Abstract

Automated text analysis methods have become increasingly popular for analyzing texts in the social sciences. While researchers routinely and conveniently resort to some forms of additional human coding in validating their results from automated procedures, the issue of “quality assurance” for manual annotation in such a “gold standard” is often go unchecked in practice. Against this backdrop, this study ﬁrst presents a systematic review of articles published in major social science journals in the past 20 years, showing that the current practices of validation utilizing manual annotations are far from being acknowledged in the literature, and that the reporting and interpretation of validation procedures differ greatly. In a second step, we assess the connection between the quality of human judgment in manual annotations and relative performance evaluations of automated procedures (against true standards) by relying on large-scale, systematic Monte-Carlo simulations. Results conﬁrm the expectation that there is a greater risk of a researcher to reach an incorrect conclusion regarding the performances of proposed automated procedures when a quality of manual annotations used in validation of automated procedures is at suspect. Our contribution should therefore be read as a call for a systematic application of manual validation materials in any social science publication drawing on automated text analysis procedures.

*Keywords: Automated text analysis, reliability, validation, Monte-Carlo simulations*

**When does garbage start to stink? The impact of imperfect human annotations as a gold standard on the validation of automated content analysis**

Automated text analysis methods have become increasingly popular for analyzing texts in the social sciences and in communication science in particular, 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 readily available, while available resources to analyze these are usually limited, research in the social sciences nowadays eagerly turns to automated approaches. 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 and careful validation may result in misleading or even plainly wrong findings; a principle famously illustrated by the phrase “garbage in, garbage out.”

In this regard, such “text-as-data” approaches squarely depend on a proper triangulation of applied techniques against some gold standard (Grimmer & Stewart, 2013).[[1]](#footnote-2) Typically, this is done by using some human inputs (“human coding” or “manual annotations”) as a benchmark to systematically compare and evaluate against the proposed supervised methods-based or dictionary-based classifications. Acting under the assumption 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.

“[T]he 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. 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 (Krippendorff, 2004; Ennser-Jedenastik & Meyer, 2018; Hayes & Krippendorff, 2007; Lombard, Snyder-Duch, & Bracken, 2002). Therefore, ensuring the acceptable “quality” of manual coding (e.g., a proper coder training and achieving an acceptable level of intercoder reliability) in traditional manual content analysis is now regarded as *the* standard practice in the field. However, as we will argue below, there has been relatively a lack of parallel attention to the issue of ensuring the acceptable quality in human coding when such manual annotations are further utilized as a gold standard material in automated content analysis.When it comes to actual research practice, many of the human-coded materials are not properly evaluated and checked before being utilized in automated procedures, and frequently conveniently misused as such by many researchers. By conditioning the relative performances of a given automated method against such “imperfect” human inputs as a gold standard, we argue, such practice greatly percolates additional systematic errors unbeknown to researchers in automated procedure. Yet, the implications of using such imperfect human judgments as the benchmark in validating the results of automated procedures are, until today, insufficiently addressed.

Against this backdrop, we first present a systematic review of articles published in the major communication science and general social science journals in the past 20 years, showing that the actual practices of systematic validation are far from what is being acknowledged as a standard in the literature. We also show that the reporting and interpretation of validation procedures also differ greatly across studies, and most importantly, human-coded materials are frequently misused by not being properly evaluated and checked before being utilized in such validations. In a second step, we argue that using “imperfect” human coding as the benchmark has systematic consequences for evaluating the performances of the proposed automatic procedures. We assess this previously unexplored connection by relying on a large-scale, systematic Monte-Carlo simulation. To our knowledge, our contribution is among the very first to provide thorough, systematic evidence pertaining to a well-known, yet extremely scarcely discussed topic. While our contribution should be read as a call for a systematic evaluation of the use of human coding in its application for validation procedures in automated content analysis, this study also serves to benchmark the combination of reliability in gold standard and validity scores, and warns against improper use of both in demonstrating the validity of the automated approach.

**Logics and Procedures of Automated Content Analysis**

We define “automated content analysis” (or automated text analysis) as the collection of content-analytic approaches that utilize automated methods of coding a large amount of unstructured textual data, in a way that the coding process itself (i.e., the text classification) is not performed manually but rather done by predefined computational algorithms (Grimmer & Stewart, 2013). While the usage of the term “automated content analysis” in general in the extant literature encompasses a wide variety of forms (e.g., Riffe, Lacy, & Fico, 2014; Grimmer & Stewart, 2013; Hopkins & King, 2010; Krippendorff, 2013), here we mainly refer to supervised learning methods and a simple dictionary approach. We choose to do so since many of the automated content analysis applications in communication science in general still heavily rely on those two broad approaches. Given that such methods are “designed to automate the hand coding of documents…[in a way that] it will directly replicate the hand coding” (Grimmer & Stewart, 2013, p. 13), they are an attractive option for overcoming the limitations of manual approaches.[[2]](#footnote-3)

Although any specific method and application of automated content analysis comes in many flavors (for broad overviews, see: Boumans & Trilling, 2016; Grimmer & Stewart, 2013), they are all based on similar principles: they all aim to identify and classify predefined categories (i.e., a discovery and measurement). A simple *dictionary approach* generally relies on a pre-defined “dictionary” of words or phrases by a researcher where the computer counts the number of such instances in a given document. Assuming a given dictionary is appropriate to be applied in a given domain (Boumans & Trilling, 2016; González-Bailón & Paltoglou, 2015), a researcher applies the dictionary on a set of documents, typically to a word-level, and then resulting “scores” are aggregated to the document level (e.g., a news article or tweet). This “dictionary” therefore represents an explicit coding rule to be applied by an algorithm, where (classified) categories may represent the visibility of a topic (i.e., the presence or the absence of a category) or a net tonality (positive vs. negative) of an article (e.g., Aaldering & Vliegenthart, 2016; Boomgaarden & Vliegenthart, 2009; González-Bailón & Paltoglou, 2015; Rooduijn & Pauwels, 2011; Young & Soroka, 2012).

