Online Appendix For:

**In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis**

**1. Variables coded in Study 1, detailed coding instructions, and reliability estimates**

Following variables utilized in Study 1 were coded by 5 trained coders.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition & Coding instructions** | **Reliability** |
| Relevance | Whether empirical text analysis is conducted and reported (Yes = 1, No = 0) | Alpha = 1 |
| Method Used | 1 = Search string based / Dictionary Approach  2 = Machine Learning  3 = Topic Modeling (excluded from further analysis)  4 = Other (excluded from further analysis) | Alpha = 1 |
| Refer to gold standard | 1 = Yes, a “gold standard” is used, and info is reported  0 = No is not used reported | Alpha = 1 |
| Report reliability | Whether intercoder-reliability of human-coded materials are reported?  (1 = Yes, reported, 0 = Not reported) | Alpha = 1 |
| Refer to validation / Report validation measures | Whether validation of automated procedures are mentioned, and if so, whether either one of validation metrics (e.g., Recall, Sensitivity, Precision, Accuracy, F1, or other measures) is reported?  (1 = Yes, mentioned, 0 = Not mentioned) | Alpha = .750 |

A total of ﬁve highly qualiﬁed coders tested the initial coding scheme by independently coding 10 sample articles (approximately 5% of the total retrieved sample) and collectively discussed any coding problems and disagreement. Coding instructions were iteratively revised until the coding schemes would produce reliable results.

**2. Detailed Setup of MC simulations**

**Supervised ML Scenario**. We set three independent variables to be sampled from a multivariate normal distribution, with a randomly generated variance-covariance matrix Σ for each simulation run. This ensures that idiosyncratic values of the covariance matrix do not skew the overall results of the simulation. The true values of (which is the binary variable) are then sampled from a Binomial distribution, with the probability parameter having a very simple linear functional form as follows:

with being Gaussian noise added to ensure that each simulation run is not completely deterministic. The , the true population parameter, was fixed throughout the simulation runs (specifically, = 0.5, = 0.2, and = 0.6, which were randomly chosen).

**Dictionary-based Scenario.** For a dictionary (i.e., bag-of-words) method, we assume a very similar approach as discussed above, but assume values of independent variables that were sampled from a Categorical distribution (i.e., a discrete value range from -5 to 5), where they represent some “features” of given textual data (e.g., a word or N-grams). Since this requires discrete rather than continuous values, we use a slightly different setup as follows:

with being the total number of observations, and and being hyper-parameters governing the shapes of the categorical distribution, and being a set of independent variables (with *K* being the number of textual features). The two negative and positive Dirichlet priors and were randomly selected for each independent variable, effectively treating such independent variables as systematic, recurring “features” of given textual data based on which is generated in a similar fashion as in Equations (1). Yet for the dictionary-based approach, the vector was extended to *K* = 5 and their values were fixed to 0.2. This enables us to better approximate the multidimensionality of textual data, while treating effectively as a function of the simple sum of the chosen textual features.



Figure A1. Overall classification accuracy against true value across MC simulation conditions, Study 2 (reference line is the overall mean).





Figure A2. Percentage of decision error and relative bias in F1-score (over 1000 simulations per each scenario), GLM classifier.

Note: Upper panel = Proportion of cases (each simulation run) incorrectly conclude on classiﬁcation performances. Lower panel = Relative bias in F1 scores among 1000 replications, with their median and 95% percentile conﬁdence intervals.





Figure A3. Percentage of decision error and relative bias in F1-score (over 1000 simulations per each scenario), Bag-of-words approach.

Note: Upper panel = Proportion of cases (each simulation run) incorrectly conclude on classiﬁcation performances. Lower panel = Relative bias in F1 scores among 1000 replications, with their median and 95% percentile conﬁdence intervals.