tested that my model is resilient to slight perturbations in the text data? | Evaluating performance metrics (i.e., F1) for new dataset of adversarial examples | (Ribeiro et al., 2018) |
| V.3 | Discriminant tests |  | Have I tested that my model is able to distinguish between the construct of interest and similar, but unrelated concepts (e.g., and sexist language)? | Inspecting output scores for a sample of “discriminant” examples | Fang et al. (2023) |
| V.4 | Out of domain tests |  | Have I tested that my model is able to generalize to out-of-domain examples? | Evaluating performance metrics (i.e., F1) for new dataset of out-of-domain examples | (Sen et al., 2022) |

# External Evidence

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Validation Step | Documentation | Considerations | Performance Criteria | Source / References |
| Construct Definition and Operationalization | | | | | |
| III.1 | Comparison of measures with human-annotated test set ("gold-standard data") |  | Have I reached sufficient predictive performance on a test set of held-out human annotations? Did I apply cross-validation to calculate average performance metrics? | Evaluating performance metrics (i.e., F1) for dataset of human annotations | (Samory et al., 2021) |
| Surrogate Label Comparison (context-specific) | | | | | |
| V.5 | Comparison of measures with surrogate labels |  | Have I reached sufficient predictive performance on the surrogate labels? | Evaluating performance metrics (i.e., F1) for the surrogate labels | Grimmer et al. (2022) |
| Criterion Prediction (context-specific) | | | | | |
| V.6 | Criterion Prediction |  | Have I been able to accurately predict real-word phenomena? | Evaluating predictive metrics (i.e., regression coefficient) for the criteria | Grimmer et al. (2022) |

# Robustness Checks

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Validation Step | Documentation | Considerations | Performance Criteria | Source / References |
| Construct Definition and Operationalization | | | | | |
| IV.1 | Rerunning the analysis using alternative text models |  | Have I rerun the analysis with alternative text-based methods, such as a baseline model? | Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative measurements | (Samory et al., 2021) |
| IV.2 | Rerunning the analysis using different hyperparameter settings |  | Have I rerun the analysis with alternative hyperparameter settings? | Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative hyperparameter settings | (Arnold et al., 2023) |
| IV.3 | Rerunning the analysis using different cutoff thresholds |  | Have I rerun the analysis with alternative cutoff thresholds? | Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative cutoff thresholds | Grimmer et al. (2022) |
| IV.4 | Rerunning the analysis using different text cleaning and preprocessing steps |  | Have I rerun the analysis using alternative data cleaning and preprocessing settings (e.g., removing certain phrases or features of the data)? | Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative preprocessing steps | Grimmer et al. (2022) |
| C.2 | Rerunning the analysis using different prompts |  | Have I rerun the analysis with alternative prompt designs? | Displaying performance metrics (e.g., F1 score on human annotated test set (see III.1)) for alternative prompts | (Gilardi et al., 2023) |

# Literature

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.

Boukes, M., Velde, B., Araujo, T., & Vliegenthart, R. (2019). What’s the Tone? Easy Doesn’t Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, *14*, 1–22. https://doi.org/10.1080/19312458.2019.1671966

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412.

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

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

Grimmer, J., 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.

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

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

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin’words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.

Müller, S. (2020). Media Coverage of Campaign Promises Throughout the Electoral Cycle. *Political Communication*, *37*(5), 696–718. https://doi.org/10.1080/10584609.2020.1744779

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

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

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

Van Der Velden, M., Schumacher, G., & Vis, B. (2018). Living in the Past or Living in the Future? Analyzing Parties’ Platform Change In Between Elections,The Netherlands 1997–2014. Political Communication, 35(3), 393–412. https://doi.org/10.1080/10584609.2017.1384771

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

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, *38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. (2018). Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, *25*(1), 364–373.