For *supervised methods*, specific coding rules in manual annotations (as an input to an algorithm) are, in general, rarely explicitly articulated. Yet the algorithm takes such implicit judgments derived from manual coding as the point of reference, and tries to infer the features of data that best classify the text into different predefined categories. Currently, a variety of supervised learning algorithms is available – ranging from a simple regression framework or Naïve Bayes to more sophisticated methods such as Support Vector Machines (SVM) or random decision forest (for an overview, see Hindman, 2015). This machine-learning process may involve many iterations of “training” and comparisons to some gold standard materials in order to achieve a desired level of classification quality (e.g., Scharkow, 2013). If the algorithmic coding is later determined to be sufficient enough in capturing the underlying quantities of interest through such a “training,” then a researcher applies such “trained” coding rules to a larger corpus of unseen documents to perform their analyses (e.g., Burscher, Vliegenthart, & De Vreese, 2015; Burscher et al., 2014; González-Bailón & Paltoglou, 2015; Scharkow, 2013).

**The Use of Human Annotations in Validation of Automated Content Analysis**

For automated content analysis methods in particular, the discussion of “validation” has traditionally evolved around the logic of establishing a correlative or a convergent validity (Krippendorff, 2008). This is because any automated content analysis – at least the ones we are focusing on in this contribution – essentially can be regarded as a classification problem (i.e., a measurement of the predefined categories in data). In this respect, the most straightforward method of “validating” a given measurement is to compare with another measurement of the same concept. Therefore, applications of validation procedures in automated content analysis have traditionally relied on some sort of human inputs measuring the same concept serving as a benchmark, or a “ground truth,” as discussed above.

In dictionary-based approaches, the actual coding process itself – using the existing dictionary – does not involve any human inputs. Instead, the use of manual coding in dictionary approaches may additionally involve comparisons of the derived results against *post-hoc* manual coding. A recent work by Rooduijn & Pauwels (2011) on the automated measurement of “populism” in election manifestos well exemplifies this strategy. Using a traditional manual content analysis, they show that dictionary-based automatic coding of “populism” categories in party manifestos produces essentially very similar results compared to manual coding of a trained researcher. Similarly, Young & Soroka (2012) compared the manually coded newspaper content against the results based on Lexicoder Sentiment Dictionary (LSD), and found that results using LSD and manual content analysis are largely comparable to each other.[[3]](#footnote-4) Nevertheless, while it appears that some additional post-hoc validations – in particular using proper human coding – could ensure the soundness of conclusions drawn from dictionary approaches, it is exceptionally rare to see a validation of results *after* such classification tasks (yet for notable exceptions, see Muddiman et al., 2018; González-Bailón & Paltoglou, 2015; Young & Soroka, 2012).

In supervised methods, the role of human input is more central in fine-tuning the algorithm (i.e., an “inner” cross-validation). As exemplified in Scharkow’s work (2013), a researcher typically produces manually annotated sample materials, or a “training set,” and the algorithm is later trained on sample material in order to develop certain statistical models (that effectively aim to reproduce implicit coding rules of human coders) to predict and classify unseen materials, the “test set.” As such, human inputs are more directly utilized in such learning procedures in supervised methods, as opposed to being utilized post-hoc as in dictionary approaches. For instance, Burscher et al. (2014) have developed a supervised machine-learning algorithm to automatically classify certain generic media frames in news articles, based on a random subset of data that were manually coded by trained coders.[[4]](#footnote-5) Yet due to inherent resource constraints, such validations typically rely on only a very small subset of held-out samples that have to be coded *both* by human coders and by trained algorithm. As such, additional post-hoc validation as described in dictionary-based approaches (i.e., an “outer” validation) appears to be also rarely offered in practice.

**The Myth of Perfect Human Coding in Validation of Automated Content Analysis**

Regardless of its specific orientations briefly reviewed above, the use of human coding as a gold standard is often regarded as *the* principal method of ensuring the validity and soundness of conclusions derived from the proposed automated procedures (DiMaggio, 2015; Grimmer & Stewart, 2013). The purpose of such validation against a gold standard is, as Krippendorff (2008) notes, “to confer validity on the otherwise uncertain research results” (p. 6). Admittedly, this logic requires the chosen benchmark (i.e., manual annotations by human coders) to be of “objective” and “unquestionable” truth – or at its minimum, to be sufficiently intersubjective enough to establish a firm common ground (Krippendorff, 2008). Yet, as much of the traditional manual content-analytic methodology literature suggests (e.g., Krippendorff, 2004; Ennser-Jedenastik & Meyer, 2018; Hayes & Krippendorff, 2007; Lombard et al., 2002), manual coding more often easily produces unreliable judgments under a lack of proper coder training, and especially so when the judgment in hand requires a nontrivial degree of inferences and subjectivity to classify latent information (Krippendorff, 2013; Riffe et al., 2014). For this reason, at least for many recent *manual* content analysis applications, there has been relatively little disagreement regarding a need for proper “quality assurance” in a form of developing unambiguous coding categories, better coder training, and achieving an acceptable level of intercoder reliability (Krippendorff, 2004, 2013). Yet when researchers’ primary interests are in applying some forms of automated procedures, even if researchers routinely and conveniently resort to some forms of human annotation in validating their results (if at all) from proposed automated procedures, “quality assurance” for manual annotation is often go unchecked. Given a nontrivial attention to the issue of a proper validation in text-as-data approaches (Boumans & Trilling, 2016; Grimmer & Stewart, 2013), a widespread practice of conveniently utilizing quick and easy manual annotation without proper quality assurance in typical automated text analysis applications is especially surprising. “[W]henever ... principles by which humans generate ratings are heterogeneous across raters” (DiMaggio, 2015, p. 4), any automated procedures that are evaluated upon such imperfect manual annotations would be systemically biased to the degree that can be found in such imperfect human judgments.

