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

Hyunjin Song, Petro Tolochko, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, Sebastian Galyga, and Hajo G. Boomgaarden

Department of Communication, University of Vienna, Austria

Draft date: November 29, 2018

Paper prepared for Inaugural issue of Computational Communication Research

Draft in progress. Please do not cite without permission.

A replication code for this manuscript can be found at:

http://github.com/revelunt/CCR-When-does-garbage-stink

Author Note

Please direct any questions and inquiries to: hyunjin.song@univie.ac.at

Abstract

Automated text analysis methods become increasingly popular for analyzing texts in the social sciences. However, with the growing popularity of such approaches, the issue of the validity of obtained results and the conclusions drawn from them becomes crucial: Blindly applying automated approaches, and hence feeding algorithms into the machine without proper validation, may result in misleading or even plainly wrong findings. Against this backdrop, this study first presents a systematic review of articles published in the major social science journals in the past 20 years, showing that the current practices of validation are far from being acknowledged in the literature and that the reporting and interpretation of validation procedures differ greatly. In a second step, we assess the previously unexplored connection between the quality of human judgment and relative performance of automated procedures (against true standards) by relying on large-scale, systematic Monte-Carlo simulations. Results confirm the expectation that any decision taken in terms of data preparation weight toward the quality of results one is able to obtain from automated text analysis. Our contribution should therefore be read as a call for a systematic application of validation procedures in any social science publication drawing on automated text analysis procedures.

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

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

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

In this regard, such "text-as-data" approaches squarely depend on a proper validation of applied techniques against some gold standard (Grimmer & Stewart, 2013).[1] Typically, applications of validation procedures using gold standards rely on some human inputs ("human coding") as a benchmark to systematically compare and evaluate against proposed supervised methods-based or dictionary-based classifications. Acting under the assumption that humans' understanding of text outperforms that of machines and that, if trained correctly, humans will make most correct and valid classifications of texts, human coded data is treated as a gold standard against which the performance of the computer is judged. However, "the quantities we seek to estimate from text [. . .] are fundamentally unobservable" (Lowe & Benoit, 2013, p. 299), and human judgment is, in fact, no exception to this general rule, as we know from the methodological content analysis literature. The consequences of, for example, human biases, predispositions and situational disturbances resulting in differing levels of "reliability" in human judgment in evaluating texts are well documented in traditional content-analytic applications (e.g., Ennser-Jedenastik & Meyer, 2018; Hayes & Krippendorff, 2007; Krippendorff, 2004;

[1] Here, we use the term "gold standard" and "ground truth" largely interchangeably, denoting some forms of objective data that serve as the reference.

Lombard, Snyder-Duch, & Bracken, 2002). Hence, the gold standard might not be as shiny as oftentimes assumed. By conditioning the relative performances of a given automated method against human inputs, we argue, the imperfect human coding as a "gold standard" greatly percolates a potentially (already) imperfect automated procedures. Yet, the implications of using such imperfect human judgments as the benchmark for evaluating the validity of the results of automated procedures — and especially the consequences of different levels of reliability in the gold standard manually materials are, until today, still insufficiently addressed.

Against this backdrop, we first presents a systematic review of articles published in the major social science journals in the past 20 years, showing that the actual practices of systematic validation are far from what is being acknowledged as a standard in the literature, and that the reporting and interpretation of validation procedures also differ greatly across studies. More importantly, in a second step, we argue that using imperfect human judgments as the benchmark effectively "tolerates" any mistakes or classification errors of automated procedures to a degree comparable to imperfect human judgments, which may have systematic consequences for the evaluation of the proposed automatic procedures. We assess this previously unexplored connection by relying on large-scale, systematic Monte-Carlo simulations. To our knowledge, our contribution is among the very first to provide a thorough, systematic evidence pertaining to a well-known, yet extremely scarcely discussed topic in automated content analysis. While our contribution should be read as a call for a systematic application of validation procedures, this study also serves to benchmark the combination of reliability in gold standard/ground truth and validity scores, warns against improper use of both in demonstrating the validity of the automated approach.

Logics and Procedures of Automated Content Analysis

We define "automated content analysis" (or automated text analysis) as the collection of content-analytic approaches that utilize automated methods of coding a large amount of unstructured textual data, in a way that the coding process itself (i.e., the text classification) is not performed manually but rather done by predefined computational algorithms (Grimmer & Stewart, 2013; Trilling & Jonkman, 2018). While the usage of the term "automated content analysis" in general in the extant literature encompasses a wide variety of forms (e.g., Grimmer

& Stewart, 2013; Hopkins & King, 2010; Krippendorff, 2013; Riffe, Lacy, & Fico, 2014), here we mainly refer to supervised learning methods and a simple dictionary approach. We choose to do so since many of the automated content analysis applications in the social sciences in general still heavily rely on those two broad approaches. Given that such methods are "designed to automate the hand coding of documents. . . [in a way that] it will directly replicate the hand coding" (Grimmer & Stewart, 2013, p. 13), this can be an attractive option for overcoming the limitations of manual approaches.[2]

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

For *supervised methods*, specific coding rules in manual annotations (as an input to an algorithm) are, in general, rarely explicitly articulated. Yet the algorithm takes such implicit

[2] Whereas dictionary-based or supervised learning methods assume already well-defined categories, the unsupervised methods such as LDA topic modeling (DiMaggio, Nag, & Blei, 2013; Maier et al., 2018) generally aim to inductively "discover" new classifications without any human input. Since our primary interest lies in the interplay of human input and machine output, we do not consider unsupervised methods such as scaling or topic modeling here. Also, our definition (inevitably) exclude automatic approaches of merely *acquiring* data, or traditional manual content analysis but make use of automated procedures in tasks other than the actual coding process (such as in data entry or data management: e.g., Lewis, Zamith, & Hermida, 2013).

