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

Journal:	*Political Communication*
Manuscript ID	UPCP-2019-0138.R1
Manuscript Type:	Original Article
Keywords:	Automated text analysis, reliability, validation, Monte-Carlo simulations

Abstract

Political communication has become the central arena of innovation in the application of automated analysis approaches to ever-growing quantities of digitized texts. However, although researchers routinely and conveniently resort to certain forms of human coding to validate the results derived from automated procedures, in practice the actual "quality assurance" of such a "gold standard" often goes unchecked. Contemporary practices of validation via manual annotations are far from being acknowledged as best practices in the literature, and the reporting and interpretation of validation procedures differ greatly. We systematically assess the connection between the quality of human judgment in manual annotations and the relative performance evaluations of automated procedures against true standards by relying on large-scale Monte Carlo simulations. The results from the simulations confirm that there is a substantially greater risk of a researcher reaching an incorrect conclusion regarding the performance of automated procedures when the quality of manual annotations used for validation is not properly ensured. Our contribution should therefore be regarded as a call for the systematic application of high-quality manual validation materials in any political communication study, drawing on automated text analysis procedures.

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

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 has become the central arena of innovation in the application of automated text analysis to ever-growing quantities of digitized texts. Understanding politics today requires a comprehensive understanding of the ways in which diverse political texts constitute or signify complex political processes. Indeed, given the sheer amounts and the ready availability of such texts, automated analysis procedures have become increasingly useful and necessary. The growing popularity of automated approaches to text analysis is mirrored in dynamic and extensive methodological developments (Grimmer & Stewart, 2013; Hopkins & King, 2010; Van Atteveldt & Peng, 2018), as well as in the application of such methods to a wide range of political communication contents, such as analyses of political newspaper coverage (Young & Soroka, 2012), parliamentary debates (Proksch & Slapin, 2010; Spirling, 2016), congressional bills (Wilkerson et al., 2015), political speeches (Rauh et al., 2017), or millions of social media posts (Barberá et al., 2019).

However, with the growing popularity of such "text-as-data" approaches within the field of political communication, the issue of ensuring the validity of the results has become crucial. To arrive at valid results, text-as-data approaches require proper triangulation of the applied techniques against some "gold standard" or "ground truth" (as some forms of "objective," or intersubjectively valid, measurements that serve as a reference: Grimmer & Stewart, 2013). This is typically achieved by using human inputs ("human coding" or "manual annotations") as a benchmark. This procedure is based on the assumption that humans' understanding of texts (still) outperforms that of machines and that, *if trained correctly*, humans will make the most correct and valid classifications of texts. However, "the quantities we seek to estimate from text […] are fundamentally unobservable" (Lowe & Benoit, 2013, p. 299), and as documented in traditional content-analytic applications (Ennser-Jedenastik & Meyer, 2018; Hayes & Krippendorff, 2007;

Krippendorff, 2004), human judgments are in fact no exception to this general rule. In this regard, devising a high-quality measurement instrument and ensuring good coder training, thereby ensuring an adequate level of data quality in traditional manual content analysis, is deemed *the* standard practice in the field.

Nevertheless, as we argue and empirically demonstrate below, there has been a relative lack of parallel attention to ensuring an acceptable level of quality in human coding when such manual annotations are utilized as a gold standard for the validation of text-as-data procedures. It seems to be relatively rare that human-coded materials are properly evaluated before being utilized as validation materials, and relatedly, the methodological details of validation procedures are not consistently reported in a transparent manner. Arguably, such practice holds major risks for both the interpretation of analyses and consequent conclusions, especially in terms of potential bias when evaluating the performance of automated procedures. However, the precise implications of using imperfect human judgments remain insufficiently addressed. Accordingly, this study concerns both the actual practices of the usage and reporting of human-coded gold standards in automated procedures and the possible implications of different qualities of such material. We assess this issue by relying on a large-scale Monte Carlo simulation. Our contribution should be regarded as a call for the systematic application of high-quality human coding for validation procedures in automated content analysis. To this end, it warns against the improper use of human coding as the benchmark in demonstrating the performance of an automated text analysis approach, and additionally formulates recommendations to improve validation practices. Give the social and political relevance of political communication research, it is vital that text-as-data analyses yield valid results.

The Use of Human Annotation in Automated Content Analysis

Following the standard techniques often employed in political communication research, we define "automated content analysis" (or automated text analysis) as a collection of content-analytic approaches that utilize automated methods to code a large amount of textual data in such a way that the coding itself (e.g., the text classification) is not performed manually, but rather through computational algorithms (Grimmer & Stewart, 2013). Although the term "automated content analysis" in general encompasses a wide variety of forms (e.g., Grimmer & Stewart, 2013; Hopkins & King, 2010; Krippendorff, 2013), our definition inevitably excludes automatic approaches of merely *acquiring* data, data *entry*, or data *management* other than the actual coding or classification process (e.g., Lewis et al., 2013). Instead, we concentrate on two broad and rather common forms – a *dictionary* (lexicon-based) approach and a *supervised machine learning* (SML) approach (see Boumans & Trilling, 2016; Grimmer & Stewart, 2013) – because, as we shall elaborate below, they are rather sensitive to the issue of imperfect gold standards. In this contribution, we focus on their application in "classification tasks," i.e., minimizing the human labor required to classify a large collection of documents into known categories.[1] We also assume that our discussion is mainly applicable to methods that classify texts at the *document* level (e.g., aggregating some word- or sentence-level textual features *within a document* as the approximation of a document membership). We choose to do so since many of the applications used in political communication heavily rely on such approaches (e.g., Boomgaarden & Vliegenthart, 2009; Burscher et al., 2015; Rooduijn & Pauwels, 2011).

There are two ways of utilizing "human coding" in dictionary and SML analysis: on the one hand in the initial development stage (i.e., in constructing dictionaries, or in training SML algorithms), and on the other hand in the "validation" stage, evaluating the classification performances of the procedures (i.e., *post-measurement* validation). Dictionary approaches and SML methods considerably differ in their use of human coding in the initial development stage,

[1] Other frequent aims of automated procedures include: (a) to estimate the location of actors or documents that belong to actors (i.e., *scaling*) in n-dimensional space; or (b) to inductively "discover" new classifications with the help of automated procedures (i.e., unsupervised methods). We do not consider these two in the present manuscript, as the architecture of validation procedures requires a very different setup from what is described here.

which requires problem-specific validation as advocated in the extant literature (see Grimmer & Stewart, 2013; Hopkins & King, 2010).[2] However, regarding *post-measurement validation*, the literature often suggests that post-measurement validation based on "out-of-sample" data represents an ideal architecture of validation (e.g., DiMaggio, 2015; Grimmer & Stewart, 2013; Lowe & Benoit, 2013). Once a researcher has developed an algorithm (or simply used an existing one), separate held-out samples (i.e., data not used to model developing and training) are coded *both* by human coders and by automated procedures, evaluating whether the results from the latter converge into the former. Given that "the validity of a method or tool is dependent on the [specific] context in which it is used" (van Atteveldt & Peng, 2018, p. 87), such post-measurement validation may provide convincing evidence of how well a given tool performs in a specific domain and task at hand, especially when off-the-shelf dictionaries or existing SML classifiers are used in a context in which they were not initially developed or trained. For instance, the results from existing, off-the-shelf dictionaries may be validated against the manual coding of highly trained researchers (e.g., Muddiman, McGregor, & Stroud, 2018; Rooduijn & Pauwels, 2011; Young & Soroka, 2012). Likewise, as exemplified in Scharkow (2013) or in Burscher et al. (2014), a similar approach can be taken for SML methods in evaluating the performance of an algorithm.[3] Although convergent validity against external standards is not the only criterion for establishing the validity of content analytic methods (Krippendorff, 2013),[4] this

[2] A dictionary approach generally relies on extensive human input in developing an explicit coding rule (e.g., simple lists of keywords, Boolean expressions, syntactic parsers, or regular expressions). In contrast, for SML, specific coding rules in manual annotations are in general rarely explicitly articulated. Nevertheless, the algorithm takes such implicit human judgments as the point of reference, and tries to infer the features of data that best classify the text into different predefined categories. See Grimmer and Stewart (2013) or van Atteveldt and Peng (2018) for details.

[3] In unsupervised methods, validations are *conditional* on the classification or scaling produced by unsupervised methods (e.g., evaluating whether direct hand coding or supervised methods can reproduce the suggested findings: e.g., Lowe & Benoit, 2013). Nevertheless, the use of human coding as a benchmark is not an uncommon practice in unsupervised methods as well.

[4] For instance, content analysis can also be validated when the actual sources of analyzed text concur with a researcher's findings (i.e., source-based, *postdictive* validity), or when some theoretically predicted effects of contents actually occur among the audiences of text (i.e., an audience-based, *predictive* validity) when such texts are used in experiments or in the real world.

practice of utilizing human coding in validation primarily owes to the general motivation behind automated approaches (i.e., automating "human coding"; Grimmer & Stewart, 2013), which evidently implies a clear standard for evaluation (i.e., against human coding). However, due to inherent resource constraints, it is rare to see validation occur *after* such classification tasks in practice (nevertheless, for notable exceptions, see the aforementioned studies).[5]

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

The purpose of using a manually annotated gold standard in validation is, as Krippendorff (2008, p. 6) notes, "to confer validity on the otherwise uncertain research results." However, this logic essentially requires that the validity of the chosen benchmark itself (i.e., manual annotations by human coders) already be *well-established*: that such human annotations are, at the very least, sufficiently intersubjectively valid (Krippendorff, 2008). As much of the traditional manual content analysis literature suggests (e.g., Ennser-Jedenastik & Meyer, 2018; Hayes & Krippendorff, 2007; Krippendorff, 2004), manual coding often produces unreliable judgments under a lack of proper instructions or coder training, especially when the judgment at hand requires a nontrivial degree of inferences and subjectivity to classify latent information (Krippendorff, 2013). For this reason, there is little disagreement within the traditional content analysis literature regarding the need for proper "quality assurance" in the form of developing unambiguous coding categories and coder training (Krippendorff, 2004, 2013), as well as the more transparent reporting of those steps (Lacy et al., 2015). In recent years, the political communication literature has embraced these steps in order to seek a standardization of research practices, and this now constitutes a compulsory aspect of any manual content analysis study.

However, when applying automated procedures, researchers appear to put too much trust in naïvely coded human annotations, and the exact methodological details of such validation

[5] Due to its labor-intensive nature in producing manually annotated data sets, recent applications in this area increasingly turn to "crowdcoding" (Haselmayer & Jenny, 2017; Lind et al., 2017). We will return to this point in the discussion section.

procedures involving human inputs are not consistently and clearly reported. As an illustration, in Table 1 below we present a review of studies published in peer-reviewed journals indexed in EBSCOhost from 1998 to 2018, the majority (approximately 74%) of which were belongs to political communication broadly defined. We searched for studies with either dictionary- or SML-based methods (based on titles, abstracts, and keywords), and examined whether they employed human coded materials as a benchmark in validation. If so, we also considered whether methodological details such as intercoder reliability were adequately reported.[6]

Of the 73 studies examined, only about 58% (*N* = 42) referred to some sort of validation. Among these, five studies did not use human-annotated materials, whereas 37 (88% of the studies that reported any type of validation, and about 51% of all studies) relied on human annotated materials for validation. Yet only few of those 37 studies (that reported the use of human-annotated validation) adequately reported the quality of human annotated data: indeed, only 14 studies (37.83% of studies reporting human-based validation, and 19.17% of all studies) provided any measures of intercoder reliability, whereas 23 studies did not report *any* intercoder reliability despite the fact that they relied on human coding. Among the 14 studies that did report the intercoder reliability, six adequately reported Krippendorff's alpha, whereas three reported either the percentage agreement or the Holsti agreement measure alone (in the remaining studies, we found other measures such as Scott's Pi or correlation coefficients).[7] In addition, out of 37 studies, only about half (*N* = 18) reported the number of coders, whereas 19 studies did not report the number of coders and/or the total size of the validation data set, rendering it impossible to judge the quality of the validation procedures.

-- Table 1 About Here --

[6] See the online appendix for detailed information regarding data, coding procedures, coded variables, and detailed reliability estimates.

[7] Given that intercoder reliability assesses the *replicability* of resulting data independently from the extraneous circumstances of the data-making process (Krippendorff, 2013), a proper reliability index requires considerations of coder agreement due to *chance*. However, simple percentage agreement (such

Regarding the reported validation metrics of automated approaches, the results were very similar. Certainly, the most commonly used measures of validity were the widely accepted measures of *precision* (13 cases, *M* = 0.74) and *recall* (9 cases, *M* = 0.60). However, other rather uncommon metrics -- such as intercoder reliability (e.g., Holsti, Cohen's Kappa, Krippendorff's alpha) or correlation coefficients -- were also widely used to report validation. There were only three studies (4% of all studies) that reported *both* Krippendorff's alpha (which signals the proper quality assurance of human coding) and a proper reporting of validation metrics.

The observations of published practices show that reported measures of validation, and especially the quality of human-coded data, are far from "best-practice" recommendations from the traditional content analysis literature (e.g., Lacy et al., 2015; Krippendorff, 2013). It is puzzling and discomforting that such best practices from traditional content analysis seem to be somewhat ignored when it comes to establishing ground truth, and that frequent calls for the proper validation of automated procedures are far from being common practice. Importantly, we *do not* claim that the lack of reporting of methodological details necessarily means a lack of proper quality insurance *per se*. However, a lack of proper reporting of important methodological details nevertheless severely undermines the trustworthiness of the reported validation involving human-coded data as the "ground-truth." While prior contributions on this topic – whether theoretically (e.g., Grimmer & Stewart, 2013; González-Bailón & Paltoglou, 2015; Hopkins & King, 2010) or empirically (e.g., such as in corpus annotations: Hovy & Lavis, 2010; Lease, 2011) – have stressed the need for a proper validation of techniques, there is still strikingly little consistency in *whether* and *how* validation is approached and reported. Indeed, a seemingly widespread practice of conveniently utilizing manual annotation without proper quality assurance – as can be seen in Table 1 – reveals a conspicuous lack of attention to this issue in

as Holsti) or Scott's Pi lack such methodological properties, as do correlation-based measures (see Krippendorff, 2013).

actual research practice.