The use of (potentially) imperfect, low-quality manual annotation as an ultimate benchmark against automatic techniques may have at least two systematic consequences. First, while automated methods themselves are perfectly reliable, imperfect human judgments (on validation materials) essentially makes the ultimate “target” of such reliable measurements radically deviate from the “true” target of inference, making them “reliably wrong” on-target. Second, and relatedly, systematically flawed human judgment can introduce unknown errors to learning algorithms, leading to biased conclusions against a true standard. Therefore, using suboptimal manual annotation makes it harder to evaluate the relative trustworthiness of such validation procedures. However, most of the empirical studies appear to pay insufficient attention to the issue of proper validation. Indeed, we know only one existing study that suggests some tentative evidence of the relationship between quality of human coding and (machine-based) classification accuracy (Burscher et al., 2014).[[5]](#footnote-6)

In sum, there appears to be good reasons to suspect a systematic relationship between the use of low-quality manual annotation and the relative bias [[6]](#footnote-7) and errors regarding the ultimate conclusions from the automated content analysis – especially when using such imperfect human coding as the “gold standard.” While many prior contributions on this topic – both theoretically and empirically – stress a need for a proper validation of applied techniques (e.g., Grimmer & Stewart, 2013; González-Bailón & Paltoglou, 2015; Hopkins & King, 2010) and despite the fact that such issues are gaining more attention in various applied research domains (e.g., such as in corpus annotations: Hovy & Lavis, 2010; Lease, 2011), we do not know much about where the field in general stands in terms of standard validation in actual research practices and the use of imperfect human coding in particular. Therefore, we first conduct a systematic review of relevant articles published in the top communication science and social science journals in the past 20 years, focusing on the use of manual annotations, if any, on standard validation procedures against the proposed automated methods, and further on the assessment of performances of such automated methods (Study 1). While Study 1 should provide an overall picture of how validation is typically approached in automated content analysis in practice, at the same time it serves to benchmark our simulation (Study 2), where we warn against *improper* use of low quality “gold standard” materials in demonstrating the validity of the automated approaches.

# Study 1: A Content Analysis Study

## Sample and Procedures

We have identified relevant studies using the EBSCO host databases “Communication & Mass Media Complete,” “Humanities Source,” and “SocINDEX with Full Index,” querying all titles, abstracts, and keywords using the following Boolean search string: *("computer assisted" OR "automated" OR "automatic" OR "computational" OR "machine learning") AND ("content analysis" OR "text analysis")*. This resulted in a total of 192 identified English-language communication science and the general social science journal articles between January 1, 1998 and November 7, 2018. Among them, 119 articles were determined not relevant (e.g., non-empirical overview/introduction articles, qualitative analyses, studies using unsupervised methods, or simple keyword frequencies, etc.)[[7]](#footnote-8) and 7 articles were either duplicates or did not contain full texts. These articles were excluded from further analyses. Using the remaining 73 articles, we systemically examined whether extant applied studies a) employ any empirical validation of their primary findings, b) if so, whether they use human coded materials as a benchmark, and c) if so, whether intercoder reliability and other methodological details are adequately reported.

A total of five highly trained 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. Intercoder reliability (based on Krippendorff’s alpha) above 0.75 was ensured for each of the variables coded.

## Results

The results of the systematic literature review are presented in Table 1 below. Among 73 articles, a total of 55 studies used dictionary approaches while 18 used supervised machine learning methods.[[8]](#footnote-9) Yet only about half of the papers using a dictionary approach referred to some sort of manual gold standard in their text, typically relied on 3.4 coders with on average a total of 532.86 annotation materials, yielding a little over than 150 manual annotations per coder. Among them, even less – only 16 percent – actually reported any measures of inter-coder reliability for such materials – only 2 studies ever reported Krippendorff’s alpha values, and 3 studies report either percentage agreement or Holsti agreement measure (the rest of the studies reported other measures such as Scott’s Pi or correlation coefficients). Notably, shares are much higher when coming to papers using supervised machine learning; a total of 16 studies (out of 18 studies, 88.8%) referred to manual gold standard materials, and among them, 10 studies typically relied on 5.7 coders. Excluding one outlier (used more than 20000 manually coded materials), they on average a total of 1640 annotation materials, yielding approximately little less than 300 manual annotations per coder. Interestingly, 6 studies did not report the number of coders and/or a total size of validation dataset, making impossible to judge the soundness of the involved validation procedures. Nevertheless, the most common measure of inter-coder reliability (if reported) were percentage agreement (*N* = 7, 38%) and Krippendorff’s alpha (*N* = 6, 33%). In sum, among 73 articles we have examined, only 9 studies (which is about 20% of studies that ever refer to some human-coded gold standards) have used Krippendorff’s alpha, with an average alpha value of 0.68.

When it comes to articles referring to procedures of validation of their automated approaches, results were very similar. However, only around 40 percent of papers using a dictionary approach also reported some validation measures, compared to 67 percent of papers using supervised machine learning. The most commonly used measures of validity are indeed the widely accepted measures of precision (in 13 cases, *M* = 0.74) and recall (in 9 cases, *M* = 0.60), yielding an average of F1 score of 0.63 (in 10 cases, based on either directly reported F1 scores or indirectly calculated F1 scores from reported precision and recall). However, we also observe that other metrics were widely used as well in reporting validation. This is somewhat disconcerting insofar as these tend to be indeed either the intercoder reliability measures (e.g., Holsti, Cohen’s Kappa, Krippendorff’s alpha) or correlation coefficients, both of which are *not* designed for the validation of a given automated procedure. Interestingly, there were only three studies which reported *both* Krippendorff’s alpha in human-coded materials (which signals the proper coder training and sufficient quality assurance of manual annotations) *and* proper validation metrics at the same time.

Generally, the results of Study 1 reveal that reported measures of validation and the quality of human-coded data used thereof are far from what is being acknowledged as the best-practice in extant literature.While we believe *the* proper use of manual annotations in automated approaches should entail *both* sufficient quality assurance of manual annotations and the proper use of validation metrics based on such manual annotations at the same time, the fact that only a total of three studies (about 4% of total studies) satisfy such criteria reveals a striking lack of attention to this issue in actual research practice.

