How to Stop Ignoring Automated Classification Errors

Differential Measurement Error & Intercoder Reliability in Measurement Error Models

Nathan TeBlunthuis (*Northwestern University*), Valerie Hase (*LMU Munich*), & Chung-hong Chan (*GESIS*)

Northwestern | SCHOOL OF COMMUNICATION

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

gesis Leibniz-Institut für Sozialwissenschaften

Addressing Misclassification Errors in Text-as-Data Approaches

- Automated text classifiers (ACs) based on Supervised Machine Learning (SML) gain traction (Baden et al., 2022; Hase et al., 2022; Jünger et al., 2022)

- *Misclassification errors* cause bias in downstream regression analyses

- Existing approaches to correct this bias include:

    - Regression Calibration (Fong & Tyler, 2021)
    - Multiple Imputation (Blackwell et al., 2012)
    - Pseudo-Likelihood based on Confusion Matrix (Zhang, 2021)
    - Maximum Likelihood Models – we provide a tailored implementation (Carroll et al., 2006)

Key Questions

1

RQ1: Which error correction methods do scholars employ for SML-based text classification?

2

RQ2: Which error correction methods *should* scholars employ for SML-based text classification?

Misclassification is a largely ignored threat

1

Identification of SML-based text-as-data studies (*N* = 49) via existing reviews
(Baden et al., 2022; Hase et al., 2022; Jünger et al., 2022; Song et al., 2021)

- Of empirical applications, 41.9% create covariates & 29% outcome variables via SML-based ACs
- **Only 16.1%** of studies mention misclassification errors, only **single study** employs correction method

Using Validation Data to Correct Estimates

2

We desire a method that will:

- provide consistent estimates
- generalize to studies where an AC measures a covariate or outcome
- handle bias and differential error[1]

We use monte-carlo simulations to test proposed methods.

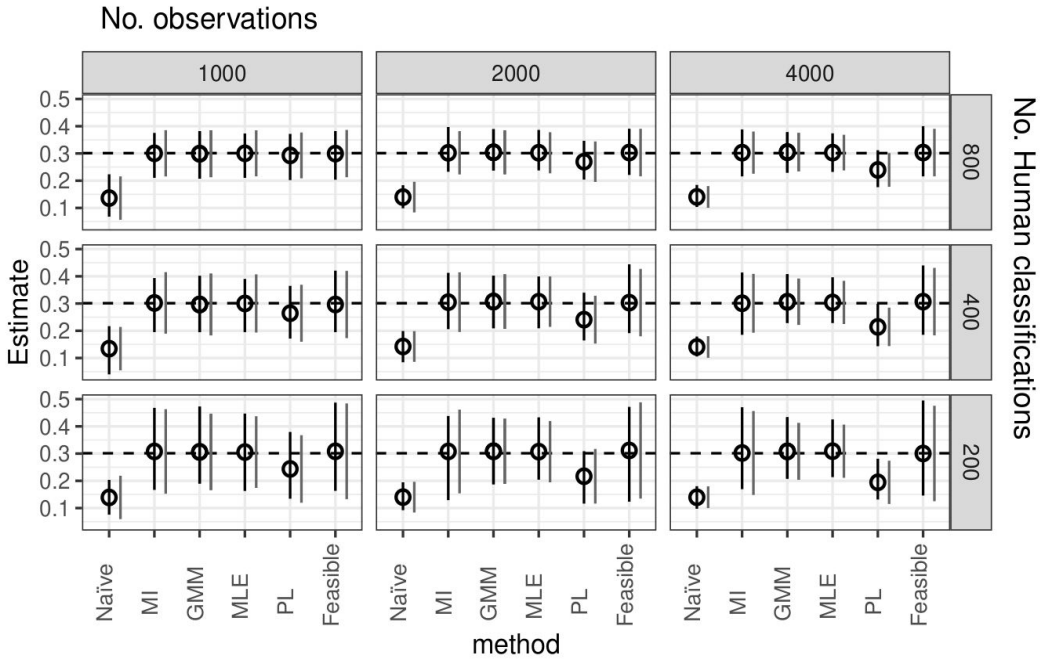
[1] Differential error is when misclassifications directly depend on the outcome.

Results: Nondifferential error

Our *likelihood modeling* approach and *GMM calibration* (Fong and Taylor) is consistent and efficient when error is nondifferential error.

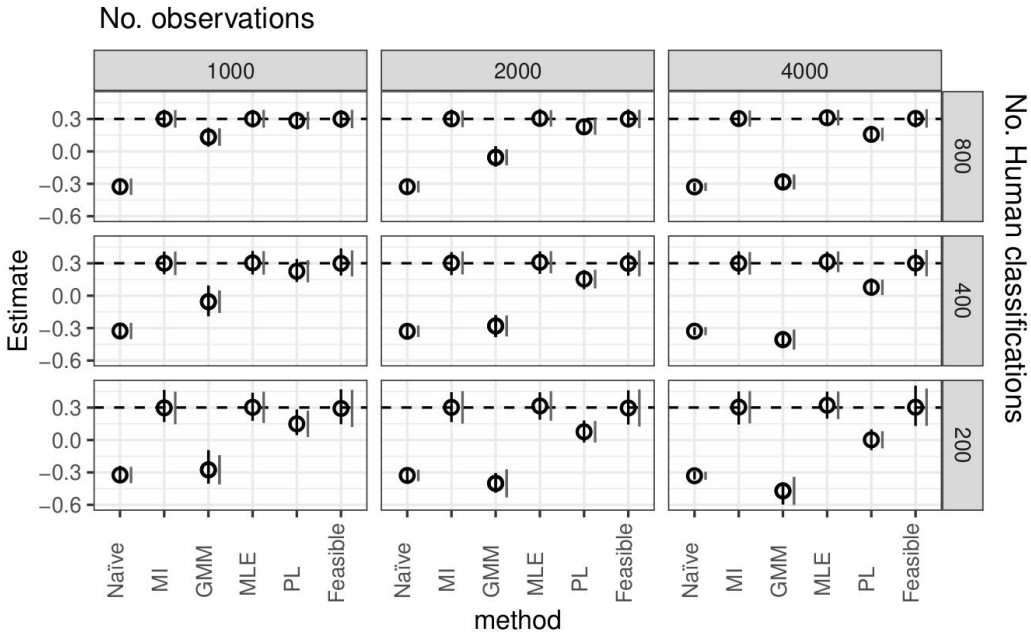
Notably, Zhang 2021's *pseudo-likelihood method* (PL) is inconsistent.