What Price Are We Paying? The Consequences of Low-quality Annotations

Evaluating the quality of algorithms' automated classification performances is typically undertaken by calculating precision, recall, and F1 scores against manually annotated materials (hereafter termed "observed" performance). If the observed performance of the algorithm is not sufficiently satisfying, such as against some *a priori* chosen threshold, additional steps are sought to improve the quality of the automated procedures (e.g., retraining algorithms, or changing the dictionary).[8] Implicitly, however, this practice treats the observed performance as the *unbiased estimates* of "true" performance (*that could have been observed* against the unknown, "true" standard). From this follows the important question of how well the "observed" performance predicts the true classification performance when researchers use imperfect manual annotations, as well as the size of the potential bias.

The use of (potentially) imperfect, low-quality manual annotation as a benchmark for automatic techniques may have at least two systematic consequences. First, although automated methods themselves are perfectly reliable, imperfect human judgments of validation materials essentially make the ultimate "target" of such reliable measurements radically deviate from the true yet unknown target of inference, rendering them "reliably wrong" on-target. Second, and relatedly, systematically flawed human judgments of validation materials can introduce unknown bias when evaluating algorithms' performance *vis-à-vis* true performance. Nevertheless, most empirical studies appear to pay insufficient attention to this issue. In a recent study, González-Bailón and Paltoglou (2015) compared five available sentiment dictionaries against human annotations, yet they do not directly deal with the implications of imperfect reliability in human coding. Scharkow and Bachl (2017), the only existing study looking at imperfect reliability in

[8] 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 the argument we advance here, that a researcher (often incorrectly) assumes that human coding is perfectly reliable and valid.

human coding, mainly deal with its implication on “linkage analysis,” but not on validation of automated content analysis. Indeed, we are aware of only a handful of studies suggesting tentative relationship between the quality of human coding of validation materials and the evaluations of machine-based classification accuracies (Burscher et al., 2014; Snow et al., 2008).

A Monte-Carlo Simulation Study

Although the general intuition regarding the impact of imperfect human judgment in validating automatic procedure is rather clear, elucidating the factors that affect potential bias, and especially its exact magnitude thereof, in the conclusions of such validation is far less straightforward. In order to systematically evaluate the implications of and to provide a more concrete benchmark for the improper use of human-coded materials as gold standards, we set up an extensive set of Monte Carlo (MC) simulations. MC simulations offer a convenient yet flexible tool for systematically evaluating the relative bias and coverage of a given statistic under certain scenarios (Leemann & Wasserfallen, 2017; Scharkow & Bachl, 2017).

We designed our procedures in a way that would largely mirror the typical approaches used in political communication research, while we systematically varied factors that might affect the quality of human annotation in *post-measurement* validation (see Table 2 below). Here, we assume that the size of the text data is sufficiently large to warrant an automated approach. Therefore, the data in question (exemplarily and hypothetically) cover ten news articles per day for ten news outlets, spanning a total of five years. Accordingly, every single simulation is set to generate 130,000 observations of media articles.

Design and Setup of Monte Carlo Simulations

Ideally, in the creation of ground truth data, two or more trained human coders are assigned to – and independently code – a set of sample documents in order to produce data for intercoder reliability assessment. Once an acceptable level of reliability is reached among coders (typically Krippendorff’s alpha equal to or greater than 0.7), the validation materials are often evenly, yet rarely *randomly*, divided into *k*-subsets, each of which is then annotated only by a single coder (Grimmer, King, & Superti, 2018). Indeed, it is still a very common practice to evenly divide coding tasks by some non-random, natural grouping variables (e.g., by media outlets or simply by order of documents) in manual annotations. Given that coders are treated as interchangeable, any (potentially) remaining coder idiosyncrasies (either coder-specific systematic errors or random measurement errors) are in effect no longer considered, neither in the analyses nor in the interpretations of the findings (see Bachl & Scharkow, 2017, for a detailed discussion on this issue). When there is a sufficiently large number of coders, or each materials are coded by multiple coders ("duplicated coding" as in some SML applications or in crowdcoding: see Lind et al., 2017; Scharkow, 2013), the impact of coder idiosyncrasies – especially random errors – would diminish, as they will cancel each other out as long as the number of coders/duplicated coding instances increases. Nevertheless, remaining systematic errors in coder idiosyncrasies may still introduce bias in gold standard materials with respect to the target of inference, especially for data with a higher level of intercoder reliability (i.e., a systematic deviation from the true target). This is even more likely to constitute a serious issue when validation data systematically differ from the training/test data (or from all data) in terms of their textual features, such as when validation sets are biased subsets of the entire data set, or even come from a different context that only remotely relates to the task at hand.

Following this logic, we specified the following factors that may affect the "quality" of the gold standard (i.e., manual annotations by human coders) and therefore the evaluations of the performance of automated algorithms (also see Table 2): (a) whether validation materials are not randomly divided among coders ("sole coding") vs. all coders independently evaluating the entire body of material ("duplicated coding");[9] (b) the proper sampling variability of validation

[9] In case of duplicated coding of entire materials, we assume the gold-standard materials are determined by the "majority rule" (assuming equal qualification of coders at given reliability level) following the common practice in the field (see Haselmayer & Jenny, 2016; Lind et al., 2017).

materials (e.g., whether validation materials are a random sample of test sets vs. systematically biased subsets); (c) the number of human coders (k = 2, 5, 10, which roughly corresponds to minimum, intermediate, and a large number of human coders); (d) the levels of intercoder reliability (Krippendorff's *alpha* = 0.5, 0.7, 0.9, deemed either low, acceptable, or high); and (e) the size of the validation data set (n = 650, 1300, 6500, 13000, approximately corresponding to 0.5%, 1%, 5%, and 10% of the total data set, N = 130,000). Although one must make practical and logistical decisions regarding these factors in any real-world application (typically by resource constraints), they are indeed crucial in terms of properly ensuring the acceptable quality of manual annotations. The specific cases in these scenarios were chosen to reflect typical procedures and their common variations.

– Table 2 About Here –

Using the above setup, we compared different scenarios in terms of their F1 scores based on imperfect validation materials vs. one based on a "true" standard. In practice, knowing the true performance of an algorithm would be impossible, because the true outcome value of textual data can never be known independently from (potentially imperfect) human coding. However, because we were simulating textual data, we could systematically evaluate the *true* classification performance of automated procedures against "observed" performances from scenarios using varying qualities of human-annotated gold standard materials. In so doing, we could illustrate how different practices of utilizing human coding in automated content analyses adversely affect the relative trustworthiness of the procedures' conclusions.

The final Monte Carlo simulation used 3 (number of human coders, k) × 3 (target Krippendorff's alpha levels in human annotations in validation data) × 4 (N of total annotations) × 2 (sole coding vs. duplicated coding) × 2 (random sample vs. biased subset for validation) full factorial design, with 1000 replications per scenario (N = 144,000). As dictionary and SML methods require different data structures and implementations of coding rules for algorithms, we separately performed two MC simulations for each (see the online appendix for detailed rationales and descriptions of our simulations).[10]

Simulation Results

We first present the mean absolute prediction error (MAPE), defined as $\Sigma\ (F1_{validation} - F1_{true})\ /\ N$, which measures the average degree of absolute bias in observed F1 scores (from the human-coded validated data) *vis-à-vis* true, unknown F1 scores. In essence, this can be framed as a prediction problem – to what degree observed F1 scores deviate from true F1 scores – when using the observed F1 score as the best possible "prediction" of the true F1 score. The MAPE is a commonly used metric for measuring the predictive accuracy of continuous measures and is regarded as one of the most robust measures of such (Hyndman & Koehler, 2006). Table 3 reports the results of ANOVA predicting MAPE (i.e., the mean of APE per 1000 runs) as the outcome of interest, along with the marginal means and contrast for every experimental factor. Finally, we report ω^2, a robust effect size measure commonly used for ANOVA, whose interpretation can be regarded as the percentage of total variance explained by the variance of a factor in question.

– Table 3 About Here –

Both for SML and dictionary-based approaches, three out of five experimental factors appeared to reduce the mean absolute prediction error of observed F1 scores in approximating the true F1 score levels (see Table 3 for details). For the SML approach, we see no overall gain when relying on duplicated coding ($df = 1$, $F = .00$, $p =$ n.s.) to produce human-annotated validation materials, as seen when contrasting *sole coding* (MAPE = .0389) *and duplicated coding* (MAPE = .0390) in Table 3. Similarly, we see no discernible effect of the *number of coders* ($df = 2$, $F = .01$, $p =$ n.s.) across simulation scenarios.[11] In contrast, the level of intercoder

[10] A set of replication codes for this manuscript can be found at [redacted for review].

[11] Importantly, we also failed to find interaction effects of these factors together with the rest of the factors (see online appendix).

reliability (i.e., Krippendorff alpha) presented the largest independent overall effect ($df = 2$, $F = 491.75$, $p < .001$, $\omega^2 = .620$, or 62% variance explained) for SML scenarios, such that hand-coded data with the highest reliability level had approximately half of the MAPE (=.0262) compared to that of the lowest level (MAPE = .0550). When the sampling variability of validation samples accurately reflected the variability of the entire body of data of interest ($df = 1$, $F = 399.73$, $p < .001$, $\omega^2 = .252$, or 25.2% variance explained), the magnitude of error also slightly decreased from .0466 (non-random) to .0313 (random subset). Lastly, the size of the validation data set ($df = 3$, $F = 22.00$, $p < .001$, $\omega^2 = .040$) had small yet discernible consequences, such that the prediction error gradually diminished from .0435 (with $N = 600$) to .0365 (with $N = 6{,}500$). However, beyond that point, the size of the annotation ($N = 13{,}000$), whose marginal mean was indistinguishable from $N = 6{,}500$ cases, had no practical gain of reducing MAPE (we will return to this point in the discussion section).

As Panel A of Figure 1 shows, the intercoder reliability also had an interactive effect on other factors (see also the online appendix for associated ANOVA results). Figure 1 presents the point estimates of the MAPEs, along with their 68% and 95% confidence intervals (respectively in thin and bold vertical lines, corresponding to 1SD and 2SD plus and minus from the MAPE) across total simulation runs per scenario. The pattern suggests that validation data with higher reliability levels had far fewer MAPEs when the size of validation data were larger and were derived from proper random samples (a rather straightforward result), such that the mean levels of the expected discrepancy between observed and true F1 scores were as high as 0.061 (SML) and 0.091 (dictionary) for the smallest non-random validation data with the lowest reliability, but as low as .001 (SML) and 0.016 (dictionary) for the largest random validation data with the highest reliability.[12] Although the benefit of greater reliability in human-coded validation data did not appear to be evident in the non-random sample, increased intercoder reliability was

[12] We present the equivalent plots for the full combination of factors in the online appendix.

readily visible in even the smallest set of validation data (e.g., $N = 600$) whose sampling variability was properly ensured. While both the size of the validation data and the intercoder reliability generally reduced prediction bias, it appears that these two factors largely compensate for one another, albeit only at the moderate (K alpha = .7) to high (= .9) reliability level with proper sampling (we will return to this point in the discussion section).

-- Figure 1 About Here --

For the dictionary approaches, largely similar patterns emerged (see Table 3 and Panel B of Figure 1). Indeed, the size of the validation data set (df = 3, $F = 62.37$, p <. 001, $\omega^2 = .130$), the level of intercoder reliability ($df = 2$, $F = 34.07$, p <. 001, $\omega^2 = .047$), and the sampling variability of the validation data set ($df = 1$, $F = 1025.31$, $\omega^2 = .723$) were all significant in explaining the MAPEs across simulation scenarios. Therefore, the results for both SML and dictionary scenarios consistently suggested that largely the same factors systematically affect uncertainties of the conclusions one can draw from proposed automated procedures.

Our next set of descriptions focuses on a researcher's "decision accuracy" regarding the overall performance of the algorithms. Researchers usually set an *a priori* threshold for desired performance levels (e.g., F1 score equal to or greater than .624), and if the observed performance of the algorithm is higher than this threshold, it is deemed acceptable. Within this context, we defined a decision based on the observed F1 score as "accurate" when this is consistent with a decision based on the true F1 score regarding the performance of an algorithm. Given that this is effectively also a function of the specific threshold that one initially targets, we considered three exemplary values (which we have conveniently chosen) here for the sake of presentation: -1SD (= .483), mean (= .624), and +1SD (= .766) of the true F1 scores in our simulations.[13] For each chosen threshold, we then evaluated the potential decision's (in)accuracy by calculating the

[13] Although the range of F1 threshold values were somewhat arbitrarily chosen, such ranges are nevertheless substantially plausible in practice. Indeed, in our earlier reported review of relevant literature, the overall mean of reported F1 was 0.644, with a range of 0.33 (min) to 0.9 (max).

percentages of simulation runs that a researcher's decision would fall into four mutually exclusive categories of true positive, true negative, false positive (Type I error) and false negative (Type II error), as a function of our experimental factors. Effectively, the proportion of Type I and Type II errors we present below can be regarded as the mean expected error rates based on different combinations of factors one might consider when producing validation materials, providing a reference point of which one can probabilistically expect erroneous classifications under each combination of factors. It may also be seen as an analogue of simulation-based power analysis, evaluating the proportion of cases classified as false positive (alpha), true positive (1 - beta, i.e., statistical power), false negative, and true negative. As the duplicated coding and the number of coders did not have any discernible effects, we collapsed the categories when calculating the percentages.

– Figures 2 and 3 About Here –

Figures 2 and 3 present the potential decision accuracy rates for SML (Figure 2) and for dictionary scenarios (Figure 3). The results across different threshold levels indicate that most decision errors are Type II errors (i.e., false negative cases), being as high as 14.9% in SML scenarios and 9.12% in dictionary scenarios, on average. It therefore appears that the observed classification quality (against manually coded validation materials) of these applications tends to *underestimate* the true classification quality. While the false negative rate generally decreases with all of the experimental factors presented here, the biggest gains again appear to be based on increased intercoder reliability. For low K alpha (= .50), the overall false negative rates across all three F1 thresholds were 11.23% for SML and 5.25% for dictionary scenarios, decreasing to 6.78% (SML) and 3.98% (dictionary, with K alpha = .70) until 3.93% (SML) and 3.04% (dictionary, with K alpha = .90), respectively. This result suggests that if statistical power is the utmost concern when determining whether the classifier performance is acceptable, the best way to achieve such a goal is to ensure high intercoder reliability in manually annotated materials.