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

The Use of Human Coding in Cross-Validation of Automated Content Analysis

Just as the term "automated content analysis" itself, the notion of "validation" often signifies many different practices, although most of them usually involve some notions of triangulation (e.g., Neunhoeffer & Sternberg, 2018). Yet for automated content analysis methods in particular, the discussion of "validation" has traditionally evolved around the general logic of establishing a correlative or a convergent validity of a measurement (Krippendorff, 2008). This is because any automated content analysis application – at least the ones we are focusing on in this contribution – essentially can be regarded as a classification problem (i.e., a measurement of the predefined categories in data). In this respect, the most straightforward method of validating a given measurement is to compare with another measurement of the same concept. Therefore, applications of validation procedures in automated content analysis have traditionally relied on some sort of human inputs ("human coding") as a benchmark as discussed above.

In dictionary-based approaches, the actual coding process itself – using the existing dictionary – does not involve any human inputs. Instead, the use of manual coding in dictionary approaches may involve additional comparisons of the derived results against manual coding, often made *post-hoc* for validating the results. A recent work by Rooduijn and Pauwels (2011)

on the (automated) measurement of "populism" in election manifestos well exemplifies the use of human coded data in validating the results of automated approaches. Using a traditional manual content analysis, they show that dictionary-based automatic coding of "populism" categories in party manifestos produces essentially very similar results compared to manual coding. Similarly, Young and Soroka (2012) compare the manually coded newspaper content against the results based on Lexicoder Sentiment Dictionary (LSD), and found that results using LSD and manual content analysis are largely comparable to each other.[3]

Nevertheless, it is exceptionally rare to see a validation of results *after* such classification tasks (yet for notable exceptions, see González-Bailón & Paltoglou, 2015; Muddiman et al., 2018; Young & Soroka, 2012). Once the dictionary is created, one typically assumes the classification membership that has to be estimated by an algorithm to be a simple additive function of the given dictionary elements (i.e., an aggregation of positive and negative dictionary scores results in the general sentiment of the article, etc.), typically ignoring any residual textual features that are not captured by the dictionary. Such blind trust in the dictionary grossly misses the very likely possibility that a given text (i.e., a newspaper article or a social network post) is more than a simple sum of its parts. While it appears that some post-hoc validations could ensure the soundness of conclusions drawn from dictionary approaches, however, to our knowledge little to no validation has been done in substantive studies that utilize dictionaries for classification purposes.

In supervised methods, the role of human input is more central in fine-tuning the algorithm (i.e., an "inner" cross-validation). As exemplified in Scharkow (2013), in supervised methods a researcher typically produces manually annotated sample materials, or a "training set," and the algorithmic classifier is later trained on sample material in order to develop certain statistical models (that effectively aim to reproduce implicit coding rules of human coders) to predict and classify unseen materials, the "test set." As such, human inputs are more directly

[3] However, a great deal of labor-intensive human inputs is usually required when building and constructing a well-defined dictionary (Muddiman, McGregor, & Stroud, 2018; Young & Soroka, 2012). Due to its labor-intensive nature, recent applications in this area increasingly turn to "crowdcoding", where the manual labor of highly trained yet few numbers of human coders are replaced with a large number of (little or untrained) crowd-coding workers (Haselmayer & Jenny, 2017; Lind, Gruber, & Boomgaarden, 2017).

utilized in such "learning" procedures in supervised methods, as opposed to being utilized post-hoc as in dictionary approaches. For instance, Burscher et al. (2014) have developed a supervised machine-learning algorithm to automatically classify certain generic media frames in news articles, based on a random subset of data that were manually coded by trained coders.[4] Yet due to inherent resource constraints associated with human coding, such validations typically rely on only a small subset of held-out samples to provide annotations that establish the ground truth. As such, although in principle a researcher can further validate their final outcomes by relying on an additional post-hoc validation method as described in dictionary-based approaches (i.e., an "outer" validation), this is rarely offered in practice.

The Myth of Perfect Standard in Human Coding?