Multiple imputation is consistent, but inefficient.



Results: Differential error

Only our *likelihood modeling* approach and *multiple imputation* are consistent when measurement error is differential.

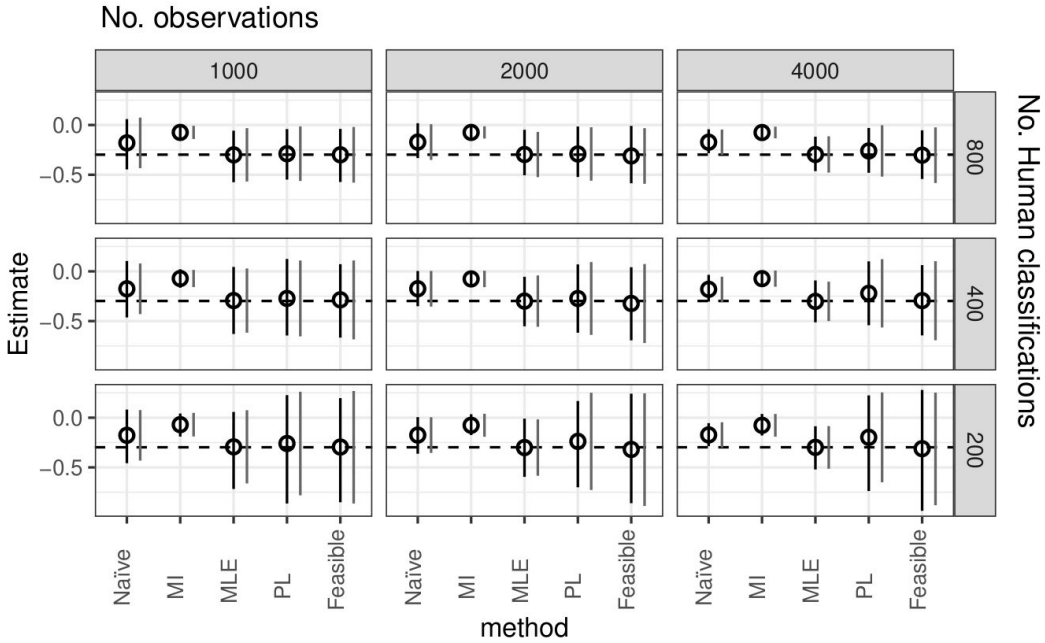


Results: (unbiased) Error in the outcome

Our *likelihood modeling* method is consistent and efficient when classification errors are in the outcome.

Multiple imputation is inconsistent.

Zhang 2021's *pseudo-likelihood model* is consistent, but inefficient.



Conclusion

(1) *"Validate, validate, validate" - yeah, but how?*
   (a) Prediction / accuracy metrics are not enough. They don't capture differential measurement errors or determine how measurement error affects inference.
   (b) Validation data can diagnose differential error.
   (c) Human coders make mistakes, use at least 2.

(2) *Validation data can be used to correct for (differential) measurement errors*

(3) *Report the "naïve", "feasible", and appropriate corrected estimates.*

Thank you!

- Bachl, M., & Scharkow, M. (2017). Correcting Measurement Error in Content Analysis. *Communication Methods and Measures*, *11*(2), 87–104. https://doi.org/10.1080/19312458.2017.1305103
- Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574
- Blackwell, M., Honaker, J., & King, G. (2012). *Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data*. 50.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models* (2nd Aufl.). Chapman & Hall/CRC.
- Fong, C., & Tyler, M. (2021). Machine Learning Predictions as Regression Covariates. *Political Analysis*, *29*(4), 467–484. https://doi.org/10.1017/pan.2020.38
- Jünger, J., Geise, S., & Hännelt, M. (2022). Unboxing Computational Social Media Research From a Datahermeneutical Perspective: How Do Scholars Address the Tension Between Automation and Interpretation? *International Journal of Communication*, *16*, 1482–1505.
- Hase, V., Mahl, D., & Schäfer, M. S. (2022). Der „Computational Turn“: Ein „interdisziplinärer Turn“? Ein systematischer Überblick zur Nutzung der automatisierten Inhaltsanalyse in der Journalismusforschung. *Medien & Kommunikationswissenschaft*, *70*(1–2), 60–78. https://doi.org/10.5771/1615-634X-2022-1-2-60
- 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
- Zhang, H. (2021). *How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/453jk