For a higher threshold of F1 score, the results indicated an increased risk of Type I errors (i.e., false positive), the magnitude of such error is as high as 5.33% for SML scenarios and as high as 4.28% for dictionary scenarios (both with scenarios with F1 = .766, K alpha = .9, N = 600, non-random sampling). This suggests that the potential decision error in performance evaluation based on observed F1 scores is a much greater issue when a researcher tries to target a higher threshold with a low amount of validation data. Comparing identical scenarios across SML vs. dictionary approaches, it seems that SML scenarios generate slightly more optimistic results in terms of their potential decision error rates. Although speculative, this may be explained by the differences in decision rules between SML and dictionary approaches, as the former can carefully calibrate their predictions, whereas the latter make rather monotonic, deterministic predictions. Regardless, as the size of the validation data set increases, the false positive rate generally decreases in all scenarios.

However, those false positive cases disproportionately *increase* in scenarios with non-random validation data (i.e., a biased subset of entire data for validation materials) whose intercoder reliability is higher. Although it may initially seem counterintuitive, it makes sense that highly calibrated but biased validation materials would “reliably” deviate from the true (yet unknown) target of inference, making them “reliably wrong” on-target. Under the same setup but with randomly sampled validation data, however, the results may suggest that such tests offer the most powerful (in terms of statistical power) *and* the most accurate (in terms of minimizing Type I errors) results for performance evaluation. While this appears to be almost self-evident, ensuring sufficient sampling variability is a rather difficult issue in practice. Indeed, validation materials often come from different contexts or points in time, and researchers seldom worry about whether potentially relevant yet unobserved factors in their validation data accurately resemble equivalent features in the test/training (or entire) data set. Considering the fundamental uncertainty of proper sampling variability for the validation data at hand, we suggest that one must strive to exercise greater caution when evaluating an algorithm's performance, especially with small (in terms of total *N*) yet high-quality (i.e., high reliability) hand-coded data. Nonetheless, all of the false positive rates were generally less than 5% (i.e., traditional 95% confidence level) in such cases.

Discussion

As a subfield at the forefront of methodological developments in text-as-data approaches, as well as its active use of such techniques to answer various questions, political communication has emerged as the central arena regarding the use of automated content analysis techniques. Therefore, the actual practices of the proper validation of automated text analysis in extant research, as well as continued discussions surrounding the issue of possible best practices, have profound implications for the field as a whole. Yet, our review of published research within the field has shown that there is still strikingly little consistency in *whether* and *how* the validation of automated procedures is approached and reported. Indeed, studies often fail to report *any* validation metrics when relying on automated methods. Moreover, even when they do, the quality of human coding (when utilized as the gold standard) is generally not properly evaluated. Against such practice, we attempted to provide further insights into the consequences of using suboptimal quality manual annotations as a gold standard in automated procedures. The results of our Monte Carlo simulation revealed that intercoder reliability (Krippendorff's alpha), the size of the validation datasets, and the proper sampling variability of such validation datasets produce systematic consequences for a researcher's ability to correctly evaluate the classification accuracy of the proposed algorithms. In sum, our results give good reasons for concern about the quality (or rather the *validity*) of conclusions drawn from automated content analyses if the proper quality assurance of gold standard data is not guaranteed.

To be clear, the current study *does not* make the argument that we should exclusively rely on human validations, or conversely, that human validations in general are a problem. To the contrary, one of the main points being advanced here is that humans are not perfect. Therefore, proper validation to ensure the "quality" of manual annotation is essential, especially when it is utilized as the gold standard in automated procedures in political communication research. While the (potentially imperfect) human gold standard is often the best we can get, the results of this study suggest that extra caution must be taken.

Some Recommendations for Improving Validation Practices

No single foolproof solution applicable to every situation exists, and what counts as "best practices" for automated text analysis remains in the early stages of development. However, based on our observations from the literature review and from our simulation results, we can offer some recommendations to help improve the practice of utilizing human annotations in the validation of automated approaches. First, we believe that not every study that relies on automated content analysis needs to use human-involved validation, given that different research contexts and different research questions ultimately dictate whether one should use human validation. However, if the main motivation behind using an automated classification algorithm is to efficiently replace costly human judgments, automated procedures should be validated against equivalent forms of human coding (DiMaggio, 2015; Grimmer & Stewart, 2013). In such cases, researchers should adhere to rigorous methodological standards to the degree expected for traditional manual content analysis (e.g., Hayes & Krippendorff, 2007; Krippendorff, 2013) in preparing a manually annotated validation data set. As with all research, the methodological details (such as measurement details, coding instructions, sampling, coder training, and intercoder reliability) of such validation data should be fully disclosed, enabling readers to independently judge the soundness of the validation procedures, as well as to increase the transparency and replicability of the research process (for instance, see Muddiman et al., 2018; Rooduijn & Pauwels, 2011; Young & Soroka, 2012).

Second, we recommend that researchers pay closer attention to the issue of the proper sampling variability of validation data *vis-à-vis* all the data to which one wishes to apply a given algorithm (especially for dictionary approaches), as well as the intercoder reliability of human coding (especially for SML approaches) for validation data. We have observed that these two factors most strongly explain (in terms of ω^2) the mean prediction error in predicting true F1 scores based on the observed performance of the algorithm on such validation data. Although this is almost self-evident, in practice ensuring proper sampling variability *vis-à-vis* the entire data set is not an easy task, especially when validation datasets are not collected simultaneously (e.g., forward in time, or from different contexts from the data at hand).

Regarding the total size of the validation sample, we noted in both SML-based and dictionary-based scenarios that beyond $N = 6{,}500$ there were no discernable *independent* effects of increased sample size (see Table 3 for details). However, increasing the sample size appeared to have compensating effects on other factors, especially on intercoder reliability. Depending on practical situations, researchers may therefore prioritize a different factor based on relative trade-offs, given the benchmarks we suggest here. For instance, in many political communication applications such as social media postings, the nature of judgments in human coding is very subjective (hence coding practices might not be sloppy, but still have lots of intercoder unreliability), yet easy to scale up in order to achieve a large number of annotations. In such cases, researchers may want to prioritize a larger size of validation dataset over higher reliability levels, as demonstrated in recent examples of analyses of moral intuitions (e.g., see Weber et al., 2018, experiments 5-6), of short texts on Twitter or of comments on news sections (e.g., González-Bailón & Paltoglou, 2015). Nevertheless, we advise researchers to strive to increase the sizes of manually coded validation dataset as large as possible, preferably to more than $N = 1{,}300$ (i.e., more than 1% of all data to be examined), assuming acceptable reliability (equal to or higher than .7). The results of our simulation study suggest that the risk of potential decision errors is substantially higher in smaller manual annotation data, especially when targeting higher

F1 scores for the performance threshold. Nevertheless, the percentages of decision errors regarding the true performance of algorithms appear to reach the acceptable range (less than 5% of Type I errors and less than 10% of Type II errors, or statistical power greater than .9) when the size of the manually annotated validation dataset is equal to or greater than *N* = 1,300 in all scenarios with a reliability level equal to or higher than 0.7. In contrast, when intercoder reliability is low, increasing the total size of validation dataset to more than *N* = 6,500 (i.e., more than 5% of all data) may help to reach the equivalent level of maximum error rates of 10%. Of course, this is a somewhat arbitrary threshold, yet the maximum error rate of 10% roughly represents a balance of considerations between a typical false positive rate (i.e., α = 0.05) and maintaining sufficient statistical power (i.e., 1 – β of 0.90, where β is a false negative rate) employed in the field, as advocated in a recent study by Holbert et al. (2018). While this number only reflects a very rough rule of thumb for highly simplified cases (i.e., a binary judgment involving only one variable), an informed judgment of whether and which types of errors to prioritize – the false positive rate, the false negative rate, or the combined error rates –could be based on the simulation evidence presented here. For more complex and subtle judgments, such as ambiguous latent contents or non-binary judgments, it would make sense to use larger amounts of manual data in validation tasks.

Third and relatedly, we have observed that improving intercoder reliability in human coding offers the greatest net benefit in reducing the magnitude of potential bias in automated tasks relative to a true, unknown standard. While this seems encouraging for many researchers, it simultaneously warns against the prevalent practice of only prioritizing the minimally acceptable level of intercoder reliability without considering other factors. Certainly, such supposed improvements may introduce *increased* systematic errors – as evident in our decision error analyses – especially in cases with very small (e.g., *N* = 600), non-randomly sampled validation dataset. Recent developments in "crowdcoding" for content analysis (Haselmayer & Jenny, 2017; Lind et al., 2017) could alternatively offer promising ways of scaling up manual annotation tasks, thereby increasing manual annotations in validation tasks, if resource constraints that preclude the use of trained coders are high. Nonetheless, the additional issue of appropriate quality control for crowdworkers (e.g., selection of workers based on task-relevant background knowledge, designing proper material presentation and option formats for an online environment, choosing optimal workload/workflow and compensations) can quickly become important (see Lease, 2011, and Lind et al., 2017, for a related discussion on this issue).

Fourth, one should also bear in mind that without proper coder training in improving reliability, human coders often experience substantial "coding drift" over time (i.e., low *intracoder* reliability), and this often goes hand in hand with low intercoder reliability, too. In such cases, the risk of introducing additional random errors due to low intracoder reliability runs very high. However, given that our simulation setup did not take intracoder reliability (rather, only intercoder reliability) into account, our simulation would have produced more "optimistic" results than most real-world, low intercoder reliability scenarios. Consequently, our results should *not* be interpreted as the indication that intercoder reliability or human coders do not matter in the validation of automated procedures.

Having addressed the proper quality assurance of human-coded validation data, we also briefly consider the issue of choosing appropriate validation metrics when reporting final classification performance based on such data. In our review, we have seen that coder-classifier reliabilities or correlation coefficients are widely used as validation metrics. However, as Krippendorff (2013) notes, intercoder reliability itself (or indeed any correlation-based measure) does not concern "truth" (i.e., deviations from a given standard, or *performance* against some benchmark) in evaluating coder (dis)agreement, rendering it an inappropriate metric for validation. In contrast, considering general motivations behind automated procedures (i.e., replacing "human coding"; Grimmer & Stewart, 2013), precision and recall would provide a direct quantification of such "performance" (i.e., the degree of deviations from the standard). While an informed argument can be made for the use of coder-classifier reliability for reporting validation, as there is currently no universal standard for how algorithmic coders should be treated in conjunction with human coders, regarding the computer as the n-th coder is only appropriate under the assumption that human coders *already* produce "reliable" and therefore intersubjectively valid results in coding classification tasks. Furthermore, algorithmic coding requires many pre-processing steps that are unique to computers (i.e., human coders do not require such pre-processing steps), violating the basic assumption of the interchangeability of coders and identical procedures/data in reliability assessment. Nevertheless, without proper human coder training, which would ensure the quality of data produced, reliability assessment with computers as n-th coders does not guarantee the ultimate validity of findings from automated procedures (yet the same would be true of precision and recall when the validity of the human-coded gold standard itself is questionable).[14]

Lastly, our results can be further used for benchmarking the expected discrepancy between observed and true F1 scores in performance evaluation of a given automated algorithm based on the combination of factors we examined here. For instance, for the smallest (N = 600), non-random validation dataset with the lowest reliability of K = .5, the mean expected discrepancy of observed versus true F1 scores was as high as 0.061 (for SML) and 0.091 (for dictionary), according to the summary presented in Panel A in Figure 1. Given that this number represented the absolute difference between the two, one might utilize this information to

[14] For nominal judgements, coder-classifier confusion matrix (i.e., precision and recall) can be directly converted to K alpha level (see Krippendorff, 2013, Ch. 12). As such, coder-classifier reliability (while treating an algorithm as the n-th coder) indeed can be indicative of validity of automated procedures. Yet this nevertheless assumes that human coders produce reliable and acceptable judgements at first place. When human coders and automated classification produce both conceptually "wrong" judgements, showing high reliability between the two does not necessarily mean findings from automated classification is "valid." Nevertheless, the objection we raise here with such practice is not about its use per se, but rather, without assuring the quality of human annotations, it may complicate the conceptual validity issue with a mere reliability between coder-classifier.

construct the lowest bound of target F1 scores by additionally considering such average errors. For instance, if one sets the *a priori* performance cut-off at 0.624 for the smallest (0.1% of all data), non-randomly sampled validation dataset with the lowest reliability of $K = .5$, one would regard a given SML algorithm as good enough *only when* the observed F1 score is equal to or above the .685 (.624 plus .061). Furthermore, if one wishes to apply different algorithms, then one can effectively re-estimate our simulation models but with a proposed algorithm instead, thereby deriving the expected errors under such scenarios.

Limitations and Conclusions

A few limitations should be noted. First, our literature review was limited to prior studies that explicitly contain our search terms in their title, abstract, or keywords. Such a sample, by definition, does not include all potentially relevant articles. Nevertheless, it allows us to make a sensible selection of extant articles that not only use these methods but also advertise that they do so. Second, in the simulations we only considered a binary classification task and a single dimension of validation metrics, specifically recall, precision, and resulting F1 scores. Although this greatly simplified our main arguments and complex simulation setups, non-trivial numbers of existing applications that go beyond such simple classifications, dealing with numerical forms of predictions (i.e., document scaling). Regardless, we reason that our core arguments are equally applicable to more complex forms of automated content analysis as well. Given the additional complexity and difficulties for human coders in such non-binary predictions, achieving acceptable intercoder reliability for manual validation materials in such applications is likely to prove even more difficult than in simple binary counterparts. Therefore, we suspect that the potential problems of "imperfect quality" in human-coded validation materials should be greater (or at least identical) in those applications.

Designing a simulation-based study has afforded us a unique opportunity to observe numerous potential counterfactual scenarios of research processes. When carefully designed, such an approach allows researchers to robustly explore potentially important variability in research practices and their consequences. Furthermore, one of the core advantages of relying on such a simulation-based approach is that it provides an opportunity to formally check the sensitivity of one's findings, or enables one to explore possible counterfactual scenarios. Nevertheless, such clairvoyance comes at a price: the degree of abstraction and simplification. This simplification is done not only for a computational, but also for a conceptual reason -- designing a simulation "that is simple enough to be comprehensible, yet realistically models the underlying process of interest" (Scharkow & Bachl, 2017, p. 330). Although we acknowledge that our setup could have been constructed in a more realistic but more complex way, our ability to accurately simulate human behaviors does not necessarily depend on very complex models.

