|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Validation Step** | **Documentation** | **Considerations** | **Performance Criteria** | **Source / References** |
| **Construct Definition and Operationalization** | | | | | |
| I.1 | Documentation of the conceptual background | * Reference to existing definitions for sexism * Reference to previous attempts to measures sexism (misogyny, benevolent vs. hostile sexism, etc.) * Discussion of implications of definitional unclarity * Reflection on previous models capturing spurious artifacts of the datasets instead of sexist language. * Systematic engagement with survey scales to measures sexist language. * Considerations of sexist phrasing (not only content), i.e., offensive language (no explanation how this decision might relate to literature) | 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 | * Development of a detailed codebook based on four dimensions identified in the psychological literature. * Discussion of coding inconsistencies + Adaption of the coding instructions | Have I sufficiently explained how the construct should manifest itself in the textual data? Have I documented my operationalization in a codebook? | Providing definition and conceptualization of the construct | Krippendorf (2018) |
| I.3 | Manual Precoding | * Test of the codebook using 5 MTURKERS (86% agreement for majority verdict (at least 3 out of 5 agreement) on the survey scales | Have I reached sufficient interrater agreement for a subsample of the textual data? Have I ensured that the construct can be detected in the textual data? Have I outlined my rules of coding uncertainty across coders? | Reaching sufficient interrater agreement (e.g., Krippendorff’s alpha α) | Krippendorf (2018), Plank (2022) |
| **Design Decisions** | | | | | |
| I.4 | Justification of data collection decisions | * Combination of different data sources with different characteristics   + Scale items from psychological scales   + Twitter data collected by keywords and human-annotated.   + Twitter data collected by “call me sexist but” phrase (quite experimental approach)   + Creation of adversarial examples (“crowd workers to generate adversarial examples, i.e., examples that are a valid input for a machine learning model, strategically synthesized to put the model to test”) * Discussion on the features of the annotated datasets | Have I selected a dataset that is representative and relevant to the research question and population of interest? Have I justified the data selection decisions (e.g., using keywords)? Have I assessed the quality and completeness of the dataset and checked for potential biases or inconsistencies? | Outlining the rationale behind data selection / collection decisions; Documenting potential limitations and data quality issues | Krippendorf (2018) |
| I.5 | Justification of method choice | * Application of different models with increasing complexity (Logit/CNN/Bert) including state of the art methods * Systematic Comparison of these methods | Have I selected the appropriate type of method based on the operationalization of the construct and data characteristics? Have I justified the concrete selection of a particular model? | Outlining the rationale behind method selection; Documenting potential limitations in comparison to alternative methods | Grimmer et al. (2022) |
| I.6 | Justification of the level of analysis | * Not explicitly mentioned, but the focus on the sentence level (based on the survey scales) appears plausible in regard to the literature. | Have I selected the appropriate level of analysis? Have I considered potential problems when aggregating measures from lower to higher levels (e.g., sentence to paragraph level)? | Outline the rationale behind the selected level of analysis (e.g., token, sentence, or paragraph level). | Jankowski & Huber (2022) |
| I.7 | Justification of preprocessing decisions | * Only for the Logit model, a short reference to the adoption of preprocessing decisions similar to Jha and Mamidi is provided. | Have I justified relevant changes to the text prior to the analysis, such as removing certain words or phrases? | Outlining the rationale behind preprocessing decisions | Grimmer et al. (2022) |