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, and Hajo G. Boomgaarden

Department of Communication, University of Vienna, Austria

Draft date: November 22, 2018

Draft in progress. Please do not cite without permission.

Author Note

Please direct any questions and inquiries concerning this manuscript to

hyunjin.song@univie.ac.at

Abstract

TBA

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 and have to be analyzed with limited resources, 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. The most common practice here is to rely on precision (share of relevant items in all selected items), recall (share of selected items in all relevant items), and the resulting F1 score (ratio of precision and recall) of the automated procedure, using human coded material as a benchmark. Acting under the assumption that humans' understanding of text outperforms that of machines and that, if trained correctly, humans will make mostly correct and valid classifications of texts, human coded data is treated as a gold standard against which the performance of the computer is judged. Consequently, the validation of computer-assisted coding based on human coded gold standards is grounded in the assumption that human readers' placements or evaluations of a given text are indeed as close as we can get to an error-free, faultless and faithful representation of the quantities being estimated.

---

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

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; 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 (imperfect "gold standard") greatly percolates a potentially already imperfect estimation of the automated procedures. Yet, the implications of using such imperfect human judgments as the benchmark for evaluating the validity of the results of automated text analysis – and especially the consequences of different levels of reliability in the gold standard are, until today, still insufficiently addressed.

Against this backdrop, we first posit that a systematic validation of automated text analysis approaches is still rare in the social science literature, in general, and in communication research, in particular. Substantiating this argument, this study first presents a systematic review of articles published in the top social science journals in the past 20 years, showing that standards of validation are far from being acknowledged in the literature and that the reporting and interpretation of validation procedures differ greatly. More importantly, in a second step, we argue that by conditioning the performances of an automated method against human inputs, such validation procedures essentially mirror imperfect human performances. Therefore, using imperfect human judgments as the benchmark for automated text analysis validation effectively "tolerates" any mistakes or classification errors of automated procedures to a degree comparable to imperfect human judgments at researcher's chosen degree of reliability levels. Thus, the choice of using human coding – and in particular the "quality" of such benchmarks – as a ground truth or gold standard may have systematic consequences for the evaluation of the proposed automatic procedures.

For this second step, we assess this previously unexplored connection between the

reliability of human judgment and relative performance of automated procedures (against true standards) by relying on large-scale, systematic Monte-Carlo simulations. Specifically, we illustrate how different levels of intercoder reliability scores in benchmark materials affect the assessment of overall (machine) accuracy/recall combinations (against such imperfect standards) and the ultimate conclusions drawn from such results. 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. Our contribution should be read as a call for a systematic application of validation procedures in any social science publication drawing on automated text analysis procedures. At the same time, the study serves to benchmark the combination of reliability in gold standard/ground truth and validity scores and warns against improper use of both to demonstrate the validity of the 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., see: Grimmer & Stewart, 2013; Hopkins & King, 2010; Krippendorff, 2013; Riff, Lacy, & Fico, 2014), here we mainly supervised machine learning methods and a simple dictionary approach, wherein computational algorithms require a nontrivial degree of human decisions and inputs before or during the coding process itself (Trilling & Jonkman, 2018). We choose to do so based on the consideration that many of the automatic content analysis applications in social science in general and in communication science in particular still heavily rely on those two broad approaches, although it is not uncommon to find the use of more sophisticated, yet highly flexible applications that have a slightly different focus (such as LDA topic models: DiMaggio, Nag, & Blei, 2013; Maier et al., 2018). This study does not include unsupervised methods, since our primary interest are the interplay of human input and machine output. Also, our definition (inevitably) exclude automatic approaches of *acquiring* textual data, or (more

traditional) manual content analysis but merely make use of computers in tasks other than the actual coding process (such as in data entry or data management: e.g., Lewis, Zamith, & Hermida, 2013), although such aspects are, in practice, tightly intertwined with automated methods as we define here.