Notwithstanding the aforementioned limitations, we believe that our contribution can further prompt political communication researchers to pay closer attention to issues associated with the systematic and proper validation of automated content analytic methods, especially those involving manually annotated materials as a gold standard. It is worth stressing that automated content analysis also takes – and indeed should take – a considerable amount of time and resources in designing a study and evaluating its performance. To this end, a little extra effort in designing proper validation is highly worthwhile, enabling valid inferences and conclusions to be drawn. Only in this way is it possible to ensure that results can be considered meaningful for understanding the political processes that are signified through textual data.

References

Bachl, M., & Scharkow, M. (2017). Correcting measurement error in content analysis. *Communication Methods & Measures, 11*, 87-104. doi: 10.1080/19312458.2017.1305103

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review, Online first,* 1-19. doi:10.1017/S0003055419000352

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 & 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, 2(2).* 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 U.S.

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 & 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

Holbert, R. L., Hardy, B. W., Park, E., Robinson, N. W., Jung, H., ... & Sweeney, K. (2018). Addressing a statistical power-alpha level blind spot in political-and health-related media research: Discontinuous criterion power analyses. *Annals of the International Communication Association, 42*, 75-92. doi: 10.1080/23808985.2018.1459198

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

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.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679-688. doi: 10.1016/j.ijforecast.2006.03.001

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.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, *92*, 791-811. doi: 10.1177/1077699015607338

Lease, M. (2011, August). On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Leemann, L., & Wasserfallen, F. (2017). Extending the use and prediction precision of subnational public opinion estimation. *American Journal of Political Science*, *61*, 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 & Measures*, *11*, 191–209. doi:10.1080/19312458.2017.1317338

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

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.

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

Slapin, J. B., & Proksch, S. O. (2010). Look who's talking: Parliamentary debate in the European Union. *European Union Politics*, *11*(3), 333-357. doi: 10.1177/1465116510369266

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254-263). Honolulu, Hawaii: Association for Computational Linguistics.

Spirling, A. (2016). Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832–1915. *The Journal of Politics, 78*(1), 120-136. doi: 10.1086/683612

van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods & Measures*, *12*, 81-92. doi: 10.1080/19312458.2018.1458084

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., ... & Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods & Measures, 12*(2-3), 119-139. doi: 10.1080/19312458.2018.1447656

Wilkerson, J., Smith, D., & Stramp, N. (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science, 59*(4), 943-956. doi: 10.1111/ajps.12175

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

Table 1. Results of the literature review (percentages in parentheses)

	Total
Total record retrieved	192
Excluded	119

Total eligible record	**Total**	**Dictionary**	**SML**
	73 (100)	55 (100)	18 (100)
Refer to any validation?	42 (57.5)	27 (49.1)	15 (83.3)
↳ ***refer to human-coded gold standard?***	**37 (50.6)**	**23 (41.8)**	**14 (77.7)**
↳ (1) *report any intercoder reliability?*	14 (19.2)	6 (10.1)	8 (44.4)
↳ *report K alpha?*	6 (8.2)	1 (1.8)	5 (27.7)
↳ (2) *report N of coders?*	18 (38.3)	10 (18.2)	8 (44.4)
↳ (3) *report N of validation data?*	34 (46.6)	21 (38.2)	13 (72.2)
↳ (4) *report validation metrics?*	32 (43.8)	20 (36.4)	12 (66.6)
↳ *report proper val metrics?*	14 (19.17)	6 (10.1)	8 (44.4)

Note: Percentages denote the share of articles that satisfy the given criteria among all articles employing respective methods.

Table 2. Simulation input parameters

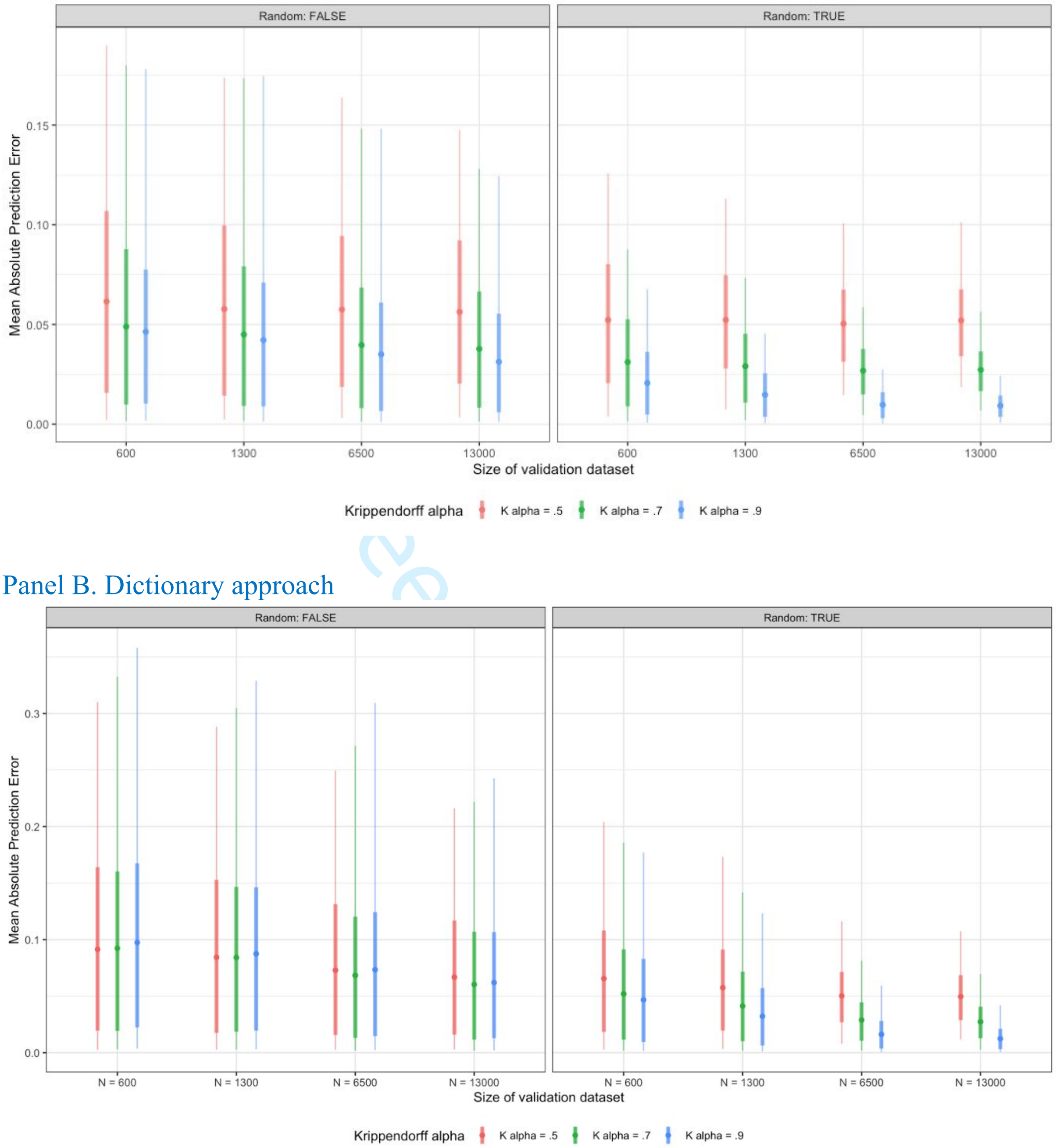
Factors	**Input parameters**
N of human coders	2 (minimum), 5 (intermediate), & 10 (large manual coding)
Intercoder reliability	0.5 (low), 0.7 (acceptable), & 0.9 (high levels of reliability)
N of validation data	600 (0.5%), 1300 (1%), 6500 (5%), & 13000 (10%) of total data
Sampling variability	Random sample vs. non-random (biased) subset for validation
Coding per entry	Sole coding vs. duplicated coding for each entry

Table 3. Simple ANOVAs predicting MAPEs and mean comparisons across factors

Factors	SML	Dictionary
Random sample vs. not	(df=1, F=399.73)***	(df=1, F=1025.31)***
Non-random (biased) sample	.0466 ^{a}	.0410 ^{e}
Random subset for validation	.0313 ^{b}	.0191 ^{f}
Duplicated vs. sole coding	(df=1, F=.00)	(df=1, F=.32)
Duplicated coding	.0390 ^{a}	.0299 ^{e}
Sole-coding	.0389 ^{a}	.0303 ^{e}
No. of coders (k)	(df=2, F=.01)	(df=2, F=.02)
k = 2	.0389 ^{a}	.0301 ^{e}
k = 5	.0389 ^{a}	.0301 ^{e}
k = 10	.0390 ^{a}	.0300 ^{e}
Target Krippendorff's alpha value	(df=2, F=491.75)***	(df=2, F=34.07)***
K alpha = 0.5	.0550 ^{a}	.0340 ^{e}
K alpha = 0.7	.0357 ^{b}	.0289 ^{f}
K alpha = 0.9	.0262 ^{c}	.0273 ^{f}
Size of validation data (N)	(df=3, F=22.00)***	(df=3, F=62.37)***
N = 600	.0435 ^{a}	.0362 ^{e}
N = 1,300	.0401 ^{b}	.0323 ^{f}
N = 6,500	.0365 ^{c}	.0270 ^{g}
N = 13,000	.0357 ^{c}	.0244 ^{g}
Residuals	df = 134	df = 134

Note: *** $p < .001$. Cell entries are marginal estimates of mean absolute prediction error (MAPE) per contrast of factors (total N =144). Within each set of factors by model, different (same) superscripts denote statistically (in)distinguishable MAPEs based on Tukey's post-hoc tests. For instance, for *duplicated versus sole coding* factors, two MAPEs are statistically the same (i.e., their mean difference is not significant), and is therefore denoted with the same superscript. In contrast, for *random versus non-random sample* factors, two MAPEs are statistically different (i.e., their mean difference is significant), and is therefore denoted with different subscripts.

Panel A. Interactive effects of experimental factors, SML scenarios



Panel B. Dictionary approach

Figure 1. Interactive effects of *N*, *K alpha*, and *random* factors predicting mean absolute prediction error (MAPE), SML (Panel A) and dictionary (Panel B) approaches.

Note: All combinations of two- and three-way interactions among *N*, *K alpha* and *random* factors were significant for both SML and dictionary scenarios. All three-way ANOVA results are reported in the online appendix.

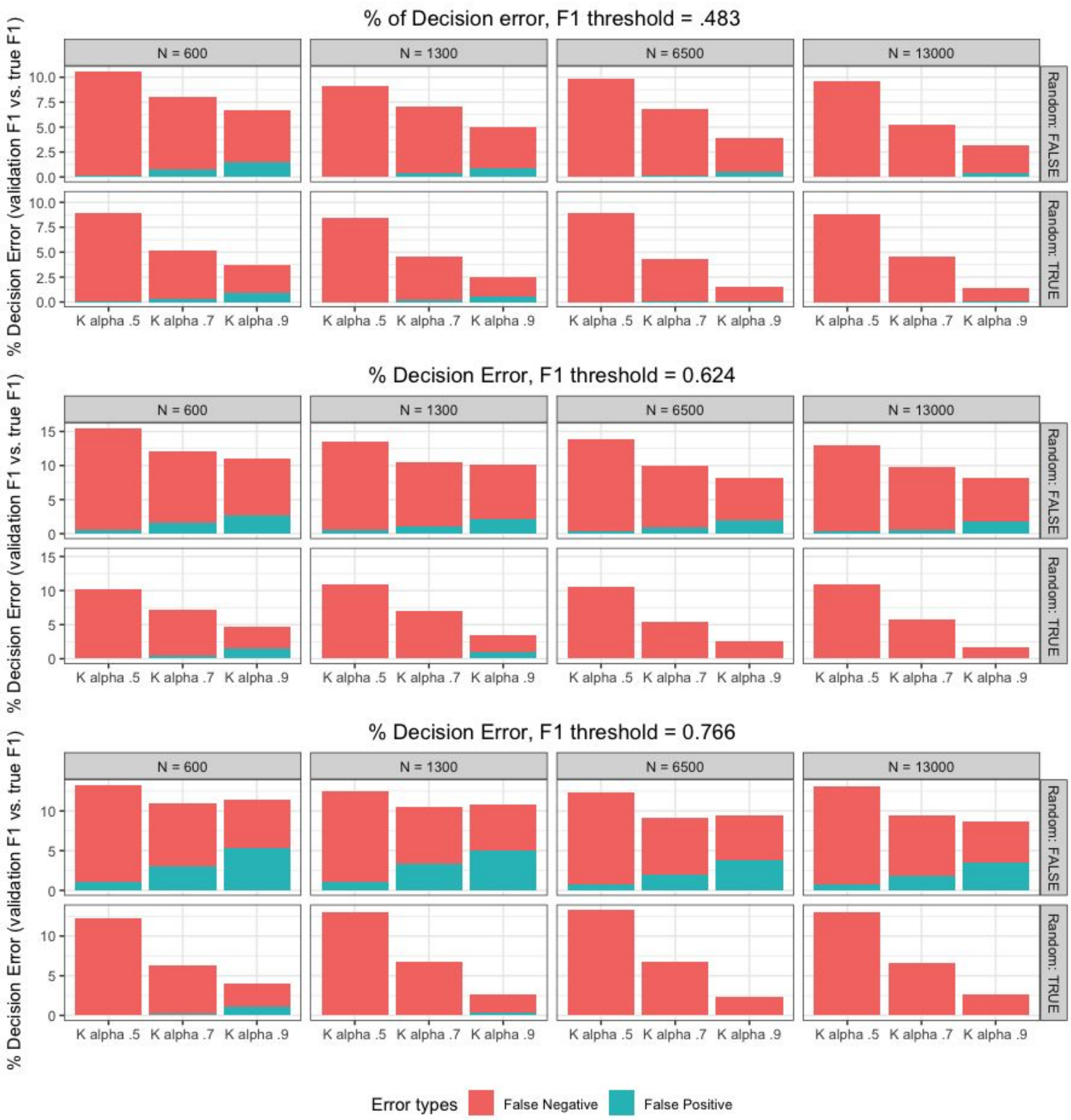


Figure 2. Decision error rate as a function of F1 thresholds, observed F1 score, and true F1 score (models for SML scenarios, N = 6,000 in each)