-- Table 1 About Here --

# Study 2: A Monte-Carlo Simulation Study

The result of Study 1 reveals that there is still strikingly little consistency in *whether*, and if so, *how* validation is approached and reported. About half of studies did not report *any* validation metrics when relying on automated methods. Even when they do, metrics related to (human-coding based) validations and their qualities are generally not consistently reported, and are often limited in providing actual methodological details. In order to systematically evaluate the implications of such improper use of human-coded materials in validation of automated procedures, we setup an extensive set of Monte Carlo (MC) simulations. MC simulation offers a convenient tool for systematically evaluating the relative bias (against the true standard) and coverage of a given statistic under certain scenarios (Leemann & Wasserfallen, 2017; Scharkow & Bachl, 2017).[[9]](#footnote-10)

While we based our procedures in a way that largely mirrors the typical approaches in this area, dictionary approaches and supervised learning approaches considerably differ in how they utilize manual coding, as well as their data requirements and the logic underlying each of the techniques. Therefore, we have developed two separate approaches in simulating their behaviors. Yet in all of the cases, we have broken down our approach into three stages – data generation, human coding, and automatic classification – where we systematically varied the intercoder reliability of the “gold standard” material, along with a number of related factors such as the number of coders and the number of manual annotations per coder (see below section). Then we systematically compared different scenarios in terms of their classification accuracy and F1 scores (i.e., precision and recall) based on the “true” standard (i.e., a quantity of interest that is typically unknown to researchers) in order to illustrate how different practices of human coding in automated content analyses affect the overall results and the relative trustworthiness of conclusion drawn from such results.

## Data Generation Stage

We create data (e.g., textual data to be analyzed by a researcher) with the “true” outcome value of interest; the goal of any quantitative text analysis method is to somehow approximate this true value *y*, either by human coding, machine algorithms, or some combinations of both. For the data generating process, we set the true value of *y* to be randomly generated from three hypothetical independent variables (, , and ), the values of which were also randomly sampled either from a multivariate normal distribution (for supervised learning) or from a categorical distribution (for dictionary approach – see below for a detail). This ensures that the results of our simulations are not completely deterministic, as well as not analytically driven to arrive our conclusion.

**Supervised ML Scenario**. For supervised learning scenarios, we set independent variables to be sampled from a multivariate normal distribution, with a randomly generated variance-covariance matrix 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 follow:

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

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

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 bag-of-words approach, the vector was extended to *K* = 5 and their values were fixed to 0.2. This enable us to better approximate the multidimensionality of textual data, while treating effectively a function of the simple sum of chosen textual features.

We assume the size of text data is sufficiently large to warrant an automated approach. As such, the data in question (hypothetically) covers 10 news articles per day per each of 10 outlets, spanning a total of 20 years. Accordingly, each single simulation run is set to generate 730,000 observations of media content data (10 x 10 x 365 x 20).

## Human Coding Stage

In a typical content analysis, in practice at least two or more trained human coders are assigned to a small set of sample documents, and independently code such documents. This process of coder training is then repeated until the satisfactory level of intercoder reliability is achieved (typically Krippendorff’s alpha equal or greater than 0.7). Once a coder interchangeability is ensured by such training, validation materials are often divided evenly and annotated only by a single coder (Grimmer, King, & Superti, 2018). By definition, this means that any potential remaining coder idiosyncrasies in human coding are no longer an issue for the validity of conclusions as long as the satisfactory level of intercoder reliability is achieved. However, when the quality of manual annotations is not sufficiently ensured, such idiosyncrasies may systematically bias the procedures to a unknown degree.

Following this logic, we specified following factors that may affect the quality of the “gold standard” (i.e., manual annotations by human coders) and therefore evaluations of overall performance of automated algorithms. Those factors include: the number of human coders (k = 2, 4, 7, 10), the predetermined target values of intercoder reliability (Krippendorff’s alpha = 0.5, 0.6, 0.7, 0.8, 0.9), as well as the number of annotations per coder (n = 50, 100, 250, 500). While in any typical real-world application a researcher has to make practical and logistical decisions on these factors (typically guided by a researcher’s resource constrains), these factors are indeed crucial in terms of properly ensuring the acceptable “quality” of manual annotations produced by human coders. These numbers were chosen to reflect typical procedures and their common variations, as can be seen in our Study 1 above.

In all scenarios, human coders classify a given observation as “1” (e.g., a news article contains the quantity of interest, such as a certain actor, frame, or tonality) or “0” (e.g., does not contain this quantity of interest). This human coding () can be, in principle, either correct or incorrect against the (unknown) “true” value of , therefore behaviors of human coders were modeled by a Binomial distribution with varying probability of successfully categorizing the true data. The heterogeneity in each of the coders’ coding behaviors (e.g., expertise) was also modelled by a beta distribution with varying shape (using predefined hyperparameters for each target reliability level), which effectively enables us to simulate a situation where a certain number of coders produce more “correct” judgment () whereas other coders produce more “false” judgment ().[[10]](#footnote-11) Yet the overall human annotation patterns at the chosen level of a beta distribution parameters were ensured to produce an associated target reliability coefficient. Depending on the specific application (dictionary-based vs. supervised learning), the “hand-coded data” from this stage is then used as either for post-hoc validation or as an input for the later automated classification.

## Algorithm-based Classification and Validation Stage

In this final stage, a researcher uses a certain algorithm to predict the values of () in each of the observations. For this purpose, we set up three different classification algorithms – binomial GLM, Naïve Bayes, and a bag-of-words dictionary approach – in simulations. Although these are not exhaustive, these methods are among the most frequently utilized classification algorithms in automated content analysis (as in footnote 6 above).