Regardless of its specific orientations briefly reviewed above, the use of human coding as a gold standard is often regarded as *the* principal method of ensuring the validity and soundness of conclusions derived from the proposed automated procedures (e.g., DiMaggio, 2015; Grimmer & Stewart, 2013). The purpose of such cross-validation against a gold standard is, as Krippendorff (2008) notes, "to confer validity on the otherwise uncertain research results" (p. 6). Admittedly, this logic requires that the chosen benchmark (i.e., manual annotations by human coders) to be of "objective" and "unquestionable" truth — which, anecdotally, is more common among natural language processing and sentiment analysis literature (DiMaggio, 2015). Yet, as much of the traditional manual content-analytic literature suggests (e.g., Ennser-Jedenastik & Meyer, 2018; Hayes & Krippendorff, 2007; Krippendorff, 2004; Lombard et al., 2002), manual coding more often easily produces unreliable judgments as well, and especially so when the judgment in hand requires a nontrivial degree of inferences and subjectivity to classify a latent information (Krippendorff, 2013; Riffe et al., 2014).

The issue of (inter-coder) reliability is one of the most essential concerns in the extant literature concerning the quality of manual content analysis. At least for many recent manual content analysis applications, there has been relatively little disagreement regarding the

[4] In unsupervised methods, the use of human-coding as a benchmark is a common practice as well. For instance, Lowe and Benoit (2013) directly compare direct human-coding based party position scaling (which serves as a benchmark) against unsupervised scaling methods; they found that the proposed parametric scaling methods can produce largely similar results on par with human judgments.

consequences of sub-optimal reliability for key measurements (Krippendorff, 2004, 2013), and the need for better reliability for that matter. Yet to date, little attention has been devoted to the question of how the "quality" of manual coding affects the validity and conclusions derived from automated procedures. This is especially surprising, since "whenever ... principles by which humans generate ratings are heterogeneous across raters" (DiMaggio, 2015, p. 4), any automated procedures that are systematically evaluated upon such imperfect human input will be biased to the degree that can be found in such biased human judgments.

The use of (potentially) imperfect human coding as an ultimate benchmark against automatic techniques may have at least two systematic consequences. First, while automated methods themselves are perfectly reliable, imperfect human judgments (on validation materials) essentially makes the ultimate "target" of such perfectly reliable measurements radically deviate from the "true" target of inference, making them "reliably wrong" on-target. Second, and somewhat relatedly, systematically flawed human judgment can bias the performance of learning algorithms, leading to biased conclusions against a true standard. Therefore, using imperfect human coded data as an ultimate benchmark makes it harder to evaluate the relative trustworthiness of such validation procedures. However, most of the studies do not provide any validation at all (please see our later empirical section). Even among studies which *do* compare automated procedures with manual coding as a validation tool, they often misreport or misspecify the validation metrics and hardly ever report the quality of such manual coding itself but essentially treat such imperfect human judgment as *the* unquestionable truth (e.g., González-Bailón & Paltoglou, 2015; Lowe & Benoit, 2013; Young & Soroka, 2012). Indeed, we know only one existing study that suggests some tentative evidence of the relationship between quality of human coding and (machine-based) classification accuracy (Burscher et al., 2014).[5]

In sum, there appears to be a sufficient reason to suspect a systematic relationship between the quality of human coding and the relative bias and errors regarding the ultimate

[5] In a recent study, González-Bailón and Paltoglou (2015) compare five available sentiment dictionaries against human annotations, yet they do not directly deal with the implications of imperfect reliability in human coding. Scharkow and Bachl (2017), the only one of existing studies that examines the consequences of imperfect reliability in human coding, mainly deal with its implication on "linkage analysis", but not on automated content analysis.

conclusions from the automated content analysis – especially when using such imperfect human coding as the "gold standard". While many of prior contributions on this topic – both theoretically and empirically – stress a need for a proper validation of applied techniques (e.g., González-Bailón & Paltoglou, 2015; Grimmer & Stewart, 2013; Hopkins & King, 2010), we do not know much about how the field in general stands in terms of standard validation practices and the use of imperfect human coding in particular. Therefore, we first conduct a systematic review of relevant articles published in the top social science journals in the past 20 years, focusing on the use of manual coding (by trained human coders), if any, on standard validation procedures against the proposed automated methods, and on the assessment of performances of such automated methods (Study 1). While Study 1 should provide us an overall picture of how typical valdiation is approached in automated content analysis, at the same time this further serves to benchmark our simulation modeling strategies (Study 2), where we warn against *improper* use of manually annotated "gold standard" materials in demonstrating the validity of the automated approaches.

Study 1: A Content Analysis Study

Sample and Procedures

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

A total of five highly qualified coders tested the initial coding scheme by independently coding 10 sample articles (approximately 14% of total sample) and collectively discussed any coding problems and disagreement. Coding instructions were iteratively revised until the coding schemes would produce reliable results. Intercoder reliability (based on Krippendorff's alpha) above 0.75 was ensured for each of the variables coded.

Results

The results of the systematic literature review are presented in Table 1 below. Among 73 articles being identified as relevant, a total of 55 studies used dictionary approaches while 18 used supervised machine learning methods.[6] Yet only about half of the papers using a dictionary approach to automated content analysis referred to some sort of manual gold standard in their text, typically relied on 4.24 coders with a median of 232 manual materials per coder, yielding a little less than 1000 manual annotations for validation materials. Among them, even less – only 16 percent – actually reported any measures of inter-coder reliability in such materials – only 2 studies ever reported Krippendorff's alpha values, and 3 studies report either percentage agreement or Holsti agreement measure (the rest of studies reported other measures such as Scott's Pi or correlation coefficients). Notably, shares are much higher when coming to papers using supervised machine learning. The most common measure of inter-coder reliability reported were percentage agreement (in seven cases, 38%) and Krippendorff's alpha (in six cases, 33%). In sum, among 73 articles we have examined, only 9 studies (which is about 20% of studies that ever refer to some human-coded gold standards) have used Krippendorff's alpha (average alpha = 0.68).

