COMMUNICATION 2024
METHODS AND MEASURES

Nathan TeBlunthuis

Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can!

Nathan TeBlunthuis[a,b], Valerie Hase[c], and Chung-Hong Chan[d]

[a]School of Information, University of Michigan, Ann Arbor, USA; [b]Department of Communication Studies, Northwestern University, Evanstone, USA; [c]Department of Media and Communication, LMU Munich, Munich, Germany; [d]Department of Computational Social Science, Cologne, Germany; GESIS - Leibniz-Institut für Sozialwissenschaften TeBlunthuis, Hase & Chan, Mannheim, Germany

ABSTRACT

Automated classifiers (ACs), often built via supervised machine learning (SML), can categorize large, statistically powerful samples of data ranging from text to images and video. They have become widely popular measurement devices in communication science and related fields. Despite this popularity, even highly accurate classifiers make errors that cause misclassification bias and misleading results when input to downstream statistical analyses–unless such analyses account for these errors. As we show in a systematic literature review of SML applications, communication scholars largely ignore misclassification bias. In principle, existing statistical methods can use "gold standard" validation data, such as that created by human annotators, to correct misclassification bias. We introduce and test such methods, including a new method we design and implement in the R package MISCLASSIFICATION_MODELS, via Monte Carlo simulations designed to reveal each method's limitations, which we also release. Based on our results, we recommend our new error correction method as it is versatile and efficient. In sum, automated classifiers, even those below common accuracy standards or those making systematic misclassifications, can be useful for measurement with careful study design and appropriate error correction methods.

Automated classifiers (ACs) based on supervised machine learning (SML) have rapidly gained popularity as part of the *automated content analysis* toolkit in communication science (Baden et al., 2022). With ACs, researchers can categorize large samples of text, images, video, or other types of data into predefined categories (Scharkow, 2013). Studies, for instance, use SML-based classifiers to study frames (Burscher et al., 2014), tonality (van Atteveldt et al., 2021), or civility (Hede et al., 2021) in news media texts or social media posts.

However, there is an increasing concern that automated classifiers' imperfections may compromise measurement validity for studying theories and concepts from communication science (Baden et al., 2022; Hase et al., 2022). Research areas where ACs have the greatest potential –e.g., content moderation, social media bots, affective polarization, or radicalization–are haunted by the specter of methodological questions related to misclassification bias (Rauchfleisch et al., 2020): How accurate must an AC be to measure a variable? Can an AC built for one context be used in another (Burscher et al., 2015; Hede et al., 2021)? Is comparing automated classifications to some external ground truth sufficient to claim validity? How do biases in AC-based measurements affect downstream statistical analyses (Millimet & Parmeter, 2022)?

TeBlunthuis, N., Hase, V., & Chan, C.-H. (2024). Misclassification in automated content analysis causes bias in regression. Can we fix it? Yes we can! Communication Methods and Measures, 18(3), 278–299. https://doi.org/10.1080/19312458.2023.2293713