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 *ω2* 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 validityof 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.