When it comes to the papers referring to procedures of validation of their automated approaches, results were very similar. However, only around 40 percent of papers using a dictionary approach also reported some validation measures, compared to 67 percent of papers using supervised machine learning. The most commonly used measures of validity are indeed the widely accepted measures of precision (in 13 cases, $M = 0.74$) and recall (in 9 cases, $M =$

[6] 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 the social sciences in general.

0.60), yielding an average of F1 score of 0.6311 (in 10 cases, based on either directly reported F1 scores or indirectly calculated F1 scores from reported precision and recall). However, we also observe that other metrics were widely used as well in reporting validation. This is somewhat disconcerting insofar as these tend to be indeed either the intercoder reliability measures (e.g., Holsti, Cohen's Kappa, Krippendorff's alpha) or correlation coefficients, which is *not* designed for the validation of a given automated procedures. Interestingly, there were only three studies which reported *both* Krippendorff's alpha in human-coded materials *and* proper validation metrics. Generally, the result of Study 1 reveals that reported measures of validations and the quality of human-coded data used in such validations are far from what is being acknowledged as the best-practice in extant literature.

Table 1
Results of the systematic literature review

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

Note: Percentages refer to share of articles using that method.

Study 2: A Monte-Carlo Simulation Study

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

While we based our procedures in a way that largely mirrors the typical approaches in this

area, dictionary approaches and supervised learning approaches considerably differ in how they utilize manual coding, as well as their data requirements and the logic underlying each of the techniques. Therefore, we have developed two separate approaches in simulating their behaviors. Yet in all of the cases, we have broken down our approach into three stages – data generation, human coding, and automatic classification – where we systematically varied the intercoder reliability of the "gold standard" material, along with a number of related factors such as the number of coders and the number of manual annotations per coder (see below section). Then we systematically compared different scenarios in terms of their classification accuracy and F1 scores (i.e., precision and recall) based on the "true" standard (i.e., a quantity of interest that is typically unknown to researchers) in order to illustrate how different practices of human coding in automated content analyses affect the overall results and the relative trustworthiness of conclusion drawn from such results.[7]

Data Generation Stage

We create data (e.g., textual data to be analyzed by a researcher) with the "true" outcome variable of interest; the goal of any quantitative text analysis method is to somehow approximate the true value *y*, either by human coding, machine algorithms, or some combinations of both. For the data generating process, we set the true value of *y* to be stochastically generated from three hypothetical independent variables (x_1, x_2, and x_3), the values of which were sampled either from a multivariate normal distribution (for supervised learning) or from a categorical distribution (for dictionary approach – see below). We assume the size of text data is sufficiently large to warrant an automated approach. As such, the data in question (hypothetically) covers 10 news articles per day per each of 10 media outlets, spanning a total of 20 years. Accordingly, each single simulation run is set to generate 730,000 observations (10 x 10 x 365 x 20).

Supervised ML Scenario. For supervised learning scenarios, we assume values of independent variables are sampled from a multivariate normal distribution, with a randomly generated variance-covariance matrix $\sum$ for each simulation run. This randomly generated variance-covariance matrix ensures that idiosyncratic values of the covariance matrix do not

[7] A set of replication code for this manuscript can be found at: [redacted for a review].

skew the overall results of the simulation. The *true* values of y (which is the binary variable) are then sampled from a Binomial distribution, with the probability parameter having a very simple linear functional form as follow:

$$\begin{aligned} y &\sim Bernoulli(\pi) \\ \pi &= logistic(\mu) \\ \mu &= \boldsymbol{X}\boldsymbol{\beta} + \epsilon \end{aligned} \tag{1}$$

with ϵ being Gaussian noise added to ensure that each run of the simulation is not completely deterministic. The $\boldsymbol{\beta}$, true population parameters, were fixed throughout the simulation runs (specifically, β_1 = 0.5, β_2 = 0.2, and β_3 = 0.6, which were randomly chosen).

Bag-of-Words Scenario. For a dictionary (bag-of-words) method, we assume a very similar approach to data generation as discussed above, however we assume values of independent variables were sampled from a Categorical distribution with two separate Dirichlet priors for negative and positive draws (i.e., a discrete value range from -5 to 5), where they represent some "features" of a given textual data (e.g., a word or N-grams). These values are later to be matched with a "dictionary" (of similar discrete values) in order to determine the estimated values of y for each observation. Since this requires discrete rather than continuous values, we use a slightly different setup as follow:

$$\begin{aligned} \pi_{neg} &\sim Dirichlet\big(N, \alpha_{neg}\big) \\ \pi_{pos} &\sim Dirichlet\big(N, \alpha_{pos}\big) \\ \boldsymbol{X}^{T}_{k \in K} &\sim Categorical\big(N, \big[\pi_{neg} \text{ or } \pi_{pos}\big]\big) \end{aligned} \tag{2}$$