In typical a real-world scenario, researchers would validate the results obtained from the automated analysis based on a manually coded test set (produced in a human coding stage) in a post-hoc manner (for a bag-of-words dictionary approach), or use such manually coded data to develop prediction algorithms (for supervised learning, such as binomial GLM or Naïve Bayes), evaluating precision, recall, and F1 scores against such “human-coded” materials (hereafter “observed” classification performance). Yet more importantly, since we are effectively simulating the typical media contents data, we can also systematically evaluate the “*true*” classification performance of automated procedures as well, and then further compare this true performance with the observed performances from scenarios using varying quality of “human-coded” materials. Admittedly, this would be impossible for practical applications since the true value y would never be known. However by doing so, we can precisely estimate the relative bias when using imperfect human coding as “gold standard” while we can observe how the overall accuracy is adversely affected based on which factors.

The final Monte Carlo simulation used 4 (number of human coders, k) 5 (target Krippendorff’s alpha levels in human coding) 4 (N of annotation per each human coder in producing validation data) 3 (classification algorithms) full factorial design with 1,000 replications per each scenario (N = 240,000), which were then summarized as below.

# Simulation Results

As an initial model check, we first look at the overall classification accuracy against true value (“true” classification performance) as a function of target reliability in validation materials. This is to check whether our simulation setup can indeed correctly reproduce common patterns of results in extant studies, ensuring the validity of our setup and its results.

In Figure A1 of the online appendix, the overall prediction accuracy against true value (defined as the sum of true positive and true negative cases over entire cases) is reported, along with their 95% confidence intervals, for every combination of experimental factors in MC simulations. Using the overall accuracy in our simulation as the reference point, the first four and the second four sets of panels in Figure A1 makes it clear that the reliability level of the training material has a nontrivial benefit in improving the accuracy of predictions based on automated procedures, especially for ML methods. Indeed, this result is expected since the ML methods takes the human input as the basis for developing the classification algorithm, therefore overall accuracy of the final classifier is dependent upon the size of human input and its quality thereof. In contrast, the last four sets of panels in Figure A1 shows that the typical Bag-of-words application does not benefit from improved (post-hoc) human coding. Yet again, this is an expected result, since the performance of a given dictionary itself does not depend at all on *post-hoc* validation. Overall, Figure A1 shows that our setup is indeed capable of correctly reproduce a common pattern in extant studies.

Next, we examine indirect consequences of relying on imperfect, less-than-desired quality manual annotations as a benchmark in evaluating the performances of automated procedures. Typically, researchers rely on a small fraction of held-out human-coded materials in this stage, deciding whether the (observed) overall accuracy or classification performance is good enough to proceed to next stage. This decision has to be based on some some *a priori* chosen threshold value: if validation (based on human coding) is purportedly not satisfying enough to pass such a threshold level, additional steps are sought to improve the quality of automated procedures (e.g., re-training algorithms, or changing the dictionary, etc).[[11]](#footnote-12) The primary interest of such validation lies in extrapolating the observed level of classification performance (based on “manual” validation materials) to the level of classification performance *that could have been observed* under the perfect, “true” quantity of interest. In other words, the “observed” classification performance based on manual annotations serves as a proxy, or an estimate for the true, yet often unknown, classification performance against the true standard. An interesting question here is thus how well the “observed” classification performance reflects the true classification performance when researchers use imperfect manual annotations, and how large or small a potential bias is compared to the true scores.

In order to illuminate this issue, we divide our entire simulations into four mutually exclusive categories as in Table 2 based on the cross-tabulation of “observed” F1 scores (against “manual” validation materials) versus “true” F1 scores (against the true values of y). We first present the proportion of simulation cases which incorrectly conclude about the overall classification quality based on “observed” classification performance, using the cutoff value of F1 score = 0.6311 (i.e., the average F1 score reported from Study 1). Second, we further present the degree of relative bias, or “error” — defined as the , where F1 score is a weighted average of the precision and recall — which captures the degree of under- or over-estimation of trueF1 scores against observed F1 scores, the latter of which is based on typical “imperfect” manually-annotated validation materials.

– Table 2 and Figure 1 to 3 About Here –

As can be seen in above Figures, it appears that utilizing more “high-quality” manual annotated materials for post-hoc validation has obvious and discernible consequences in the evaluation of classification quality of the automated procedures. Among 1000 replications of each scenario, all of our experimental factors appear to decrease the decision error rates in using observed level of F1 scores to approximate the true F1 score level. The leftmost upper panel in Figure 1 shows that in worst-case scenarios in which using only two coders with a handful of materials (annotations per coder *N* = 50), coupled with low reliability (K alpha of 0.5), such suboptimal manually annotated validation materials result in approximately 30% of cases incorrectly over- or underestimate the true F1 scores. Under the same combination of a total number of coders and independent annotations per each coder, improving intercoder reliability at best marginally decreases the overall percentage of decision errors. Yet as either a total number of coders or a number of independent annotations per each coder start to increase, we see that the total proportion of cases that incorrectly estimate the true F1 scores start to decrease substantially (the upper panel of Figure 1). Indeed, when the total size of manually annotated validation materials exceed 1000 observations, the percentages of decision error appears to be at acceptable range (less than 10%, based on all the cases of Type I and Type II error combined), although there also appears to be some variation depending on specific algorithms being utilized in proposed automated procedures.

The lower panel of Figure 1 displays the point estimate of relative bias (defined as the , or the degree of under- or overestimation of the true F1 scores based on observed F1 score), along with their 95% CIs from 1000 replications of each scenarios. We see that the distribution of relative bias also starts to converge to unbiased estimates (toward 1) as the total number of annotations start to increase, no matter of which combinations of total number of coders and no. of manual annotations per coder is being used in both in ML scenarios and dictionary approaches. In addition, while there initially appears to be no apparent overall main effect of higher reliability of manually-annotated validation materials in terms of reducing the total number of decision errors in ML approaches (as per the upper panel of Figure 1), using higher reliability in manual validation materials nevertheless tends to reduce the uncertainties of relative errors, as shown in the lower panel of Figure 1, and especially more so when coupled with a larger size of validation materials. For instance, under a scenario of *N* = 100 validation materials (e.g., 2 coders x 50 annotations per each coder: the first set of bars in lower panel of Figure 1), increasing intercoder reliability from 0.5 to 0.9 reduces the 95% CI range of relative errors from [.95, 1.61] to [.99, 1.53], or 17.45% reduction in CI range. When the total number of annotations are sufficiently large, such as *N* = 5000 (e.g., 10 coders x 500 annotations per each coder: the last set of bars in lower panel of Figure 1), the same change in reliability from 0.5 to 0.9 reduces the 95% CI range of relative errors from [.92, 1.11] to [.97, 1.03], or 69.89% reduction in CIs. This thus greatly reduces the uncertainties regarding the ultimate conclusions one can draw from proposed automated procedures. While we see the largely similar patterns for GLM classifier (Figure 2), in bag-of-words applications as presented in Figure 3, we see much clearer impact of higher reliability. It appears that “observed” classification quality (against manually coded validation materials) in bag-of-words applications tends to underestimate the true classification quality (lower panel of Figure 3). Yet, higher reliability in such validation material nevertheless greatly reduces the uncertainties regarding the ultimate conclusions based on automated procedures. In sum, this provides a very consistent picture of the impact of “proper” manually coded materials – both in terms of its size and the quality – in validating the conclusions from automated procedures.