The prevalence of simple dictionary- or supervised learning-based methods is partly at least due to the fact that such forms of automated analyses are tend to be "designed to automate the hand coding of documents. . . [in a way that] it will directly replicate the hand coding" (Grimmer & Stewart, 2013, p. 13). At the presence of a large volume of data, this can be an attractive option for overcomming the inherent time and resources constraints in manual approaches. In addition, in practice, such methods are relatively easy to implement in combination with manual approaches, which are still the dominant workhorse for many communication scholars.

Although a specific method and application of automatic 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 the predefined categories (i.e., a discovery and measurement).[2] A simple dictionary approach generally relies on the pre-defined "dictionary" of words or phrases by a researcher – ranging from simple keywords lists to complex Boolean expression, 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 of the absence of a category) or a net tonality (positive vs. negative) of an article

---

[2] In this respect, unsupervised learning methods are most clearly distinguishable from the other two: while unsupervised learning methods generally aim to inductively "discover" categories without human inputs, dictionary-based or supervised learning methods assume (already) well-defined categories. For a detailed discussion about the differences across methods, see Grimmer and Stewart (2013) or Boumans and Trilling (2016).

(e.g., Aaldering & Vliegenthart, 2016; Boomgaarden & Vliegenthart, 2009; González-Bailón & Paltoglou, 2015; Rooduijn & Pauwels, 2011; Young & Soroka, 2012).

For supervised methods, specific coding rules in manual annotations (as an input to an algorithm) are, in general, rarely explicitly articulated. Yet the algorithm takes such implicit judgments derived from manual coding as the point of reference, and tries to infer the features of data that best classify the text into different predefined categories. Currently, there exists a variety of supervised learning algorithms available – ranging from a simple regression framework or Naive Bayes to more sofisticated methods such as Support Vector Machines (SVM) or random decision forest (for an overview of different algorithms commonly used in social science applications, 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 a 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; van Atteveldt, Kleinnijenhuis, Ruigrok, & Schlobach, 2008).

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

Just as the term "automated content analysis" itself, the notion of "cross-validation" often signifies many different practices within social sciences, although most of them usually involve some notions of triangulation. According to Neunhoeffer and Sternberg (2018), there are at least four different meanings of cross-validation: (a) validating new measures or instruments (i.e., traditional measurement validity), (b) obtaining an "estimate" of *true* error, (c) a model tuning (i.e., hyperparameter tuning), and (d) a robustness check using K-fold validations (as a statistical tool). Yet for automated content analysis methods in particular, the discussion of "validation" has traditionally evolved around the first meaning of the term, following 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 measurement instrument (or rather the results thereof) is to compare one measurement 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. Since simple dictionary-based or supervised method-based automated coding essentially aims to produce human-equivalent judgment, there are indeed many steps within which a researcher can compare the manual coding of well-trained human coders and that of automated algorithms during a research process.

In dictionary-based approaches, a great deal of labor-intensive human inputs are required when building and constructing a well-defined dictionary (Muddiman, McGregor, & Stroud, 2018; Young & Soroka, 2012).[3] Yet once created, the actual coding process itself using the 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, in a sense that such comparisons are often made *post-hoc* for validating the results. Rooduijn and Pauwels's (2011) research on the (automatic) 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 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.

Nevertheless, most often such dictionaries are validated during the initial dictionary construction processes, and it is exceptionally rare to see a validation of results *after* such classification tasks (yet for notable exceptions, see González-Bailón & Paltoglou, 2015;

---

[3] Due to its labor-intensive nature of building a valid and reliable dictionary, recent applications in this topic are increasingly turn to "crowdcoding", where the manual labor of highly trained yet few number of human coders are replaced with a large number of (untrained) crowd-coding workers (e.g., Haselmayer & Jenny, 2017; Lind, Gruber, & Boomgaarden, 2017), serving as a "wisdom-of-the-crowd" approach. While a detailed discussion of such issue is beyond the scope of this article, interested readers may instead consult Lind et al. (2017).

