When Does Garbage Stink? Imperfect Gold Standards and the Validation of Automated Content Analysis

Hyunjin Song, Hajo Boomgaarden, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, & Petro Tolochko

Automated text analysis methods become increasingly popular for analyzing texts in the social sciences, ranging from large-scale analyses of decades of newspaper coverage and party manifestos to millions of social media posts. Taking advantage of the fact that ever-growing quantities of text are available and have to be analyzed with limited resources, research in the social sciences nowadays readily turns to automated approaches to investigate a great range of questions in manifold sources (Boumans & Trilling, 2016; de Graaf & van der Vossen, 2013). However, with the growing popularity of such approaches, the issue of the validity of obtained results and the conclusions drawn from them becomes crucial. Blindly applying automated approaches without proper validation may result in misleading or even plainly wrong findings; a principle famously illustrated by the phrase "garbage in, garbage out".

In that respect, such "text-as-data" approaches squarely depend on a proper validation of applied techniques against some gold-standard/ground truth (Grimmer and Stewart, 2013). Typically, applications of validation procedures rely on experimentally elicited human inputs ("human coding") as a benchmark to validate proposed (un)supervised scaling methods or dictionary-based classification approaches. The most common practice here is to rely on precision (share of relevant items in all selected items), recall (share of selected items in all relevant items), and the resulting F1 score (ratio of precision and recall), using human coding material as a benchmark. Assuming that humans' understanding of text outperforms that of machines and that, if trained correctly, humans will make most correct and valid classifications of texts, human coded data is treated as a gold standard against which the performance of the computer is judged. Consequently, the validation of computer-assisted coding based on human coded gold standards is grounded in the assumption that human readers' placements or evaluations of a given text are indeed as close as we can get to an error-free, faultless and faithful representation of the quantities being estimated. However, "the quantities we seek to estimate from text […] are fundamentally unobservable" (Lowe & Benoit, 2013, p. 299), and human judgment is, in fact, no exception to this general rule as we know from the methodological content analysis literature. The consequences of, for example, human biases, predispositions and situational disturbances resulting in differing levels of "reliability" in human judgment in evaluating texts are well documented in traditional content-analytic applications (e.g. Krippendorff, 2004; Hayes and Krippendorff, 2007; Lombard, Snyder-Duch, & Bracken, 2002). Yet, the implications of using such human judgments as the benchmark for evaluating the validity of the results of automated text analysis – and especially the consequences of different levels of reliability in the gold standard are, until today, not well understood.

Against this backdrop, we first argue that a systematic validation of automated text analysis approaches is generally still rare in the social science literature, in general, and in communication research, in particular. Substantiating this argument, this study first presents a systematic review of published works in the top journals in social sciences the past 20 years[1], showing that standards of validation are far from being acknowledged in the literature and that the reporting and interpretation of validation procedures differ greatly. In a second step, then, we argue that by conditioning the relative performances of a given automated method against human inputs, such validation procedures essentially aim to mirror human performances, therefore effectively "tolerate" any mistakes or classification errors of automated procedures to a degree comparable to imperfect human judgments at researcher's chosen degree of reliability levels. Thus, the choice of using human coding – and in particular the quality of such benchmarks – as a ground truth or gold standard may have systematic consequences for the evaluation of the validity of the proposed automatic procedures.

We assess this previously unexplored connection between the reliability of human judgment and relative performance of automated procedures (against true standards) by simulating the interconnection of standards of reliability in the gold standard material (often Krippendorff alpha of .7 or above) and standard cut-offs of precision, recall, and F1 scores. We rely on systematic Monte-Carlo simulations to illustrate how different validity coupled with different reliability affects overall research results and the ultimate conclusions drawn from such results. Specifically, we generate multiple sample "contents" of known, latent quantities of text, and then add a different level of standard normal error to the multiple "measurements" of those quantities to simulate different levels of reliability of human coding. Those multiple measurements across (purported) human coders are then averaged, generating the "benchmark score" against which automated procedures are evaluated. Next, we submit those simulated text datasets to a simple automated procedure routine, derive their relative performances against benchmark scores (from different levels of human coding reliability). We systematically compare different levels of precision, recall, and F1 scores evaluated at different levels of benchmark score, evaluating the consequences of choosing different combinations of (human) reliability and (machine) accuracy/recall combinations against the true standard. Our contribution should be read as a call for a thorough and systematic application of validation procedures in any publication drawing on automated text analysis procedures. At the same time, the study serves to benchmark the combination of reliability in gold standard/ ground truth and validity scores and warns against improper use of both to demonstrate the validity of the approach. **(Abstract: 874 words)**

[1] We will identify relevant studies using the EBSCO host databases "Communication & Mass Media Complete", "Humanities Source" and "SocINDEX with Full Index" and the following Boolean search string "("computer-assisted" OR "computer assisted" OR "automated" OR "automatic" OR "computational" OR "machine learning") AND ("content analysis" OR "text analysis")" querying all abstracts and corresponding keywords in the databases.