with N being the total number of observations, and α_{neg} and α_{pos} being hyper-parameters governing the shapes of the categorical distribution, and $\boldsymbol{X}_k$ being a set of independent variables (with K number of textual features). The two negative and positive Dirichlet priors π_{neg} and π_{pos} were randomly selected for each column of independent variable, effectively treating such independent variables as systematic, recurring features of given textual data. Given the set of textual features, y is generated in a similar fashion as in Equations (1). Yet for the bag-of-words approach, the vector $\boldsymbol{\beta}$ was extended to K = 5 and their $\boldsymbol{\beta}$ values were fixed to 0.2. This enable us to better approximate the multidimensionality of textual data, while treating y effectively a function of the simple sum of chosen textual features.

Human Coding Stage

In a typical content analysis, at least two or more trained human coders are assigned to a small set of sample documents, and independently code such documents. This process is repeated until the satisfactory level of intercoder reliability is achieved (typically Krippendorff's alpha equal or greater than 0.7). Once a coder interchangeability is ensured by such coder training, validation materials are often divided evenly and annotated only by a single coder (Grimmer, King, & Superti, 2018). While the common practice of evenly (but not randomly) dividing the validation materials across coders effectively means that any potential coder idiosyncrasies are just ignored, such idiosyncrasies are not likely to be randomly distributed across manual annotations with just a handful number of manual annotations. Therefore, it is indeed expected that the likelihood of coder idiosyncrasies being effectively cancel out each other (therefore improving the quality of the "gold standard") would gradually increases as the function of the number of total coders and the number of independent annotations per coders.

Following this logic, we specified following factors that may affect the quality of the "gold standard" and therefore evaluations of overall performance of automated algorithms. Those factors include: the number of human coders (k = 2, 4, 7, 10), the predetermined target values of intercoder reliability (Krippendorff's alpha = 0.5, 0.6, 0.7, 0.8, 0.9), as well as the number of annotation per coder (n = 50, 100, 250, 500). These numbers were chosen to reflect typical procedures and their common variations, as can be seen in our Study 1 above.

In all scenarios, human coders classify a given observation as "1" (e.g., a news article contains the quantity of interest, such as a certain actor, frame, or tonality) or "0" (e.g., does not contain this quantity of interest). This human coding ($\hat{y}$) can be, in principle, either correct or incorrect against the (unknown) "true" value of y, therefore behaviors of human coders were modeled by a Binomial distribution with varying probability of successfully categorizing the true data. The heterogeneity in each of the coders' coding behaviors (e.g., expertise or bias) was also modelled by a beta distribution with varying shape (using predefined hyperparameters for each target reliability level), which effectively enables us to simulate a situation where a certain number of coders produce more "correct" judgment ($\hat{y} = y$) whereas other coders produce more "false" judgment ($\hat{y} \neq y$). Yet the overall human annotation patterns at the chosen level of a beta

distribution parameters were ensured to produce an acceptable level of the target reliability coefficient (e.g., Krippendorff's alpha = 0.7). Depending on the specific application (dictionary-based vs. supervised learning), the "hand-coded data" from this stage is then used as either for post-hoc validation or as an input for the later automated classification.

Algorithm-based Classification Stage

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

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

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

Simulation Results

As an initial model check, we first investigate the overall classification accuracy (against true value) as a function of target reliability in validation materials, predicting how well the proposed algorithms correctly recover the quantity of interest in data (such as a certain actor,

frame, or tonality) at varying levels of reliability in validation materials. In Figure 1, different levels of target Krippendorff's alpha is stacked over y-axis, and in x-axis the overall prediction accuracy against true value (defined as the proportion of correct predictions, or the sum of true positive and true negative cases over entire cases) is reported; they are presented along with their 95% confidence intervals, for every combination of experimental factors in MC simulations.

As can be seen in Figure 1 above, within two ML scenarios, using "better" quality training material appears to directly increase true classification accuracy. Using the overall accuracy in our simulation as the reference point, the first and the second panels of Figure 1 makes it clear that the reliability level of the training material has a nontrivial benefit in improving the accuracy of predictions based on automated procedures. Indeed, this results is somewhat expected since the ML methods takes the human input as the basis for developing the classification algorithm, therefore overall accuracy of the final classifier is dependent upon human input and its quality thereof. In contrast, the last panel of Figure 1 shows that the typical Bag-of-words application does not benefit from improved (post-hoc) human coding. Yet again, this is a much expected pattern, since the quality of a given dictionary itself does not depend at all on *post-hoc* validation of human coding unless manual coding is directly utilized in developing and constructing dictionary itself. Overall, our results presented in Figure 1 show that our simulation setup can correctly reproduce a common pattern in extant studies (therefore ensures the validity of our analysis and conclusion).

