|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Validation Step | Documentation | Considerations | Performance Criteria | Source / References |
| Construct Definition and Operationalization | | | | | |
| III.1 | Comparison of measures with human-annotated test set ("gold-standard data") | * Systematic comparison with hand-annotated test set, report of F1 scores | Have I reached sufficient predictive performance on a test set of held-out human annotations? Did I apply cross-validation to calculate average performance metrics? | Evaluating performance metrics (i.e., F1) for dataset of human annotations | (Samory et al., 2021) |
| Surrogate Label Comparison (context-specific) | | | | | |
| V.5 | Comparison of measures with surrogate labels | * Not provided | Have I reached sufficient predictive performance on the surrogate labels? | Evaluating performance metrics (i.e., F1) for the surrogate labels | Grimmer et al. (2022) |
| Criterion Prediction (context-specific) | | | | | |
| V.6 | Criterion Prediction | * Not provided | Have I been able to accurately predict real-word phenomena? | Evaluating predictive metrics (i.e., regression coefficient) for the criteria | Grimmer et al. (2022) |