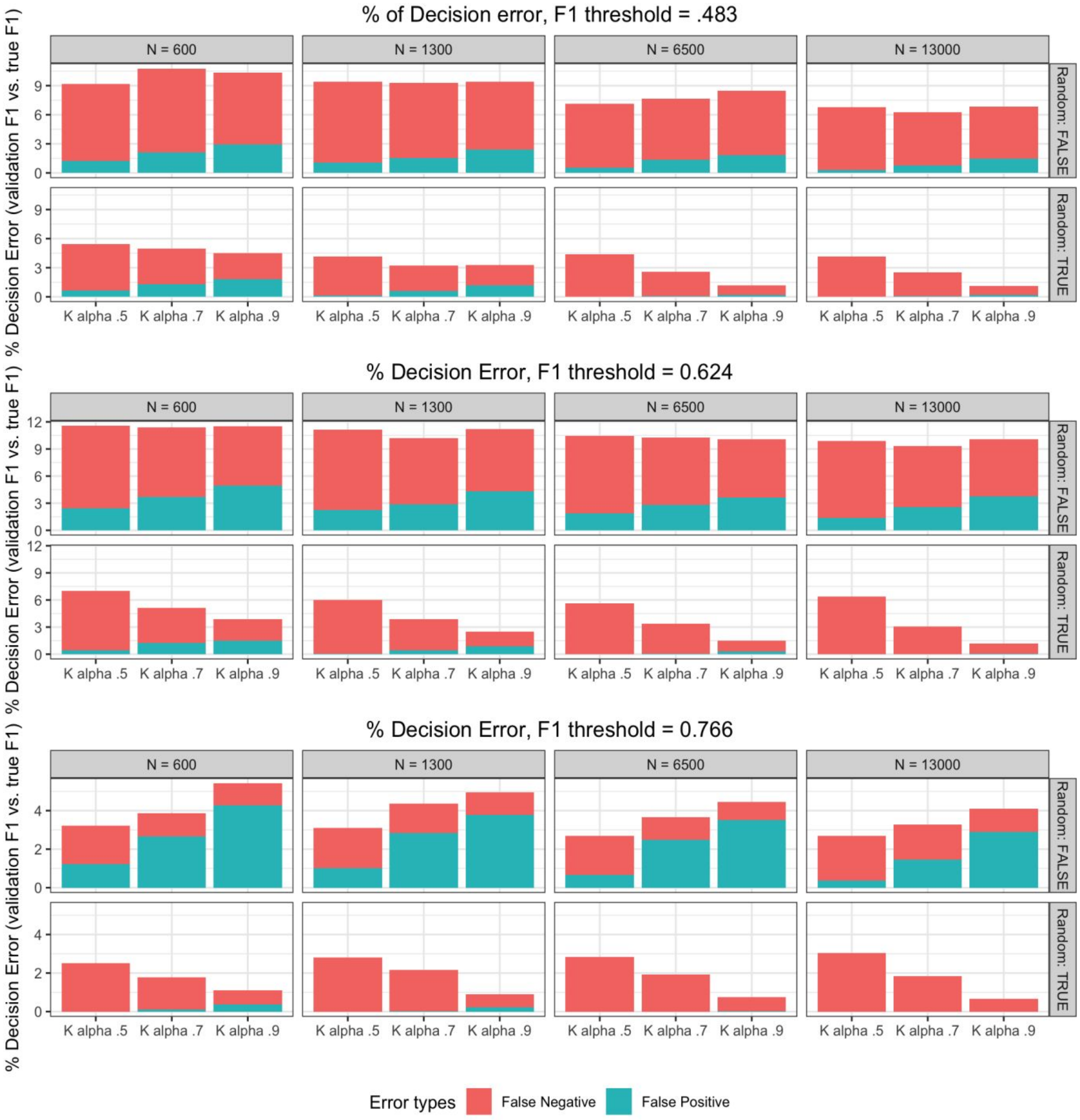


Figure 3. Decision error rate as a function of F1 thresholds, observed F1 score, and true F1 score (models for dictionary scenarios, N = 6,000 in each)

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

Using EBSCOhost databases, we searched all English-language journal articles published between January 1, 1998 and November 7, 2018, 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 was done by examining “Communication & Mass Media Complete,” “Humanities Source,” and “SocINDEX with Full Index” collections.

Among a total of 192 retrieved articles, 112 articles were determined as not relevant (e.g., non-empirical overviews/introduction articles, qualitative analyses, studies using unsupervised methods, or simple keyword frequencies, etc.) and 7 articles were either duplicates or could not be obtained as full texts. These articles were excluded from further analyses. Here, we exclude a simple keyword-frequency based study (e.g., simply counting the number of occurrences of a keyword in a given text, but not actually classifying the documents based on such frequency) since human inputs play no role other than compiling the keyword list itself. 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 general.

A total of five highly trained coders tested the initial coding scheme by independently coding 10 randomly sampled articles (approximately 5% of the total retrieved sample, N = 119) and collectively discussed any coding problems and disagreements. Traditional content analysis literature generally recommends 5% to 25% of all materials to be used for reliability assessment (Lacy & Riffe, 1996). Coding instructions were iteratively revised until the

Online Appendix

coding schemes would produce reliable results. Intercoder reliability (based on Krippendorff's alpha) above 0.75 was ensured for each of the variables coded. Following variables were independently 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

Online Appendix

2. Detailed Setup of MC simulations

Data Generation

We create data (e.g., textual data, such as newspaper articles, to be analyzed) with the "true" outcome value of interest, y (i.e., a classification membership of a given document); the goal of any quantitative text analysis method is to somehow directly approximate this value of y for each observation-level, or instead estimate the unbiased distribution of y at the aggregate level (Grimmer & Stewart, 2013). For the data generating process, we set y at each document level to be randomly generated from three hypothetical independent variables (x1, x2, and x3), all of which stand for some textual features (e.g., words or phrases) of a given document, plus a certain unobserved feature (x0) that is not evenly distributed across the dataset. The values of those variables were randomly sampled from a multivariate normal distribution. In addition, values of x0 were set to be identical across certain grouping variables of media content data, effectively simulating features that are not randomly nor uniformly distributed in the data. This ensures that the results of our simulations are not completely deterministic nor analytically driven to arrive at our conclusion.

SML Scenario. For supervised machine learning approach, we set the true values of y (which is the binary variable) are sampled from a Binomial distribution, with the probability parameter having a very simple linear functional form as follows:

$$y \sim Bernoulli(\pi)$$

$$\pi = \ logistic(\mu)$$

$$\mu = \ \boldsymbol{X}\beta + \varepsilon$$

with ϵ being Gaussian noise added to ensure that each simulation run is not completely deterministic. The β, the true population parameter, was fixed throughout the simulation runs (specifically, $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.2$, and $\beta_3 = 0.6$, which were randomly chosen).

Online Appendix

Following this setup, a single simulation run is set to generate a total of 130,000 observations of media content data.

Dictionary-based Scenario. For a dictionary (i.e., bag-of-words) method, we assume a very similar approach as discussed above, but additionally truncate the values of independent variables to its nearest integer values (i.e., a discrete value), where they represent some "features" of given textual data (e.g., a word) or a combination of such textual features (e.g., a word order or N-grams), in a similar fashion as in Equations (1). Yet for the dictionary-based approach, the vector β was extended to $K = 5$ and their β values were fixed to 0.2. This enables us to better approximate the multidimensionality of textual data, while treating y effectively as a function of the simple sum of the chosen textual features (which is a general assumption that most of the dictionary-based classification methods assumes).

This slight modification for dictionary approaches – truncating to the nearest integers – is due to the fact that each "feature" in the text (e.g., words, phrases, or boolean expressions, etc.) should be "predefined" to be matched against identical forms of dictionaries. We therefore effectively treat simulated integer numbers for three independent variables as each of the predefined categories for textual features, whose scores are simply taken from the existing dictionaries based on some rules. In contrast, for SML scenarios, we use raw continuous normal distributions as is (without rounding up/down numbers) effectively treating them as some kind of a transformed vector dimensional space wherein algorithms try to separate the observations into two categories (i.e., classification membership to be estimated) on that space.

Human Coding

In all scenarios, human coders classify a given observation as "1" (e.g., a text contains the quantity of interest, such as a certain actor, frame, or tonality) or "0" (e.g., does not contain this quantity), based on some observable features of each documents. This human coding (y) can be, in principle, either correct or incorrect against the (unknown) true value, y, therefore behaviors of human coders were modeled by a Binomial distribution with varying probability of successfully categorizing the true data. This enables us to simulate a situation where, at a given target reliability level, some coders produce "correct" judgments while other coders produce "false" judgments more often.

Algorithm-based Classification and Validation

For the dictionary approach, we assume that a researcher utilizes an off-the-shelf dictionary, based on mean valence of observed textual features (e.g., words, phrases, etc.), whose valence scores are taken from the existing dictionary. For the SML approach, we also assume that appropriate, domain-specific annotated materials for a given task already exist for the algorithm development, with a fixed number of training materials ($N = 5000$, approximately 4% of the total dataset being coded).[1]

[1] This means that researchers would only require to produce human coding for validation materials. In practice, when domain-appropriate training materials are not available, one need to produce human coding for training/testing materials as well. Doing so means the "quality" of human coding in such training/testing materials would be the same as validation materials, since one rarely employ different standards for training/testing vs. validation materials in such cases

Online Appendix

3. Additional results referred in the main results.

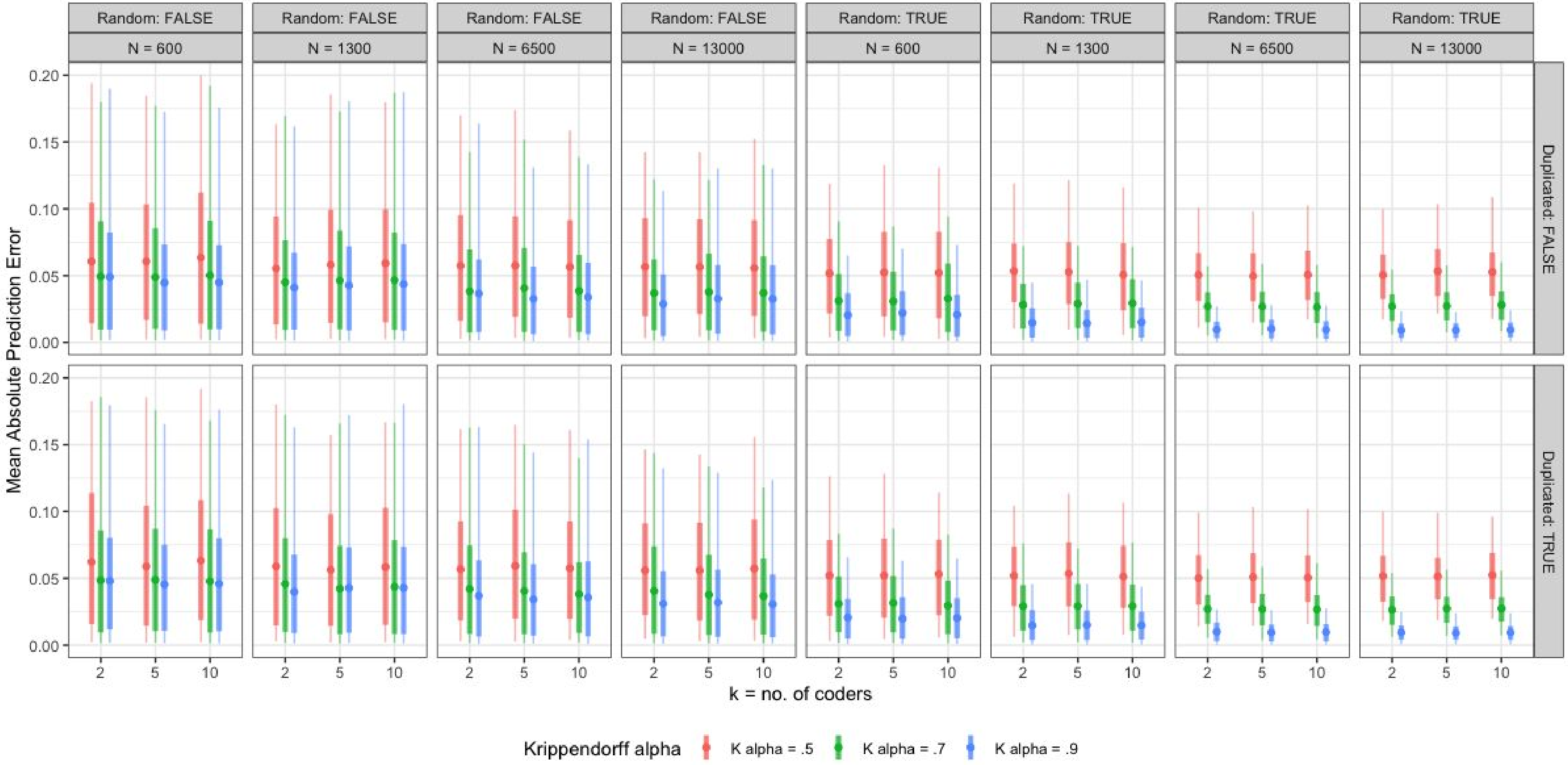


Figure A1. Mean Absolute Prediction Error (point estimate) and their 68% (±1SD) and 95% (±2SD) percentile intervals for every combination of experimental factors, **SML** scenarios (*N* = 1,000 per scenario).