Next, we examine indirect consequences of relying on imperfect human coded materials as a benchmark in evaluating the performances of automated procedures in terms of post-hoc validation. Typically, researchers rely on a small fraction of human-coded materials for validating their primary findings from automated procedures, deciding whether the overall accuracy or classification performance is good enough to proceed to further analyses. This decision has to be based on some some *a priori* chosen threshold value: if validation (based on human coding) is purportedly not satisfying enough to pass such a threshold level, additional steps are sought to improve the quality of automated procedures (e.g., re-training of algorithms, or changing the dictionary, etc).[8] The primary interest of such validation lies in extrapolating

[8] Here, we do not consider a scenario where a researcher decides to improve the quality of human

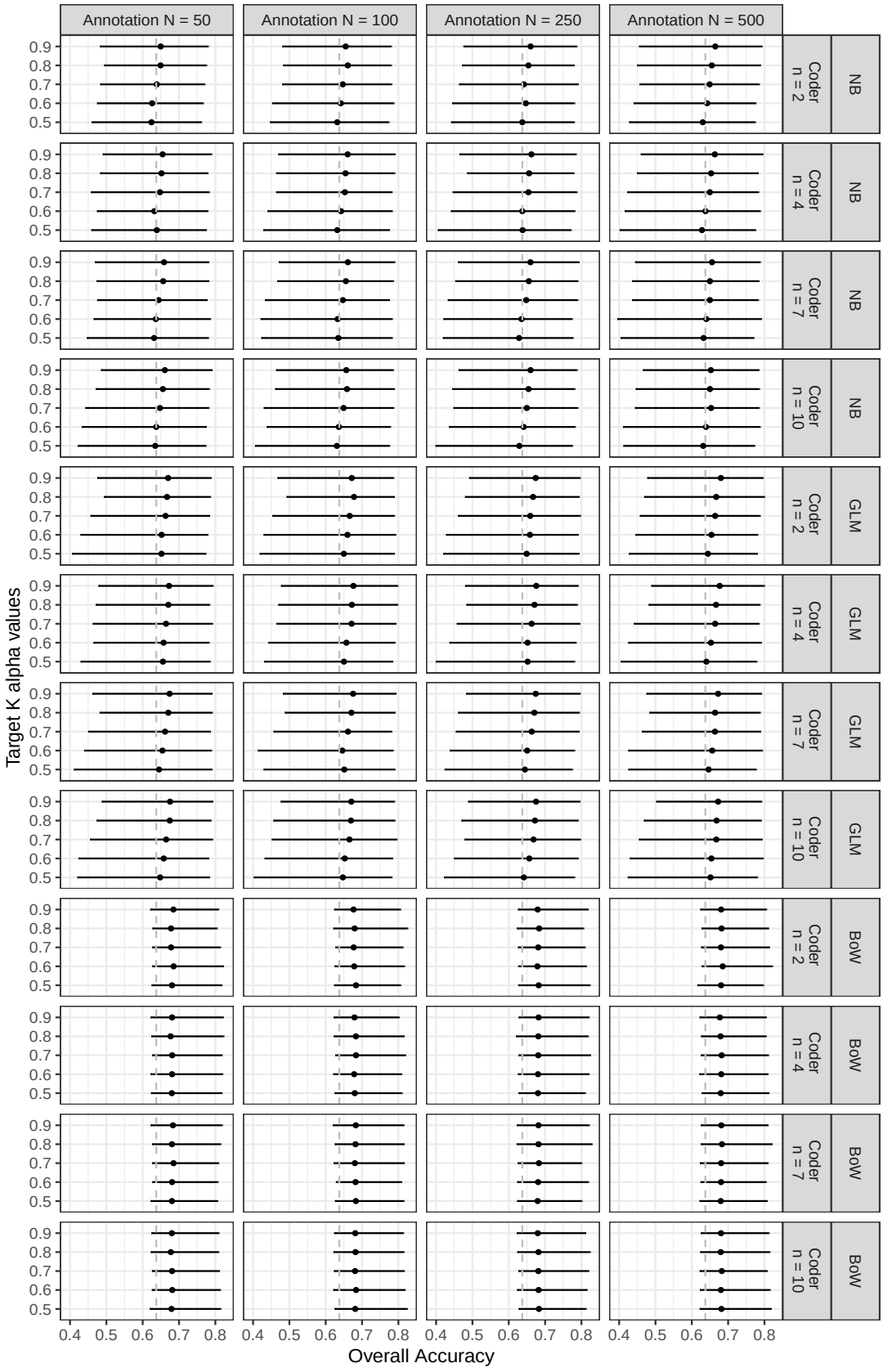

Figure 1. Overall classification accuracy against true value across conditions (reference line is overall mean).

the observed level of classification performance (based on validation materials) to the level of classification performance *that could have been observed* based on the entire range of data. In other words, the observed level of classification performance serves as a proxy, or an estimate for the true, unknown classification performance. Therefore, an interesting question here is how well the observed classification performance reflects the true classification performance under imperfect reliability, and how large or small a potential bias is. Relatedly, making decisions about the unknown, true values of the overall classification performance based on the observed performance from hand-coded validation materials, essentially, can be seen as classical decision error scenarios (i.e., type I and type II errors), as schematically presented in Table 2.

Table 2
Decision scenarios based on observed vs. true level of classification performance.