Muddiman et al., 2018; Young & Soroka, 2012). Once the dictionary is created and tested, one typically assumes the classification membership that has to be estimated by an algorithm to be a simple additive — and only exclusive — function of the given dictionary elements (i.e., 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 a given dictionary features. Such blind trust in the dictionary grossly misses the very likely possibility that a given text in question (i.e., a newspaper article or a social network post) is more than a simple sum of its parts. Admittedly, the researchers frequently make such an assumption the minute before they decided to use a dictionary, however to our knowledge little to no validation has been done in substantive studies that utilizing dictionaries for classification purposes to check how well does a given dictionary approximate the objective reality of text data.

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 "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 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 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.

_____

[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, found that the proposed parametric scaling methods can produce largely similar results on par with human judgments.

The Myth of Perfect Standard in Human Coding?

Regardless of its specific orientations briefly reviewed above, in extant literature "human reasoning is routinely described as a 'gold standard' against which algorithmic output should be judged" (DiMaggio, 2015, p. 2). Indeed, 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., Grimmer & Stewart, 2013; Lowe & Benoit, 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., manually produced annotations by human coders) to be of "objective" and "unquestionable" truth — which, we anecdotally observe, is more common among natural language processing and sentiment analysis literature (e.g., DiMaggio, 2015). Yet as much of the traditional manual content-analytic literature suggests (Ennser-Jedenastik & Meyer, 2018; Hayes & Krippendorff, 2007; Krippendorff, 2004; Lombard et al., 2002), manual coding more often than not easily produces unreliable judgments as well, and especially more so when the judgment in hand requires a nontrivial degree of inferences and subjectivity to classify a latent information in a given text (Krippendorff, 2013; Riff 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. If the reliability estimates of key constructs are not that compelling, as Krippendorff (2013) notes, "doubts as to what these data mean prevail, and their analysis is hard to justify" (p. 268). As such, there has been relatively little disagreement regarding the consequences of sub-optimal reliability for key measurements (Krippendorff, 2004, 2013), and the need for better reliability for that matter, at least for many recent manual content analysis applications. 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 – as DiMaggio (2015) notes – "whenever ... principles by which humans generate ratings are heterogeneous across raters" (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 which

automatic techniques are validated upon may have therefore at least two systematic consequences. First, while automated methods themselves are perfectly reliable in producing the same results whenever repeated, 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), or 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* true, gold-standard materials (e.g., González-Bailón & Paltoglou, 2015; Lowe & Benoit, 2013; Young & Soroka, 2012). Indeed, we know only one existing study (Burscher et al., 2014), which suggests some tentative evidence between quality of human coding and (machine-based) classification accuracy.[5]

In sum, there appears to be a convincing reason to suspect a systematic relationship between the quality of human coding and the relative bias and errors regarding the ultimate 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 – stresses a need for a proper validation of applied techniques against some gold-standard (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

---

[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 examines the consequences of imperfect reliability in human coding, mainly deal with its implication on "linkage analysis," but not on automated content analysis.

science journals in the past 20 years, focusing on the use of manual coding (by trained human coders), if any, in standard validation procedures against the proposed automated methods, and the assessment of performances of such automated methods (Study 1). While this systematic review should provide us a overall picture of typical validation procedures in any publication drawing on automated text analysis procedures in extant literature, study 1 at the same time further serves to benchmark in modeling the combination of reliability in gold standard/ground truth and validity scores against the true standard (Study 2), where we warn against improper use of both to demonstrate the validity of the approach.

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 abstracts and corresponding keywords using the following Boolean search string: "("computer-assisted" OR "computer assisted" OR "automated" OR "automatic" OR "computational" OR "machine learning") AND ("content analysis" OR "text analysis")." This resulted a total of 192 relevant papers. Among them, 95 articles were determined not relevant (e.g., an overview/introduction article, qualitative analysis, or studies using unsupervised methods or simple keyword frequencies, etc), therefore excluded from the further analysis. Using remaining 97 articles, we systemically examined whether extant applied studies using dictionary- or supervised methods a) adequately employ any cross-validation of their primary findings, b) if so, whether they use human coded materials as a benchmark, and c) if so, whether intercoder relabilities and other methodological details are adequately and consistently reported.

