ID	Validation Step	Documentation	Considerations	Performance Criteria	Source / References
Model Feature Inspection					
I.1	Inspection of predictive model features	• Conducting of feature importance analysis for predictive unigrams (Table 4)	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	• Not provided	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	• Not provided	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	• Not provided	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	• detailed discussion of misclassified examples, identification of systematic errors (e.g., varying performance of baseline model for topicality)	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	• Not provided	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	(Wu et al., 2019)

				misclassified examples	
Systematic Testing (context-specific)					
V.1	Counterfactual tests	• Conducting counterfactual tests; providing new training samples of counterfactual tests and displaying performance metrics (F1 score).	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	• Not provided	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	• Not provided	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	• Not provided	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)