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

In order to illuminate the potential consequences of relying on imperfect data in making decisions about the true, unknown classification performances of the proposed automated procedure, we divide our entire simulations into four mutually exclusive categories, as in Table 1, based on the cross-tabulation of "observed" F1 scores (from validation materials) against true F1 scores (based on the true values of y). We first present the proportion of simulation cases which incorrectly conclude about the overall classification quality based on observed quality. Second, we further present the degree of relative "bias" within such incorrectly concluded cases — defined as the $F1_{validation}/F1_{true}$, where F1 score is a weighted average of the precision and recall — which captures the degree of under- or over-estimation of (true) overall F1 scores against observed F1 scores based on human validation materials. For this application, we choose the cutoff value of F1 score to be 0.6311, which is just the average F1 score reported in Study 1.

As can be seen in above Figures, it appears that utilizing more "high-quality" materials

---

coding. This is based on the consideration, as to our argument being advanced here, that a researcher (often incorrectly) assumes that human coding is perfectly reliable and valid.

Figure 2. Percentage of decision error and relative bias in F1 scores (over 1000 Simulations per each scenario), Naïve Bayes classifier.

Note: Upper panel = Proportion of cases (each simulation run) incorrectly conclude on classification performances. Lower panel = Relative bias in F1 scores among 1000 replications, with their median and 95% percentile confidence intervals.

Figure 3. Percentage of decision error and relative bias in F1 scores (over 1000 Simulations per each scenario), GLM classifier.

Note: Upper panel = Proportion of cases (each simulation run) incorrectly conclude on classification performances. Lower panel = Relative bias in F1 scores among 1000 replications, with their median and 95% percentile confidence intervals.

Figure 4. Percentage of decision error and relative bias in F1 scores (over 1000 Simulations per each scenario), Bag-of-words.

Note: Upper panel = Proportion of cases (each simulation run) incorrectly conclude on classification performances. Lower panel = Relative bias in F1 scores among 1000 replications, with their median and 95% percentile confidence intervals.

for post-hoc validation has obvious and discernible consequences in the evaluation of classification quality of the automated procedures. Among 1000 replications of each scenario, all of our experimental factors appear to decrease the decision error rates in using observed level of F1 scores to approximate the true F1 score level. The leftmost upper panel in Figure 2 shows that in worst-case scenarios in which using only two coders with a handful of materials, coupled with low reliability (K alpha of 0.5), such suboptimal hand-coded data result in approximately 30% of cases incorrectly over- or underestimate the true F1 scores. Under the same combination of a total number of coders and independent annotations per each coder, improving reliability at best marginally decreases the overall percentage of decision errors. Yet as either a total number of coders or a number of independent annotations per each coder start to increase, we see that the total proportion of cases that incorrectly estimate the true F1 scores start to decrease substantially (the upper panel of Figure 2). At the same time, the relative bias (as can be seen in the bottom panel of Figure 2) also starts to converge to true estimates, both in ML scenarios and dictionary approaches. While there appears to be no apparent overall main effect of higher reliability in terms of reducing the total number of decision errors in ML approaches, using higher reliability in manual validation materials tends to *reduce* the magnitude of relative bias (i.e., the degree under- or overestimation of the true F1 scores based on observed F1 score). This thus greatly reduces the uncertainties regarding the ultimate conclusions based on automated procedures. In bag-of-words applications as presented in Figure 4, we see a clearer impact of higher reliability in reducing potential decision error using observed F1 scores. It appears that using the observed classification quality derived from manually coded validation materials in bag-of-words applications tends to underestimate the true classification quality (lower panel of Figure 4). Yet, higher reliability in such validation material nevertheless reduces the uncertainties regarding the ultimate conclusions based on automated procedures. In sum, this provides a very consistent picture of the impact of "quality" of manually coded materials in validating the conclusions derived from automated procedures.

Discussion and Conclusion

The aims of the current investigation were twofold; first, we attempted to provide a systematic overview of current practices in validation procedures of automated content analysis

in the social sciences. Second, we aimed to provide further insights into how the various decisions taken by researchers on various steps of the research process influence the overall quality of arguments and conclusions that could be drawn from the analyses of such data. In order to achieve these goals, Study 1 relied on a systematic review of published papers from major social science journals that purportedly have relied on "automated text analysis." The results of the first study show that, while automated content analysis procedures are widely used throughout social science applications, there is still strikingly little consistency in *whether* and *how* validation is reported. Very often, studies do not report *any* validation metrics when relying on automated methods. Even when they do, metrics related to (human-coding based) validations are generally not consistently reported, and are often limited in providing actual methodological details. For the second study, we have designed a set of Monte Carlo simulation procedures, which closely mimic multiple scenarios of data coding, classification, and data validation for the most widely-used supervised learning-based applications as well as for the dictionary classification methods. The second study revealed that any decision taken during preparation of validation materials – the number of coders, the size of the training sets produced by each coder, and the "quality cut-off point" (e.g., Krippendorff's alpha for intercoder reliability) – indeed all produces systematic consequences for the evaluation of the proposed automated procedures.