A total of five highly qualified coders tested the initial coding scheme developed by the first and third author by independently coding 10 sample articles (approximately 10% of total sample) and collectively discuss the coding problems and disagreement. Based on this iterative procedure, coding instructions were revised until the coding schemes would produce reliable results among any trained and competent coders. Intercoder reliability was assessed for each of the variables coded, using Krippendorff's alpha, which is reported in Table XX below.

Results

The results for each variables examined are presented in Table XX below.

Study 2: A Monte-Carlo Simulation Study

The result of Study 1 reveals TBA.

In order to systematically evaluate how different setups and practices of utilizing manual coding within automated procedures can yield different conclusions, we setup an extensive set of Monte Carlo (MC) simulations ($N = 240,000$). MC simulation offers a convenient tool for systematically evaluating the relative bias and coverage of a given statistic under certain scenarios (for an example, see Leemann & Wasserfallen, 2017; Scharkow & Bachl, 2017; Stephenson & Holbert, 2003).

While we setup 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 technique. Therefore, here we develop 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 vary the intercoder reliability of the "gold standard" material, along with a number of related factors such as number of coders and the number of manual annotations per coder (see below section). Then we systematically compare different scenarios in terms of their precision, recall, and F1 scores based on the "true" standard (i.e., a quantity of interest that is typically unknown to researchers) in order to illustrate how different practices of human coding in automated content analyses affect the overall results and the relative trustworthiness of conclusion drawn from such results.

Data Generation Stage

We create data (e.g., textual data to be analyzed by a researcher) with the "true" outcome variable of interest; the goal of any quantitative text analysis method is 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 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:

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

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, $\beta_1 = .5$, $\beta_2 = .2$, and $\beta_3 = .6$, 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 a separate Dirichlet priors for negative and positive draws (i.e., a discrete value range range from -5 to 5), where they represent some "features" of a given textual data (e.g., a word or a N-grams). These values then are later to be matched with a "dictionary" (which has similar discrete values that represent a predefined category) in order to determine the estimated values of *y* for each observation. Since this requires discrete rather than continuous values, here we use slightly different setup for the dictionary approach. Based on those selected textual features, the *true* values of *y* are then sampled from a Binomial distribution as discussed above, with an exception that now the vector $\boldsymbol{\beta}$ is extended to K = 10 and values are fixed to .2, to better approximate the

dimensionality of text data.

$$\pi_{neg} \sim Dirichlet\big(K, \alpha_{neg}\big)$$
$$\pi_{pos} \sim Dirichlet\big(K, \alpha_{pos}\big)$$

(2)

with K being the desired number of independent variables, and α_{neg} and α_{pos} — concentration hyperparameters with mass concentrated closer and further away from 0, respectively. The negative and positive draw probabilities π_{neg} and π_{pos} are then randomly selected for each draw and a single row of independent variables is drawn.

$$\boldsymbol{X} \sim Categorical\big(K, \big[\pi_{neg} \text{ or } \pi_{pos}\big]\big)$$

(3)

And finally, having generated the set of independent variables, we generate y in a similar fashion as discussed in (1).

$$y \sim Bernoulli(\pi)$$
$$\pi = logistic(\mu)$$
$$\mu = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

(4)

This means that y becomes effectively a function of the simple sum of chosen textual features.