# Discussion

Based on a thorough discussion of the importance of proper validation in automated text analysis, the aims of the current investigation were twofold; first, we attempted to provide a systematic overview of current research practices concerning validation procedures of automated content analysis in communication science and in the social sciences. Second, we aimed to provide further insights into the consequences of using suboptimal-quality manual annotation as a “gold standard” material on the researchers’ ability to reach a correct conclusion regarding the performance of proposed automated procedures.

In order to achieve these goals, Study 1 relied on a systematic review of published papers from major communication and general social science journals that purportedly have relied on “automated text analysis.” The results of the first study show that, while automated content analysis procedures are widely used, there is still strikingly little consistency in *whether* and *how* validation of automated procedures is reported. Very often, studies do not report *any* validation metrics when relying on automated methods. Even when they do, validation metrics utilizing human-coding (if any) are generally not consistently reported, and are often limited in providing actual methodological details. They frequently misreport or misspecify the validation metrics, and often, they essentially not properly evaluate the quality of human coding when manual annotations are utilized in such validation procedures. For the second study, we have designed a set of Monte Carlo simulation procedures, which closely mimic typical real-world scenarios and their plausible variations of validation practices for the most widely-used automated procedures. The results revealed that any practical and logistical factors for standard “quality assurance” – the number of coders, the size of the training sets produced by each coder, and the “quality cut-off point” (e.g., Krippendorff’s alpha for intercoder reliability) – indeed all produces systematic consequences for the evaluation of the proposed automated procedures. Coupled together, the results from both studies give good reason for concern about the quality (or rather the *validity*) of conclusions drawn from automated content analyses in the social sciences. In order to make sensible conclusions, a proper validation of the automated methods is much needed; yet such a proper validation – especially when manual annotations are involved – is rarely done in practice.

To be clear, the current study *does not* make a case for the argument that we should exclusively rely on human validations, or conversely, human validations in general are a problem. Quite the contrary, one of the main points being advanced here is that humans are not perfect, therefore, *proper* validation ensuring the “quality” of manual annotations is essential especially when they are utilized as the gold standard in automated procedures. While the (potentially far from being perfect) human “gold standard” is often the best we can get, the results of this study suggest that extra caution must be taken. Often, imperfect manual annotations can percolate to machine coding to inferences if a researcher turns a blind eye to such imperfect quality of human judgment. We therefore advise researchers to pay closer attention to the issue of ensuring proper quality of manually coded validation materials.

**Some Tentative Recommendations for Best Practices**

While there exists no single fool-proof solution applicable to every situation, based on our observation from our systematic reviews and from our simulation results, here we offer some general recommendations towards the best practice of utilizing human annotations in validating the automated approaches. First, while we believe not every study with automated content analysis should use human-involved validation, yet if used, researchers should adhere to rigorous methodological standard to the degree expected for a traditional manual content analysis (e.g., Hayes & Krippendorff, 2007; Krippendorff, 2013) in preparing the manually annotated validation dataset. Also, they should fully disclose the methodological details (such as measurement details, coding instructions, coder training, intercoder reliability, etc.) of such validation data (as per our Study 1). Not only such practice would enable readers to independently judge the soundness of validation procedures involved, but also it would increase transparency and replicability of the research process.

Second, we recommend researchers to strive to increase the total size of the manually coded validation datasets as large as possible, preferably always much more than *N* = 1000. In our Study 1, we observed that they typical size of the manual annotation dataset ranged from approximately a total of 500 to 600 annotations. Yet in our simulation (Study 2), we observed that the risk of potential decision errors (by relying on observed classification accuracy based on low quality validation materials) is substantially high, sometimes up to 30%, with smaller size of the manual annotations. Yet the percentages of decision errors appear to reach at acceptable range (less than 10%) when the size of the manually annotated validation data is equal or greater than 1000. Therefore, we advise researchers to use, *at least*, more than 1000 manually annotated materials as a very minimum criterion. Of course, this number only reflects a rough rule of thumb for much simplified cases (i.e., simple binary judgement involving only one variable). For more complex and subtle judgments (such as ambiguous latent contents or non-binary judgments), one therefore should use much larger size of manual data in validation tasks, although increasing the size of manual annotations in such cases would be a challenging task.

Third, and relatedly, we observed that improving intercoder reliability in human coding sometimes offers a substantially large benefit in reducing the magnitude of potential bias in automated classification tasks relative to true, unknown standard. While this seems encouraging for many researchers, at the same time this also warns against the prevalent practice of only prioritizing the intercoder reliability improvements in coder training to reach a minimally “accepted standard” with just handful of annotation materials. Quite contrary, such improvements at best offer marginal gains over potential decision errors especially in case with very small validation dataset. Recent developments in “crowdcoding” for content analysis (Haselmayer & Jenny, 2017; Lind et al., 2017) could also offer promising ways of scaling up the manual annotation tasks (therefore increase the manual annotations in validation tasks) in this regard, although additional issue of quality control for crowdworkers quickly becomes an important issue (Lease, 2011; Lind et al., 2017).