Coupled together, the results from both studies give good reason for concern about the quality (or rather the *validity*) of conclusions drawn from automated content analyses in the social sciences. In order to make sensible conclusions from data, a proper validation of the proposed automated methods is much needed. A statistical model may have an excellent fit to the data and a chosen dictionary may very well recover some pre-defined categorizations, but these metrics themselves do not provide any sensible meaning of their own if the model suffers from over-fitting the training data, or the pre-defined dictionary does not actually conceptually "align" with the categories of interest. To be clear, the current study *does not* make a case for the argument that we should *exclusively* rely on human validations. Quite the contrary, one of the main points being advanced here is that humans are not perfect. However, this study *does* argue that *proper* validation is essential in general. While the (potentially imperfect) human "gold standard" is often the best we can get, the results of this study suggest that extra cautions should

be taken in such validation procedures. Often, imperfect judgment of human coding can percolate to machine coding to inferences if a researcher turns a blind eye to such imperfect quality of human judgment. We therefore advise researchers to pay closer attention to the issue of proper reliability in manually coded validation materials in order to reduce uncertainties in potential biases, while striving to increase the overall size of the validation materials as large as possible in order to reduce overall decision errors based on manual validation materials.

Few limitations should be noted. First, and concerning Study 1, our systematic literature review is limited to studies explicitly using our search terms in their title, abstract, or keyword list. Such sample, by definition, does not include all potentially relevant articles from the last 20 years. Yet still, it allows us to make a sensible selection of extant articles that do not only use these methods but also advertise that they do so. In fact, we believe articles that are not part of our sample to be of even lesser rigor in the use and reporting of "gold standards".

Second, and concerning Study 2, we only have considered a binary classification task in our simulation setup and a single dimension of validation metrics — specifically, recall, precision, and resulting F1 scores. While this greatly simplifies our main arguments and (still complex) simulation setups, there are indeed a nontrivial number of existing applications that go beyond such simple classifications, dealing with numerical forms of predictions (i.e., scaling). Ultimately our conclusion would be bounded to our specific setup, yet we reason that our core arguments can be equally applicable to more complex forms of automated content analysis as well, and potentially to any unsupervised methods for that matter. Given the additional complexity and difficulties involved in such complex, numerical forms of predictions (e.g., document scaling) for human coders, achieving acceptable intercoder reliability in such applications likely to be even more difficult compared to that of simple binary ones. Hence we suspect the potential problems of "imperfect quality" in human-coded validation materials should be greater, if not identical, on those applications.

Designing a simulation-based study gives us a unique opportunity to see the many potential *what if* scenarios of the research processes. Such an approach, if carefully designed, allows to robustly explore potentially important variability in research processes in a probabilistic manner without having to actually spend resources on carrying out those scenarios.

Also, one of the core advantages of relying on such a simulation-based approach is that a researcher has the ability of "knowing the truth," specifically the true value of the quantity of interest, y, as well as their correct functional forms that produce each observation (of media data) in this case. This also provides a means for formally checking the sensitivity of one's findings, or enables one to provide a proper context of one's findings by exploring counterfactual scenarios. Such clairvoyance, however, comes at a cost — the degree of abstraction and simplification. This simplification is done not only for computational, but also for conceptual reasons, however. As Scharkow and Bachl (2017) note, "the challenge is to specify a simulation that is simple enough to be comprehensible, yet realistically models the underlying process of interest" (p. 330). In this regard, we have relied on somewhat idiosyncratic and simplified approaches in our simulations. While we surely acknowledge that our setup could have been constructed in a more realistic and complex way, our ability to accurately simulate human behaviour does not necessarily depends on very complex models.

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

References

Aaldering, L., & Vliegenthart, R. (2016). Political leaders and the media: Can we measure political leadership images in newspapers using computer-assisted content analysis? *Quality & Quantity*, *50*(5), 1871–1905.

Boomgaarden, H. G., & Vliegenthart, R. (2009). How news content influences anti-immigration attitudes: Germany, 1993–2005. *European Journal of Political Research*, *48*(4), 516–542.

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*(1), 8–23.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190–206.

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*(1), 122–131.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, *2*(2), 2053951715602908.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, *41*(6), 570–606.

Ennser-Jedenastik, L., & Meyer, T. M. (2018). The impact of party cues on manual coding of political texts. *Political Science Research and Methods*, *6*(3), 625–633.

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*(1), 95–107.

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

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, *51*(6), 2623–2646.

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

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.

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

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

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

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*(1), 34–52.

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures*, *11*(3), 191–209.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604.

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*(3), 298–313.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., … Häussler, T., et al. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93–118.

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*, 1–13.

Neunhoeffer, M., & Sternberg, S. (2018). How cross-validation can go wrong and what to do about it. *Political Analysis*, 1–28. Retrieved from http://www.marcel-neunhoeffer.com/pdf/papers/pa_cross-validation.pdf

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

Rooduijn, M., & Pauwels, T. (2011). Measuring populism: Comparing two methods of content analysis. *West European Politics*, *34*(6), 1272–1283.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using german online news. *Quality & Quantity*, *47*(2), 761–773.

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*(3), 323–343.

Trilling, D., & Jonkman, J. G. (2018). Scaling up content analysis. *Communication Methods and Measures*, *12*(2-3), 158–174.

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