Dear Professor Jesper Strömbäck,

We would like to thank you and reviewers once again for thoughtful comments. We are glad that, at this stage, reviewer concerns seem to be focused on making sure we are absolutely clear in describing the details of our approaches. Below, we provide our responses on the remaining reviewers’ concerns. We hope that this last set of revisions is considered suitable.

We believe that the manuscript has greatly benefited from the revisions, and look forward to your reactions.

Reviewer 2 comments:

**Q1.** “*the first sentence of the abstract and the paper now makes a claim that is a tad bolder than I would consider warranted - there are lots of fields in and beyond the social sciences that work with automated methods to study text, and even within communication, it is one thing to say there's a lot of work on this in political communication, but quite another to say that it is "the central arena of innovation"*.”

**RESPONSE:** We agree with this assessment and toned down our claim, by clearly denoting that political communication “has increasingly become *one of the*” central arenas of innovation for automated text analysis, along with other fields of studies in social science and computer science.

**Q2.** “*in footnote eight you exclude a scenario - but if I get it right, this is not so much excluded but rather conceptually prior to the point when your study jumps in, as it only applies once a researcher has decided that human coding is now good enough to serve as ground truth...?*”

**RESPONSE:** This is indeed correct, and thank you for pointing out that this is not really an excluded scenario but conceptually prior to our arguments. To reflect this point more clearly, we have revised our description in footnote eight clearly stating such.

**Q3.** “as you briefly raise a potential explanation for differences between dictionary and sml approaches in the findings, would it not be appropriate to have said something about differences that \_might\_ shape how they respond to imperfect validation already when you introduce your choice to focus on these two? it makes sense to focus on these two as they are indeed common, but you make little of the comparison, maybe three sentences early on can make this clearer?”

**RESPONSE:** As we understand, it seems that the reviewer thinks that we need to include some potential factors that differ across dictionary vs. SML approaches in motivating our selection of the two onset of the study when we introduce such choices. It is true that we make little of the comparison of the two but just include a rather brief mention of the differences in footnote 2 (as suggested by another reviewer in the previous round of the review). While we generally have no clear-cut, definitive explanation of why they exhibit different behaviors relative to imperfect validation, we think one of the potential reasons of why they \*may\* differ can be that dictionary algorithms makes rather monotonic predictions based on fixed algorithms while SMLs can make highly calibrated predictions based on additional materials on which prediction algorithms are trained. We additionally include this in our explanation of the motivations behind our choice of two methods (pp. 4) as suggested.

Q4. “can you make all figures such that they are readable as standalone (e.g., spell out what "random" means)?”

**RESPONSE:** We have added more necessary information needed to interpret the figures for its own as a stand-alone. Thank you for your suggestion.

**Q5.** “*relatedly, for comparison it may be preferable to use consistent scales for the graphics? I know that costs whitespace but visual impressions are powerful, and your argument is not only about the decrease in error (which one can see on any scale) but also the size of the error. that way, your sentence on p. 17 (comparing identical scenarios...) would also tally with the visual impression.*”

**RESPONSE:** We agree with the suggestion and modified our graphs as suggested, in a way that they all have identical scales for the graphics. We thank you for this suggestion.

Q6. “finally, one thing for the discussion, I was wondering: would you expect that the uneven distributions of categories in the ground truth should make an important difference (like, when 95% of cases are class A, such that random correct classification is likely)? No need to test this here, but if you can make an argument why it matters or doesn't, that might still be of interest (and would relate neatly to the rationales for k alpha)”

**RESPONSE:** We surely think such factors may additionally affect the overall accuracy of the validation in a practical situation, yet at least theoretically, K alpha is largely immune to unequal distributions (or lack of variability) of available categories in coding tasks (see Krippendorff, 2004). This also means that once we achieve higher K alpha, the likelihood of “random” correct/incorrect classification due to random chance generally diminishes.

In our simulation setup, we indeed kept track of overall true prevalence of y = 1 cases (relative to y = 0), and the results suggest that an uneven distribution of categories in the entire raw data (NOT ground-truth data generated via human coding) does not make any important difference in terms of the accuracy of predictions (zero-order *r* between accuracy and prevalence: -. 049 for dictionary approaches, and .102 for SML approaches, where “accuracy” is defined as the *abs(F1\_true - F1\_validation)*, as in pp. 13 of our main manuscript). While we did not keep track of the same information for the ground-truth data, at the very least we would expect largely the same pattern of results as long as the validation data well reflect the entire data to be analyzed (i.e., the *random sampling* factor). For non-randomly sampled data on which ground-truth data is generated, we would expect that uneven categories may additionally bias the ultimate accuracy of the predictions, especially under the low reliability levels. We include this (although very briefly) in the revised discussion section as suggested (please see pp. 20 for details).

Reviewer 3 comments:

**Q7**. “*I think terms like Type I/II and alpha/beta error rates as well as power should be reserved for NHST/inferential statistics. Although I agree the analogy is quite strong, I expect it will confuse readers, because we are mostly not (never?) testing F scores or other reliability coefficients against predefined thresholds. Therefore, why not stick with false optimistic/pessimistic or false positive/negative estimates (although the latter might be equally confusing since they have a known definition in precision/recall coefficients.) This is not a major concern, but rather a modest suggestion about semantics.*”

**RESPONSE:** While we do not necessarily think that TypeI/II error terminologies should be confined to NHST only, we agree that using more consistent terminology would less confuse readers. We have used the “false positive/negative error” terminology for denoting decision error rates as suggested.