Lastly, one should also bear in mind that without proper coder training in improving reliability, human coders often experience substantial “coding drift” over time, therefore the risk of introducing additional random errors runs very high. Since our simulation does not take account to such factor under low reliability condition, our simulation as a consequence would have produced more “optimistic” results than most of the real-world, low-intercoder reliability scenarios. For that matter, our results should *not* be read as a signal that intercoder reliability or human coders does not matter at all in validation of automated procedures.

**Limitations and Conclusions**

Few limitations should be noted. First, and concerning Study 1, our systematic review is limited to prior studies explicitly contain our search terms in their title, abstract, or keyword. Such sample, by definition, does not include all potentially relevant articles. Still, it allows us to make a sensible selection of extant articles that do not only use these methods but also advertise that they do so. In fact, we believe articles that are not part of our sample to be of even lesser rigor in the use and reporting of “gold standards” in validation materials.

Second, concerning Study 2, we only have considered a binary classification task in our simulation setup and a single dimension of validation metrics — specifically, recall, precision, and resulting F1 scores. While this greatly simplifies our main arguments and (still complex) simulation setups, there are indeed a nontrivial number of existing applications that go beyond such simple classifications, dealing with numerical forms of predictions (i.e., a document scaling). Ultimately our conclusion would be bound to our specific setup, yet we reason that our core arguments can be equally applicable to more complex forms of automated content analysis as well, and potentially to any unsupervised methods for that matter. Given the additional complexity and difficulties involved in such none-binary predictions for human coders, achieving acceptable intercoder reliability for manual validation materials in such applications is likely to be even more difficult than in simple binary ones, due to inherent ambiguities in making fine-grained distinctions among many similar categories (i.e., non-binary scores) or due to a inductive nature of such applications (i.e., unsupervised learning). Hence we suspect the potential problems of “imperfect quality” in human-coded validation materials should be greater, if not identical, in those applications.

Designing a simulation-based study gives us a unique opportunity to observe many potential counterfactualscenarios of the research processes. Such an approach, if carefully designed, allows to robustly explore potentially important variability in research practices and their consequences. Also, one of the core advantages of relying on such a simulation-based approach is that a researcher has the ability of (already) “knowing the truth,” or the true value of the quantity of interest (in each of media data observations in this case). This also provides an opportunity of formally checking the sensitivity of one’s findings, or enables one to provide a proper context of one’s findings by exploring possible counterfactual scenarios. Such clairvoyance, however, comes at a cost — the degree of abstraction and simplification. This simplification is done not only for computational, but also for conceptual reasons, however. As Scharkow & Bachl (2017) note, “the challenge is to specify a simulation that is simple enough to be comprehensible, yet realistically models the underlying process of interest” (p. 330). In this regard, we have relied on somewhat idiosyncratic and simplified approaches in our simulations. While we surely acknowledge that our setup could have been constructed in a more realistic and complex way, our ability to accurately simulate human behavior does not necessarily depend on very complex models.

Notwithstanding aforementioned limitations, we believe that our contribution would further prompt researchers both in communication science in particular and in the social sciences more generally to pay more close attention to the issues of systematic and proper validation of automated content analytic methods. Our contribution should be read as a call for a thorough and systematic application of validation procedures – especially the ones involving manually annotated materials as a “gold standard.” Statistical models and algorithms, while being infallible in terms of reliability, are not inherently correct; they are only useful in so far as they can properly approximate researchers’ conceptualizations. And this degree of approximation, we argue, can only be established by thorough and systematic validations. In this regard, it is worth stressing here that automated content analysis should not be just regarded as a cheap alternative to expensive manual coding, and it also takes – and indeed should take – a good amount of time and resources in designing the study and evaluating its performance. Otherwise, the risk of “garbage in, garbage out” appears to loom large. To this end, a little extra time spent on designing proper validation does quickly pay off, and eventually would ensure valid inferences and conclusion drawn from such studies.

References

Aaldering, L., & Vliegenthart, R. (2016). Political leaders and the media: Can we measure political leadership images in newspapers using computer-assisted content analysis? *Quality & Quantity*, *50*, 1871–1905. doi:[10.1007/s11135-015-0242-9](https://doi.org/10.1007/s11135-015-0242-9)

Boomgaarden, H. G., & Vliegenthart, R. (2009). How news content influences anti-immigration attitudes: Germany, 1993–2005. *European Journal of Political Research*, *48*, 516–542. doi:10.1111/j.1475-6765.2009.01831.x

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*, 8–23. doi:10.1080/21670811.2015.1096598

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*, 190–206. doi: 10.1080/19312458.2014.937527

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The ANNALS of the American Academy of Political and Social Science*, *659*, 122–131. doi: 10.1177/0002716215569441

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society.* doi: 10.1177/2053951715602908

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, *41*, 570–606. doi: 10.1016/j.poetic.2013.08.004

Ennser-Jedenastik, L., & Meyer, T. M. (2018). The impact of party cues on manual coding of political texts. *Political Science Research and Methods*, *6*, 625–633. doi:10.1017/psrm.2017.29

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science*, *659*, 95–107. doi: 10.1177/0002716215569192

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297. doi:10.1093/pan/mps028

Grimmer, J., King, G., & Superti, C. (2018). The unreliability of measures of intercoder reliability, and what to do about it. Unpublished manuscript. Retrieved from <http://web.stanford.edu/~jgrimmer/Handbib.pdf>

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, *51*, 2623–2646. doi:10.1007/s11135-016-0412-4

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89. doi:10.1080/19312450709336664

Hovy, E., & Lavid, J. (2010). Towards a “science” of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation, 22*, 13-36.