Online Appendix

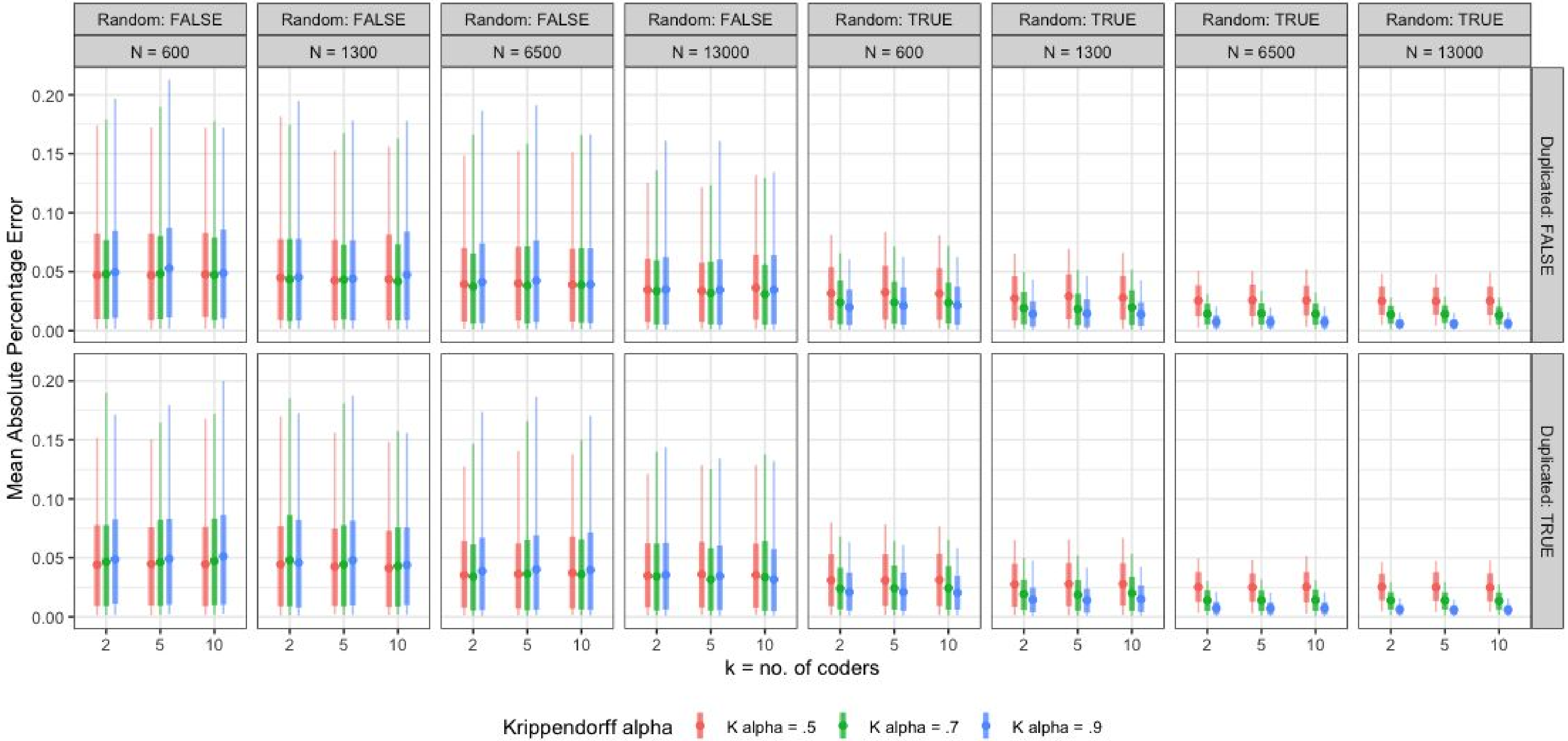


Figure A2. Mean Absolute Prediction Error (point estimate) and their 68% (±1SD) and 95% (±2SD) percentile intervals for every combination of experimental factors, **dictionary** scenarios (*N* = 1,000 per scenario).

Online Appendix

ANOVAs estimating interactions of *number of coders*, *duplicated codings*, and *intercoder reliability* with other factors, SML scenarios.

A 3-way interaction among *intercoder reliability, size of dataset, and random sample*

Factors	**Df**	**SS**	**MS**	**F**	**Pr(>*F*)**
No. of coders	2	.00	.00	.096	.908
Duplicated vs. Sole-coding	1	.00	.00	.012	.913
Size of validation data (N)	3	.0013	.0004	359.497	.001 ***
Target Krippendorff's alpha (K)	2	.0207	.0104	8036.286	.001 ***
Random sample vs. not (R)	1	.0084	.0084	6532.416	.001 ***
K * N	6	.0004	.00006	52.884	.001 ***
N * R	3	.0001	.00004	35.610	.001 ***
K * R	2	.0021	.0010	807.706	.001 ***
K * N * R	6	.00004	.000007	5.551	.001 ***
Residuals	117	.00015	.000001		

A 2-way interaction with the *number of coders*

Factors	**Df**	**SS**	**MS**	**F**	**Pr(>*F*)**
No. of coders (k)	2	.00	.00	.005	.995
Duplicated vs. Sole-coding	1	.00	.00	.001	.979
Size of validation data	3	.0013	.0004	19.853	.001 ***
Target Krippendorff's alpha	2	.0207	.0103	443.801	.001 ***
Random sample vs. not	1	.0084	.0084	360.750	.001 ***
k * Duplicated vs. Sole-coding	2	.000005	.000002	.106	.900
k * Size of validation data	6	.000011	.000002	.076	.998
k * Krippendorff's alpha	4	.000004	.000001	.041	.977
k * Random sample vs. not	2	.000002	.000001	.053	.949
Residuals	120	.002803	.000023		

A 2-way interaction with *duplicated coding*

Factors	**Df**	**SS**	**MS**	**F**	**Pr(>*F*)**
No. of coders	2	.00	.00	.006	.994
Duplicated vs. Sole-coding (D)	1	.00	.00	.001	.979
Size of validation data	3	.0013	.0004	20.769	.001 ***
Target Krippendorff's alpha	2	.0207	.0103	464.284	.001 ***
Random sample vs. not	1	.0084	.0084	377.400	.001 ***
D * No. of coders	2	.000005	.000002	.111	.895
D * Size of validation data	3	.000005	.000002	.079	.971
D * Krippendorff's alpha	2	.000001	.00	.022	.978
D * Random sample vs. not	1	.00	.00	.011	.918
Residuals	126	.002813	.000022		

Manuscript ID UPCP-2019-0138 “In Validations We Trust?”
Response to Reviewers

Dear Prof Strömbäck,

We would like to thank you and the four anonymous reviewers for taking interest in our manuscript, and in particular for the extensive, detailed, and incredibly thoughtful comments. We are glad that all of the reviewers generally agree on the merits of our manuscript, in that it “undertakes the supremely important effort to challenge computational scholars’ thinking about the matter of the validation of automatic classifications by its comparison with human coding” (R2), which is a highly relevant and important issue for the field of political communication. Addressing these critical comments from the reviewers has undoubtedly improved many aspects of our manuscript during the revision. Below we provide a point-by-point response to the reviewers’ concerns (we first address the general points raised by multiple reviewers, by presenting them together, and then follow up with individual reviewers’ remaining comments). We have done our best to respond to the questions and engage with critical remarks in great detail in this document to help alleviate any concerns regarding our general argumentation, the structure of the manuscript, analytical approaches, and further considerations that arise from the results. We have, in many instances, added additional information or clarification in the revised manuscript as well, but not in nearly the same detail that we offer here for the reviewers. Instead, we provide more detailed information in the now revised “*online supporting information*” document (further referred to as *supporting information*) and refer to this document where appropriate in the main manuscript text. In this letter we also reference the main manuscript as well as the supplementary material document, where relevant. We also note that we have highlighted major changes by using colored text in the revised manuscript for an easier comparison vis-à-vis the previous version of the manuscript.

In closing, we hope that through this round of revisions we have made a compelling case for the current approach and our arguments. We believe that the manuscript has greatly benefited from the revisions.

We look forward to your reactions.

Sincerely,
The Authors.

1. General comments raised by multiple reviewers

1.1. General orientations of the paper

The editor and reviewers generally expressed that, as it currently stands, our manuscript seems to take no special interest in political communication but rather reads as a general methods paper. In the first draft of our manuscript, we think the core message of our paper might have been obfuscated by its heavy method focus, without reflecting actual research examples or specific implications for the field of political communication. Following the suggestions, we have revised our manuscript to more centrally reflect the field-specific practices and examples. We hope these changes better highlight our contribution specifically for the political communication community.

1.2. Clarity

All of the reviewers, and particularly reviewer 2 and 3 have urged to be more explicit and clear about our argumentation and theoretical rationale regarding the experimental factors we examine here, preferable architectures of validation (including why they are more appropriate over the others), and our own approach to the review of the relevant literature. Relatedly, reviewer 1 asked what population quantity the Krippendorff's alpha estimates and why they are of interest to researchers, which also speaks to the general idea expressed in the comments of reviewer 2 and 3. Here are some verbatim comments of reviewers:

"*I understand that Krippendorff's alpha has a clear mathematical definition, and has been in the literature for some time, but if the authors are going to use it as a measure, they need to explain what population quantity it estimates. They also need to motivate why that quantity is of interest to researchers. I do not see this point addressed in this paper (or prior cited literature).*" (R1)

"*the authors imply that convergent validity is inappropriate in terms of comparing computer and manual coding (treating the computer as the n-th coder). This is of course an argument that can be made, but then it actually needs to be argued. I was in multiple places uncertain which "architectures" of validation the authors would accept as valid, and why - there seems to be a preference for a two-step procedure, wherein first, (flawed) human judgments need to converge to yield an (approximate, but still imperfect) gold-ish standard, and then, computers' classifications are compared to this standard - but it is not entirely clear why other architectures are considered structurally inferior."* (R2)

"*Hence, if we are to follow that only exactly this architecture is appropriate, that we need sophisticated measures for intercoder reliability and somewhat shallow measures then for precision and recall (which presume the truth of the gold standard and are then effectively percentage agreement measures; I am not really convinced by the very important argument in footnote 10), that needs to be expressly argued."* (R2)

"*The choice of experimental factors is not very thoroughly motivated, especially why both the number of coders and annotations per coder need to be considered, not just the overall number of annotations.*" (R3)

"*I would like to know why the authors expected different effects for both factors, given the input parameters. Moreover, why did the authors include two different supervised learning approaches? Do we have any prior information whether NB or GLM should behave differently in this regard?*" (R3)

"*As additional parameters, consider the sampling of validation and training material, as well as coders. In practice, the validation sample and/or reliability test sample is not always drawn from the same population of documents.*" (R3)

In response, we have generally restructured our literature part along with those crucial factors in presenting our arguments.

To start with Krippendorff's alpha, it estimates the *replicability* of resulting data independent of extraneous circumstances of the data-making process (Krippendorff, 2013), by measuring "chance-corrected" agreement among coders. This is the ability to rely on known interpretations by others, which means researchers must be able to reconstruct the distinctions that coders made among what they were describing, transcribing, or recording in analyzable terms. As Krippendorff (2011) puts it, especially human coders are employed in evaluating, categorizing, scaling, or systematically interpreting the objects of interest (typically textual, visual, or audible matter), numerous extraneous circumstances may affect the outcome of the coding process. When relying on human observations, researchers must worry about the appropriateness of their coding protocols and "quality" of the resulting data – whether the coding procedure itself is able to result in consistent categorization of relevant contents, therefore the resulting data correctly reflect the quantity of interest inherent in textual data, or is it the results of human idiosyncrasies? The key to reliability is the agreement – especially, *independent of* measurement artifacts such as numerical scales, level of measurements, and number of categories or scale points made available for coding – between two or more observers who describe each of the units of analysis separately from each other. The more observers agree on the data they generate, the more comfortable we can say that their data are trustworthy, replicable, and reproducible. The fact that coders reliably reproduce the data (and therefore conclusions from such data) based on well-specified coding instructions serves as the minimum standard of ensuring the proper validity of content analysis technique.

As suggested elsewhere, in actual research practice, other reliability measures are widely used (this is also the case for our own review of published journal articles). Yet as noted in Krippendorff (2011), simple percentage agreement measures (such as Holsti) or Scott's Pi lack proper methodological properties to make them a good "reliability index," as do correlation-based measures. Therefore, we generally advocate the use of Krippendorff's alpha for reporting intercoder reliability (this rather briefly appear in the footnote 7 in the revised manuscript).

For the issue of "proper architecture of validation" and the use of validation metric, we now detail our rationale why we think this particular structure is more appropriate in the revised manuscript (pp. 22-23) as well. It is true that we implicitly preferred two-step procedures instead of one-step procedures (i.e., treating algorithm as the n-th coder), although there is currently no agreed-up standard of how to treat algorithmic coders in conjunction with human coders in such a case. While taking any firm, definitive stance on this emerging issue is not our best intention, the reason why we think a two-step procedure is more appropriate, particularly within the context we are describing in our paper, is as follows:

First, as the logic of validity and validation requires, the reason why we compare one outcome to another "benchmark" is to confer the validity of the "objective" benchmark to the compared outcome under the assumption that those two are estimating the same quantity in data. Quite simply put, if we're not sure what we are comparing against, the coder-classifier reliability itself does not guarantee the ultimate validity of findings. The reason we call for a more thorough quality assurance on human coded data (which serves as the benchmark) is precisely because without such quality assurance, the human-coded benchmark may lose its "validity" it confers to the automated procedures. As we have demonstrated in our manuscript, oftentimes human coding in automated content analysis is used without the proper quality checks, which leads us to wonder against what standard given automated procedures are validated against.

To our assessment, treating the computer as the n-th coder is only appropriate under the assumption that human coders *already* produce "reliable" (therefore intersubjectively valid) results in coding classification tasks. Also, algorithmic coding requires many pre-processing steps that are only unique to computers (i.e., human coders do not require such pre-processing steps), therefore violates the basic assumption of interchangeability of coders and identical procedures/data in reliability assessment. Nevertheless, without proper human coder training (which ensures the quality of data they produce), the resulting reliability assessment with computers as n-th coders itself does not guarantee the validity of findings from automated procedures – again, if we are unsure about against which we are comparing to, we cannot say anything about the ultimate quality of automated procedures. This is conceptually a prerequisite to ensure the soundness and validity of algorithms if we ever want to use algorithms to replace the human judgments (which is often the main motivation to use automated methods in a large classification tasks), therefore the minimum validity of human coded data (which, in effect, best guaranteed by ensuring proper quality) should be established independently from and prior to the comparison of the algorithmic coders against the human coders.

While we think both of the two scenarios would be substantially plausible – as long as a proper quality assurance is guaranteed during the human-coding stage of gold-standard materials –, nevertheless the objection we raise here with practices of treating the computer as simply the n-th coder (therefore only showing coder-classifier reliability) is not about its use per se, but rather, without assuring the quality of human annotations first, it may complicate the conceptual validity issue with a mere reliability issue between coder-classifier. In contrast, the two-step procedure we suggest explicitly encourages researchers to pay attention to the issue of potential pitfalls in utilization of human coding, therefore we think such an approach is more preferable.