Human Coding Stage

We specified several typical scenarios of how the true data would be coded by human coders for the purpose of validation of automated procedures. In typical automatic content analysis applications, at least two or more number of trained human coders are assigned to a common, small set of sample documents for manual coding, and independently code such documents for intercoder reliability testing. This process is repeated until the level of intercoder reliability is deemed satisfactory (typically Krippendorff's alpha equal or greater than 0.7). Theoretically, achieving the satisfactory level of reliability means that coders can be treated as interchangeable to each other (Krippendorff, 2004), therefore any observations produced by a single human coder in validation materials are, on average, likely to be similar to each other

proportional to the chosen degree of intercoder reliability. Once such a coder interchangeability is ensured, the common practice is that remaining documents (for validation materials) are evenly — typically by some convenient standard but non randomly — divided and only coded by a single coder (Grimmer, King, & Superti, 2018). While the total number of manual annotations produced by such human coders greatly vary across studies, the practice of evenly (but not randomly) dividing the validation materials across coders effectively means that any potential coder idiosyncrasies are just ignored while potential coder idiosyncrasies are not randomly distributed across manual annotations produced for validation materials.

Following the typical application in this area as sketched above, we assume that only a small fraction of data (i.e., held-out sample) is analyzed manually by a varying number of human coders for the purpose of providing a cross-validation result (against machine algorithms) that informs the performance of such automated procedures. Following this general setup, we specified few important factors that may affect the quality of the "golden standard" and therefore of the overall performance evaluation of later automated content algorithms. Those factors include: the number of human coders (k = 2, 4, 7, 10), the predetermined target values of intercoder reliability metric (Krippendorff's alpha = 0.5, 0.6, 0.7, 0.8, 0.9), as well as the number of articles that each coder had to analyze in order to produce final human-coded data (the number of annotation per coder, n = 50, 100, 250, 500). Although somewhat arbitrary, these numbers were chosen to reflect typical content analysis procedures and their variations commonly carried out in social science applications, as evidenced in our Study 1.

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 hyperparameters, which effectively enables us to simulate a situation in which where a certain coders produce more "correct" judgment ($\hat{y} = y$) whereas other coders produce more "false" judgment ($\hat{y} \neq y$), yet the overall

human coding pattern at chosen level of a beta distribution parameters would produce an acceptable level of target reliability coefficient (e.g., Krippendorff's Alpha = 0.7). Depending on specific application (dictionary-based vs. supervised learning), the output (coded data) from this stage is then used as either a post-hoc validation data or as an input for the later automated classification stage.

Algorithm-based Classification Stage

Following the typical application, we assumed a researcher uses a certain classification algorithm to predict the values of $(y = \hat{y})$ in each of the observations. For this purpose, we used three different classification algorithms – binomial GLM, Naïve Bayes (NB) classifier, and a bag-of-words dictionary approach. Although these are not exhaustive, these methods are among the most frequently utilized applications in automated contents analysis, especially coupled with human validation of such algorithms, as can be seen in Table 1 in our Study 1 result.

Typically in a real-world scenario, researchers validate the computer-based coding based on a held-out human coded test set post hoc (for a bag-of-words dictionary approach), or use such human coded data to develop prediction algorithms and then "scale up" the automated coding process (for supervised learning, such as binomial GLM or Naïve Bayes). This means that we can also systematically compare different levels of precision, recall, and F1 scores (that are evaluated at different levels of human benchmark scores) against the true standard. For this purpose, we compare the estimated $\hat{y}$ values from scaled-up automated coding to the true values of y. 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 and coverage of proposed methods when using imperfect human coding as "gold standard." Also by doing so, we also can compare how the overall accuracy is diminished from one step to another based on which factors.

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

Simulation Results

TBA

Conclusion and Discussion: TBA

Our contribution should be read as a call for a thorough and systematic application of validation procedures in any publication drawing on automated text analysis procedures. At the same time, the study serves to benchmark the combination of reliability in gold standard/ ground truth and validity scores and warns against improper use of both to demonstrate the validity of the approach

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

Riff, 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.

Stephenson, M. T., & Holbert, R. L. (2003). A monte carlo simulation of observable versus latent variable structural equation modeling techniques. *Communication Research*, *30*(3), 332–354.

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

van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (2008). Good news or bad news? Conducting sentiment analysis on dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics*, *5*(1), 73–94.

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