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*, 48–62. doi:10.1177/0002716215570279

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*, 229–247. Doi:10.1111/j.1540-5907.2009.00428.x

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*, 411–433. doi:10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2008). Validity. In W. Donsbach (Ed.), *The international encyclopedia of communication*. Hoboken, NJ: Blackwell Publishing. doi:10.1002/9781405186407.wbiecv001

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Lease, M. (2011, August). On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Retrieved at: https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewFile/3906/4255

Leemann, L., & Wasserfallen, F. (2017). Extending the use and prediction precision of subnational public opinion estimation. *American Journal of Political Science*, *61*(4), 1003–1022. doi:10.1111/ajps.12319

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*, 34–52. doi:10.1080/08838151.2012.761702

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures*, *11*, 191–209. doi:10.1080/19312458.2017.1317338

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*, 587–604. doi:10.1111/j.1468-2958.2002.tb00826.x

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*, 298–313. doi:10.1093/pan/mpt002

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., … & S. Adam. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*, 93–118. doi:10.1080/19312458.2018.1430754

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2018). (Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, Online first, doi:10.1080/10584609.2018.1517843.

Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. New York: Routledge.

Rooduijn, M., & Pauwels, T. (2011). Measuring populism: Comparing two methods of content analysis. *West European Politics*, *34*, 1272–1283. doi:10.1080/01402382.2011.616665

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*, 761–773. doi:10.1007/s11135-011-9545-7

Scharkow, M., & Bachl, M. (2017). How measurement error in content analysis and self-reported media use leads to minimal media effect findings in linkage analyses: A simulation study. *Political Communication*, *34*, 323–343. doi:10.1080/10584609.2016.1235640

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231. doi:10.1080/10584609.2012.671234

*Table 1.*

*Results of the Systematic Literature Review, Study 1.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method**  **Used** | ***N*** | **Refer to Gold Standard** | **Report Reliability** | **Refer to Validation** | **Refer to Validation measures** |
| Dictionary | 55 | 29 (53%) | 9 (16%) | 27 (49%) | 21 (38%) |
| Supervised ML | 18 | 16 (89%) | 10 (56%) | 15 (83%) | 12 (67%) |
| Total | 73 | 45 (62%) | 19 (26%) | 42 (58%) | 33 (45%) |

***Note***: Percentages refer to share of articles using mentioned method.

*Table 2.*

*Decision Scenarios Based on Observed vs. True Classiﬁcation Performance, Study 2.*

|  |  |  |
| --- | --- | --- |
| **Observed performance** | **True classification performance** | |
| Below threshold | Above threshold |
| Below threshold | True Negative | False Negative (Type II) |
| Above threshold | False Positive (Type I) | True Positive |





Figure 1. Percentage of decision error and relative bias in F1-score (over 1000 simulations per each scenario), Naïve Bayes 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 2. 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 3. 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.

1. Here, we use the term “gold standard” and “ground truth” largely interchangeably, denoting some forms of “objective” (or intersubjectively valid) data that serve as the reference. [↑](#footnote-ref-2)
2. Whereas dictionary-based or supervised learning methods assume already well-defined categories, the unsupervised methods such as LDA topic modeling (DiMaggio, Nag, & Blei, 2013; Maier et al., 2018) generally aim to inductively “discover” new classifications without any human input. Since our primary interest lies in the interplay of human input and machine output, we do not consider unsupervised methods here. Also, our definition (inevitably) exclude automatic approaches of merely *acquiring* data, or traditional manual content analysis but make use of automated procedures in tasks other than the actual coding process (such as in data entry or data management: e.g., Lewis, Zamith, & Hermida, 2013). [↑](#footnote-ref-3)
3. However, a great deal of labor-intensive human inputs is usually required when building and constructing a well-defined dictionary (Muddiman, McGregor, & Stroud, 2018; Young & Soroka, 2012). Due to its labor-intensive nature, recent applications in this area increasingly turn to “crowdcoding”, where the manual labor of highly trained yet few numbers of human coders are replaced with a large number of (little or untrained) crowd-coding workers (Haselmayer & Jenny, 2017; Lind, Gruber, & Boomgaarden, 2017). [↑](#footnote-ref-4)
4. In unsupervised methods, the use of human-coding as a benchmark is a common practice as well. For instance, Lowe & Benoit (2013) directly compare direct human-coding based party position scaling (which serves as a benchmark) against unsupervised scaling methods; they found that the proposed scaling methods can produce largely similar results on par with human judgments. [↑](#footnote-ref-5)
5. In a recent study, González-Bailón & Paltoglou (2015) compare five available sentiment dictionaries against human annotations, yet they do not directly deal with the implications of imperfect reliability in human coding. Scharkow & Bachl (2017), the only one of existing studies that examines the consequences of imperfect reliability in human coding, mainly deal with its implication on “linkage analysis”, but not on automated content analysis. [↑](#footnote-ref-6)
6. This is a different issue from much-discussed “representational bias” (especially within the context of an algorithmic bias), but rather is more closely related to “quality control” (QC) issue in NLP literature (Lease, 2011). We use the term “bias” here to refer a degree of discrepancy between true classification membership that could have been obtained by (theoretically) perfect data and that from low-quality data. [↑](#footnote-ref-7)
7. We do not consider a simple keyword-frequency based studies (e.g., simply counting the number of occurrences of a keyword in a given text) in Study 1 since human inputs play no role other than compiling the keyword list itself. [↑](#footnote-ref-8)
8. Among excluded studies, only 15 studies have used unsupervised learning or other forms of automated content analysis, suggesting dictionary-based or supervised machine learning applications are much more frequently used in social sciences in general. [↑](#footnote-ref-9)
9. A set of replication codes for this manuscript can be found at [redacted for a review]. [↑](#footnote-ref-10)
10. It is important to note that, in reality, untrained manual coders often experience “coding drift” over time (i.e., a low *intra-coder* reliability), and this is often correlated with the low intercoder reliability (which signals a lack of a proper coder training). However, since we are only simulating the variability of intercoder reliability alone, the intra-coder reliability factor was not considered in our simulation. We return this point later in the discussion section. [↑](#footnote-ref-11)
11. Here, we do not consider a scenario where a researcher decides to improve the quality of human coding. This is based on the consideration, as to our argument being advanced here, that a researcher (often incorrectly) assumes that human coding is perfectly reliable and valid. [↑](#footnote-ref-12)