Lastly, for the experimental factors we consider here, we generally agree that we have not fully motivated why we have considered such factors. We now present a more detailed rationale in the "Design and Setup of Monte Carlo Simulations" section in the revised manuscript (pp. 10). We also note that we have changed the "the total number of annotations *per coder*" to simply "the total number of annotations" following the recommendation of the reviewers. Also, we have incorporated the sampling of validation materials as suggested by the reviewers.

In doing so, however, we did not consider the sampling of reliability test materials nor the sampling of coders. Once the target reliability level is met, the assumption in reliability estimate is to treat the coders interchangeable and such coders would produce the same quality data given the identical coding instructions. Therefore, we think that these factors are – rather indirectly yet "already" – reflected in our simulation setup.

We also note some major changes we have made in the simulation setup. First, upon reflecting the comments of reviewer 3, we have simplified our simulation into just two categories – one for SML scenarios, and the other for the dictionary approaches (previously, we have implemented three separate MC simulations – one for NB classifier, one for GLM classifier, and lastly, dictionary approaches). The reason for this change is as follows: Here, we are interested in the impact of (potentially) imperfect validation materials on potential errors (in terms of relative bias of F1 scores) and associated decision (in)accuracies, not the absolute level of classification performances of different algorithms (e.g., NB vs. GLM) within SMLs. As long as we employ the identical approaches on validation data and on the entire sample of data, we should be reasonably confident that relative errors and decision (in)accuracies are rather reflective of the quality of the validation dataset, not the absolute differences in classification performances of different algorithms.

Second, we also note that we now rely on an identical data generating process (multivariate normal distribution) for both of the approaches, with slight modification for dictionary scenarios (we first simulate continuous normal distributions, and then round up/down numbers to their nearest integers, rather than simulating them from categorical distributions). As we detailed in the online appendix, this slight modification for dictionary approaches is due to a peculiarity of them vis-a-vis SML scenarios -- that each "feature" in the text (e.g., words, phrases, or boolean expressions, etc.) should be "predefined" to be matched against identical forms of dictionaries. We effectively treat simulated integer numbers for three independent variables as each of the predefined categories for textual features, whose scores are simply taken from the existing dictionaries based on some rules. In contrast, for SML scenarios, we use raw continuous normal distributions as is (without rounding up/down numbers) effectively treating them as some kind of a transformed vector dimensional space wherein algorithms try to separate the observations into two categories (i.e., classification membership to be estimated) on that space.

Third, we now consistently note SML vs. dictionaries, in order to avoid any confusion with the use of "Bag-of-Words" terminology. As all of the reviewers correctly noted, BoW also applies to SML in general, and not all dictionaries are based on BoW either (e.g., taking the order of the words, or window within which they appear together, into account). Therefore, we think "SML vs. dictionaries" terminologies are more appropriate in light of our original intention. We are grateful to the reviewers for pointing out this crucial difference.

1.3. Concrete recommendations

Reviewers 2, 3, and 4 suggested to provide more concrete, actionable recommendations in a way that "*every user can understand the impact of different validation practices on expertable unrecognized error rates and confidence in classifications*", such as "*explain how the relative bias can be effectively interpreted in a hypothetical application*" or "*you do the extra step of explaining what this means in an actual study*" (R2). Relatedly, R3 recommended to provide some estimates of variance components (e.g., % of variance in error is due to X factors), in order to provide more practical anchor points against which one can efficiently allocate resources in producing validation data.

Indeed, one of the motivations to provide overall decision rates (in terms of Type I and Type II error) in our results section is precisely to provide such insights. We have revised the results and discussion section to provide (a) a clearer description of potential error rates arise from the use of imperfect quality manual annotations for validation, (b) and more actionable recommendations, following (roughly) the order of importance of the factors in producing errors in evaluations (from intercoder reliability, sampling variability, and lastly the total size of validation dataset). We also note that (c) researchers can additionally consider mean prediction errors in their decision (regarding the performance of an algorithm) in order to reduce the potential errors in such decisions. Please check the revised section for the details of those changes. We also note that we now present ω^2 statistics when describing our results, whose interpretation can be regarded as the % of total variance explained by a given factor in ANOVA models.

Reviewer: 1 Comments to the Author

1. "*What the authors don't really say is how much is enough. Their own coding scheme is only 75% accurate. Is that good? Is the 25% error enough to overwhelm and reverse the author's conclusions? How would we know if it were one vs the other?*"

RESPONSE: Although there is no commonly acceptable threshold, in most of the manual content analysis literature a Krippendorff alpha value of .667 to .80 is considered to be acceptable (see Krippendorff, 2004; Scharkow, 2013). Besides, our reliability estimates are far higher (alpha = 1.0) than such commonly acceptable value in all but only one variable we examined during the review. For the concerned variable (i.e., whether the study refers to validation/report validation metric), while we surely acknowledge that our characterization of the relevant literature might not be perfect, yet we think it is still good enough to illustrate our points. In the paper, we report on average 45% of studies have reported validation metrics (33 studies out of 73). If we assume that we make on average 25% errors in that variable, it would have been 33% (a worst-case) to 60% (a best-case scenario) studies reporting validation metrics. While we acknowledge that there is no agreed-upon judgmental standard, or whether 60% of all studies is a high enough number, yet we think it still illustrates our point that we as a field can do better in consistently and clearly reporting the validation procedures and how such validation has been approached.

2. "*The simulations provide one small example of what can go wrong. It is quite narrow, not particularly close to any empirical example, or especially probative. The point of it of course is clear. It also helps with the polemic.*"

"*What's the advantage of generating data from multivariate normal distributions when you can start with a real data set. That real dataset will have error in it of course, but even with the error it will be closer a natural dataset than a random number generator can create. You would also wind up making many decisions that are closer to what happens in reality. For example, you have 730,000 observations is a lot more than most datasets have.*"

RESPONSE: We understand, in some fields, the proof-of-concept and empirical validation takes the form of re-analyzing the (previously published) empirical data based on newly proposed algorithms or methods, demonstrating the new procedure outperforms the older. Surely this is one way of demonstrating our point, yet it would be even more strong polemic against the one or several particular examples, might risk us to convey a wrong impression of the motivation behind our paper, let alone such results would be "bound" to a particular context of such example. How can we be sure the results from such an approach is the artifact of the data at hand? Using existing, empirical data for our purpose may bear the risk that our findings would be driven by unknown features of such data at hand, therefore, the generalizability of such an approach would be limited. Besides, relying on an empirical example might lead us to arbitrarily determine a "true" value for each observation based on imperfect human coding, which would further reduce the validity and generalizability of such approach (i.e. our point is that human coded data is not perfect, yet relying on empirical example necessitate the use of such imperfect data in calculating the potential bias). Due to this reason, rather than explicitly single-out certain empirical example, we decided to carry out a Monte Carlo simulation study, which we think is better suited for our purpose here. As a numerical technique for conducting systematic experiments, MC simulation may approximate how certain statistics (such as estimation errors or decision accuracies) would behave under different scenarios. This is particularly useful when the quantities in question (i.e., estimation errors or decision accuracies) cannot be directly derived since its true sampling distribution is not known (which is the case here). While our MC simulation it is still fairly general, the basic set-up and factors examined here (we believe) rather well reflects the most typical research scenarios – which we think demonstrates the utility of our approach.

3. "*So what does a journal do with a polemic? Normally, you just reject it. But in this case, I would point out that the argument here really is important even though I think you could do the whole thing in 5 pages rather than 36. The paper would also be vastly improved if it were 5 pages (or 10 but certainly not longer). I don't know whether the journal would consider something like that but I'd be in favor of something like that. We basically want anyone doing automated text analysis to get this point, without having to burden them with analyses and discussions that are beside the point (if they were all wrong, for example, it really wouldn't matter much to the overall point of this paper).*"

RESPONSE: While we have tried to streamline the entire paper by moving some technical information into the appendix, we think the issue being described here deserves more thorough assessment than simple technical notes or research briefs.

4. "*Were the 10 articles randomly selected? If not how? If so, what was the precise randomization procedure? Let's apply the authors' critique to the authors: if you are making errors ¼ of the time, then we have to ask whether your results provide a clear enough signal to be seen above this noise. What's your evidence for this? What is the nature of the errors? In what way do they bias the results of this paper?*"

RESPONSE: We note that those 10 articles are randomly selected for the purpose of computing reliability. While we could have done it better, alpha of 0.75 to 1.0 considered suitable in most of the cases (also see our response above on this point). While our conclusions and description about the state of the field may have been slightly different based on what we might have found in more "reliable" coding results, the importance of our core arguments (the proper quality insurance for the standard itself) does not change.

5. *The authors need to define "best practices" and their evidence for why they are best.*

RESPONSE: We refer to extant literature advocating the sufficient quality assurance of and the proper use of such data based on quality manual annotations, as well as the consistent and thorough reporting of such methodological details (e.g., Lacy & Riffe, 1996; Krippendorff, 2011).

6. "*the goal of any quantitative text analysis method is to somehow approximate this true value of y". -- This is not the case. We need observation-level measures (which is what I take it the authors mean by y), but printouts of y are rarely of interest. Instead there is some quantity computed from all the data that is of interest (such as an average or causal effect). We don't care about bias in the individual measures if it has no effect on our quantity of interest.*"

RESPONSE: Within our simulation, y means the true classification membership (that has to be estimated via hand-coding, automated coding, or combinations of both). We surely acknowledge that directly estimating the observation-level measure of y may not be the interest of all automated approaches (e.g., estimating the distribution of y instead of direct estimations of each), yet nevertheless it is an important issue, at least as an intermediate step, considering the general motivation behind the automated approaches we consider here (i.e., *classifying* a large collection of documents into known categories).

Reviewer 2 Comments to the Author

1. "*However, there are some questions here. One concerns the matter of convergent validity: The paper (compellingly) argues that convergent validity is probably the best practically available proxy for construct validity in human coding (in fact, it states that it is "one of the principal methods" for approximating validity, which had me wondering which other, competing methods the authors would have in mind and whether there are any useful insights to be gained from these, too).*"

RESPONSE: We note that the validation may take another form (see Grimmer & Steward, 2013, for a related discussion), especially rather than convergent validity against some external standard. For instance, content analysis can also be validated when actual sources of analyzed text concur with a researcher's findings (i.e., a source-based, *postdictive* validity), or when some theoretically predicted effects of contents actually occur among audiences of text (i.e., an audience-based, *predictive* validity) when such texts are used in experiments or in the real world. This is now added at footnote 4 of the revised manuscript.

2. "*Relatedly, the paper laments that studies apply measures developed to determine reliability based on agreement (convergent reliability, if you will; measures that consider category frequencies, chance agreement, and so on) for the second stage and insist that the appropriate measures should be precision and recall - which is reasonable if, but only if, we can assume that there is a known ground truth; yet, if we take the stance that both humans and computers occasionally err, that is not necessarily self-evident, and the notion of chance agreement between a crude algorithm and an inattentive coder is anything but implausible*"

RESPONSE: It is true that we assume there is a ground truth in evaluating the validity of proposed automated procedure in our discussion. This is NOT to say that we agree with the philosophical stance of assuming that there exists unquestionable truth (that a given automated procedure can be evaluated against), yet simply saying that use of such metrics assumes such. As the reviewer correctly points out here, the use of precision and recall is based on such assumption. What we are trying to point out here is that *if* a researcher makes such an assumption, then one should also pay close attention to whether a given standard (i.e., human annotations) adhere to such assumptions by ensuring the proper validity of human annotations as well.

As we detailed in the revised manuscript (as well as our earlier answer to the general comments presented above), we indeed believe a principled argument can be made for the use of coder-classifier intercoder reliability for reporting validation. However, in such a case, one nevertheless needs to ensure validity of the human-annotated standards at the first place in order to claim the validity of automated procedures in relation to a researcher's conceptualization and theory (which is *the* ultimate purpose of such validation). Without such quality assurance, coder-classifier reliability cannot be taken as definitive evidence of the validity of automated procedures. The objection we raise here with such practice is not about its use per se but rather it complicates the conceptual distinction between reliability and validation.

3. "*A related challenge is that the study tosses dictionary based and machine learning based approaches into the same basket without really explaining where their differences and similarities lie. Sure, both enable a post hoc evaluation of how well machine classifications match with separately obtained human annotations, but the study repeatedly alludes also to validation procedures prior to this stage - e.g., by suggesting that imperfect validation of human coding can bias subsequent machine learning models; an argument that transfers rather badly to dictionaries, which are also influenced by flawed human judgments during construction, but in a very different way - and which permit and in fact require validation right during construction. Moreover, the accuracy of dictionary based classifications can be determined at the level of specific recognized constructs localized within a text, whereas SML classifications are more holistic.*"

RESPONSE: Indeed this is a fair point, yet we do not have enough space to discuss such differences in detail (instead we only focus on similarities of two methods in utilizing human coding as post-hoc validation). We instead briefly note that dictionary approaches and SML methods considerably differ in their use of human coding as an *initial input*, and direct readers to additional resources (e.g., Grimmer & Stewart, 2013, and van Atteveldt & Peng, 2018) for the related discussion. Briefly, we see that a dictionary approach generally relies on extensive human input in developing an explicit coding rule (e.g., simple keywords lists, Boolean expressions, syntactic parsers, or regular expressions, etc.), yet for SML, specific coding rules in manual annotations are, in general, rarely explicitly articulated but instead the algorithm takes such implicit human judgments as the point of reference that best classifies the text into different predefined categories. This discussion appears in footnote 2 of the revised manuscript.

4. "*In this vein, but only as a side note, I might note that the paper might also explain a bit better how these two approaches treated are different from the approach excluded (unsupervised models) - which probably has to do with the absence of deductive categories capable of sustaining a ground truth and therefore incapable of formal validation in a comparable form.*"

RESPONSE: As to the case for unsupervised methods, as suggested by the reviewer, we now briefly discuss how validations are typically approached in unsupervised methods relative to dictionary- or SML-based methods. Briefly, we note that validations can be done only *conditional* on the classification or scaling produced by the unsupervised methods (e.g., given suggested classifications, evaluating whether direct hand-coding, supervised methods, or any other methods can reproduce the findings: e.g., Lowe & Benoit, 2013). Nevertheless, the use of human-coding as a benchmark is not an uncommon practice in unsupervised methods as well. This appears in footnote 3 of the revised manuscript.

5. "*Either way, validation by comparison to known true classifications means something quite different during dictionary construction and application (why would one even validate a dictionary after application if it only gets cleared for application after every included rule has been systematically validated?) as opposed to machine learning based classification.*"

RESPONSE: To be clear, the focus of our arguments concerns the post-measurement validation in applied research settings, not the validation of dictionaries themselves *during* the construction.

6. "*A different issue concerns the overall narrative and order of the paper. In its present form, the paper first argues that there is a problem, then shows that there is a problem, and then models the shape of the problem. In my view, that's one step too many; given especially the rather small scale of the empirical verification that the problem exists, together with the a bit sketchy documentation of that part of the study (especially given the paper's otherwise consistent insistence on validation and documentation), I would suggest to not treat this as a study in its own right, but embed this within the theoretical argument: For instance, you could use your knowledge of what is done empirically to organize your critique of why these approaches are insufficient; or you could build a methodological argument and mention where appropriate how common said problems are. In its present form, by contrast, the survey of uses is methodologically massively underspecified. There is no reference to any efforts to validate the search string used to identify relevant texts or the cleaning procedure, there are no formal definitions of classified uses, and no justification beyond a broad reference to "appropriate" documentation, which is coded with high reliability but in a completely non-replicable manner, since the criteria aren't explicated (See also in the appendix and table 1). One could of course add all these things, but for the very small N and relatively straightforward findings, that appears excessive - so maybe it is better to demote this component and integrate it into the main argument. That would also enable you to much more clearly formulate the contribution of the paper, which is currently a bit double barreled - to determine the scope of a problem, and to simulate how much that matters. Why use the existence of a problem - methodologically and empirically - as the setup and then demonstrate the impact of better validation as the key contribution?*"

RESPONSE: As suggested, we have now integrated Study 1 into our theoretical discussion. We are particularly grateful for this suggestion.

7. "*In the present shape, I have some important difficulties following what exactly is entered at what stage into the model, why, and what that means for practical applications.* "

RESPONSE: As suggested, we now more clearly present the factors involving simulation setups (in Table 2), and simplified the discussion of the exact setup in order to increase the clarity. We note that more technical discussion is now moved to the online appendix.

8. "*I do get the overall setup of the model and main data and error generation procedures - but then it gets murkier: For instance, the displayed figures suggest that the propensity of false positives increases substantially with strict alpha values - why is that? Shouldn't stricter criteria depress both type I and type II errors? Is there any explanation for higher error rates for stricter alpha values for seven coders (bag of words approach), 100 annotations (GLM; by the way, why is this labeled as GLM here and the other option as BoW when the approaches were previously introduced as SML and Dictionaries?), ...? Much more importantly, can you somehow convey what that really means? For instance, if I use 100 annotations by two coders with alpha=0.09 (GLM), your model displays almost 10% error rate, mostly false positives; but this is all probabilistic, it is completely possible that these annotations are in fact near-perfect and you get a very strong ground truth, or that the annotations are somewhat off, or reliable but invalid - so how exactly does this 10% error rate enable me to tell my probability of erroneous classification if I validated my classifier as described?*"

RESPONSE: We reason that a more stricter alpha value can indeed increase the Type I error (i.e., false positive), since highly calibrated yet biased validation materials would "reliably" deviate from the true (yet unknown) target of inference, making them "reliably wrong" on-target. This is more evident in non-randomly sampled validation dataset, and this becomes more prevalent in higher thresholds for the performance standard (e.g., when a researcher sets a higher F1 score as the cutoff points for acceptable algorithm performance). In order to better illustrate this point, we now present three threshold values (instead of one) in the revised manuscript.

We also now clearly note that our results may provide some concrete reference points of which one can expect an average degree of discrepancy between observed vs. true F1 scores based on combination of factors we examined here. For instance, for the smallest (N = 600), non-random validation data with the lowest reliability of K = .5, the mean expected discrepancy of observed vs. true F1 score is as high as 0.061 (for SML) and 0.091 (for dictionary, respectively) according to the summary presented in Panel A in Figure 1 above. Considering this information, if one sets the a priori performance cutoff as 0.624 for the same combination of factors for the validation data, for instance, then one would treat a given SML algorithm to be good enough *only when* the observed F1 score is equal or above the .685 (.624 plus .061). This discussion now appears in the revised discussion section.

9. "*Also the explanation of Figure A1 is insufficient, especially the stark difference between the GLM and BoW displays, and the absence of effects in the BoW part in particular - I think I get what this means, but I am not sure, and if I am right, this is trivial, since validation feeds back into the classification in a very different way in dictionary or SML based classifiations.*"

RESPONSE: We now excluded this Figure from the manuscript.

10. "*Relating to your discussion, where you conclude that reliable human coding "sometimes" offers benefits, this vagueness massively depresses the utility of your paper for readers, who will get from it that one should validate (I do not really follow where the 1000 texts come from in your discussion; shouldn't there be something more akin to a power analysis so you can estimate based on some basic properties of your study and material how big the problem/need for validation is?; likewise, you say that not every study needs human validation, but you say nothing about when this might be mandatory or optional), but are left alone determining just how much validation is appropriate and necessary depending on the setup of their study, or how much the incurred hidden error might be if this is not done.*"

RESPONSE: We appreciate the reviewer's criticism to be more clear in our recommendations, and we have largely followed your suggestions. First, we now clearly state that when human validation should be used -- "if the motivation behind the use of automated classification is to efficiently replace the (costly) human judgements, it clearly implies that automated procedures should be validated against the equivalent forms of human coding (DiMaggio, 2015; Grimmer & Stewart, 2013)."
Second, (as correctly noted by the reviewer) we now clearly state that the decision error analysis we present is very similar to, therefore can be regarded as, the simulation-based power analysis (see page 16 of the revised manuscript about the details). Likewise, (again, as correctly noted by the reviewer) we note that MAPEs we present here can be regarded as the mean expected error rates based on different combination of factors one can consider in producing validation materials.

11. "*One, dictionaries are not all BoW. Many dictionaries include proximity or position rules, so they require word order; and there are lots of BoW approaches that are not dictionary based.*"

RESPONSE: We have changed and clarified that in our terminology. We appreciate this suggestion to be clearer in our explanations/definitions.

12. "*Two, I am unsure what you count as one error. If I have a text that contains a reference to a construct, and that construct is not recognized by a dictionary where it is referenced, but it instead falsely recognizes the construct in a different place where it is not referenced, is this no error (document correctly classified as "containing the construct"), one error (one document that contains a classification error on this construct), or two errors (one false negative instance, one false positive instance)?*"

RESPONSE: As we understand, this relates to the units of analysis -- what the reviewer describes happens when we define units of analysis as the word or "feature" level. In contrast, we assume all errors are counted at the "document" level, assuming we only predict "document membership" as the outcome of interest. So if one construct is missed (false negative) and the other is incorrectly referred (false negative) "at the feature level," yet nevertheless the document is correctly classified, then this would not be counted as any error. The assumption of document-level unit of analysis is crucial in understanding our setup, so we explicitly declare early in our manuscript that we assume the unit of analysis is document-level.

13. "*Three, in your discussion you note the difficulty of evaluating the reliability of non-binary classifications; but most SML classifications pass through a nonbinary stage, which is then binarized by just adopting the most likely category.*"

RESPONSE: The difficulty we refer to here is rather the difficulty of "human coders" in producing more fine-grained judgment (compared to yes/no answers), not algorithms' abilities to do such (also note that we refer to this "reliability" problem – implying human coders -- in such non-binary judgments).

14. "*Four, you mention crowdcoding; but you argued before that precise instructions and trained coders are essential, two properties that are very hard to achieve in crowdcoding once constructs get complex. Isn't this reliance on mass rather than accuracy and reliability exactly contradicting your argument?*"

RESPONSE: The quality control in crowd-coding generally slightly differs from manual content analysis involving few trained coders (e.g., selection of workers based on task-relevant background knowledge, designing proper material presentation and option formats for an online environment, choosing optimal workload/workflow and compensations). Also, as we present in our revised simulation setup, we see that increasing the total number of annotations (via crowd-coding) can effectively compensate the lower reliability level.

15. "*Five, I was confused in several places whether the errors you consider are random or systematic (notably, page 18, page 7), and in places where systematic errors were suggested, I was not always certain why these would be systematic - algorithms make systematic errors, but not all human error is systematic.*"

RESPONSE: We now make clear that coder idiosyncrasy may involve both random measurement errors and systematic errors.

16. "*Six, does the argument consider that there may be instances where there is no unique ground truth but where the same text supports more than one legitimate classifications?*"

RESPONSE: While we do believe there might be a situation where one text affords different interpretations, such differences would be better reflected in the distinctive coding instructions (therefore the same text would generate different data depending on coding instructions), not in multiple possible classifications given the same coding instructions. At best, such a situation would rather signify a lack of proper coder training, or at best interpretation of the same text would be highly contingent upon receiver characteristics.

17. "*Seven, wouldn't it be nice to name those few studies who did a good job validating following your survey, to mark the best practices so far?*"

RESPONSE: We appreciate this suggestion. We now have cited several studies where appropriate, explicitly acknowledging such studies as the best practice examples.

18. "*Eight, I am not convinced by the calculations of annotations per coder, do these consider that some material is coded multiple times by different coders?*"

RESPONSE: To clarify, in our previous version of the analysis, we have assumed that each of the materials are only coded by a single coder, not by multiple coders. Following the suggestions, we now have considered such cases (denoted as "duplicated coding" where each entry is coded independently by multiple coders, then consensus among coders -- such as mean or mode -- is taken as the "correct" categories).

19. "*Nine, can you report the range of coders involved? Ten, on page 9, 10 texts are not 5% of 73.*"

RESPONSE: In our own content analysis of published findings, all variables were independently coded by five different coders throughout, all of whom were adequately trained for the purpose of the analysis.

We first retrieved a total of 192 articles, and among them, we have randomly sampled 10 articles (approximately 5% of the "total retrieved" sample, not based on final sample we have analyzed) for coder training and reliability testing. This is also used for the "relevance" variable (i.e., the identification of whether empirical text analysis is conducted or not), determining whether a study should be excluded or not for further analyses. We have corrected our description in the relevant parts in the online appendix.

20. "*Eleven, most footnotes struck me as relevant enough to be part of the text, although partly shortened (fn11 is the first one I noted as a footnote that is duly a footnote).*"

RESPONSE: We have incorporated some of the previous footnotes into the main text, yet please note that the journal imposes stricter page limits, so we are bound by length limitations.

21. "*Twelve, you cannot put figures referred to in the text into the Appendix (A1)*"
22. "*There are several issues with terminology*"

23. "*The paper needs some more structuring by subtitles and paragraph breaks to better guide the reader, and there are many very long sentences that need to be broken into legible parts or otherwise clarified.*"

RESPONSE: We have restructured our main arguments and presentation, as well as done a careful rewrite to be clearer in our presentation. We thank the reviewer for the extensive suggestions.

Reviewer 3 Comments to the Author

1. "*Also, what are the three x variables? Please clarify this.*"

RESPONSE: Those refers to hypothetical independent variables that representing some textual "features" (such as certain words, syntax, etc.) which we assume to be principally related to the would-be true classification membership for each document. We have clarified this point in the revised manuscript.

2. "*Also in the analysis, you quickly move on from the direction of bias to the absolute amount and its variance. Looking at the NB (GLM) vs Dictionary approach, it seems that one is too optimistic, the other too pessimistic about classification performance. Is this an expected outcome, and can you explain why? You only briefly mention this on p. 19, but I am curious about this result.*"

RESPONSE: We agree with the observation of the reviewer that SMLs appear to generate slightly more optimistic results than dictionary approaches (please note that due to changes in the simulation, the differences are now that so dramatic than it was in the previous version). We conjecture that this is due to SMLs ability to more highly calibrate their predictions than dictionary approaches, the latter of which is rather deterministic in its decision rule (e.g., a linear combination of all available features found in the documents, etc). Yet we refrain to strongly ascertain to do so since this is hard to substantiate at this point with our particular approach here. Instead, we have briefly mentioned this in the results section, additionally noting that this is largely speculative.

Reviewer 4 Comments to the Author

1. "*Unfortunately, I didn't feel like the simulation produced actionable information. It essentially just establishes that when we have errors in our human coders we end up with errors in our outputs (either the prediction from a contaminated training set or our assessment of our own accuracy in the cases of dictionaries). The direction of the result isn't really in question and the simulation is sufficiently abstracted from the real world that I don't think it gives us a sense of how bad the problem is.*"

RESPONSE: While we agree with the reviewer's assessment that the direction of the result isn't really in question, we think the core contribution we make is to exactly quantify such bias and decision errors with a systematic investigation of relevant factors. Besides, based on the suggestion of the reviewer (and also of other reviewers), we have greatly expanded the recommendation part to provide more actionable guidelines.

2. "*Additionally, I found the study quite hard to read in numerous places. I couldn't quite figure out what was meant by 'inner' and 'outer' cross-validation. On page 10, I read the last two sentences of the first paragraph (starting with "In sum, the most common") a few times and couldn't work out the distinction between the first Krippendorff's alpha number and the second. The transition that immediately follows that "Regarding the validation of automated approaches, results were very similar" was confusing because the prior paragraph was about uses of automated approaches (dictionary methods etc.). I had these kind of confusions throughout the paper and often found the results sections hard to follow honestly.*"

RESPONSE: We apologize for the confusion and tried to make our language more accessible as possible in our description of approaches. Now all of the technical details are placed in the online appendix, and we hope this change will help better orient readers to our core message without being lost in details.

3. "*The authors write that they "know of only one study relating quality of human coding and machine-based classification accuracy." There is a whole literature in CS on crowd-based linguistic annotation (see as a starting point 'Cheap and Fast - But is it good' by Snow et al.)*"

RESPONSE: We have added suggested literature in relevant place. However, we also wish to note that those studies (e.g., the paper by Snow et al.) also may fall short in terms of ensuring proper reliability/validity of human coding based on "expert" annotations. They generally compare the quality of human coding in expert vs. crowd-based annotations, yet they nevertheless treat the coding from a "single expert" as the given standard, and rarely examine whether multiple experts indeed agree with each other. As Krippendorff (2013, in Ch. 12) notes, such single expert annotation is essentially non-reproducible, let alone two or more expect often disagree with each other more than non-expert coders.

4. "*In footnote 9 it feels like a stretch to say that your literature review tells us what is most common in "social science" after choosing databases that deliberately limit you to the communications literature.*"

RESPONSE: We believe this was a misunderstanding. We have clarified our language throughout the manuscript.