Towards a framework for creating trustworthy measures with supervised machine learning

Ju Yeon Park
Assistant Professor
The Ohio State University
Email: park.3509@osu.edu

Jacob M. Montgomery
Professor
Washington University in St. Louis
Email: jacob.montgomery@wustl.edu

July 14, 2024

Abstract

Supervised learning has become a staple in social science research for quantifying abstract concepts within textual data. However, a survey of recent studies reveals inconsistencies in reporting practices and validation standards. To tackle this issue, we introduce a framework that delineates the process of converting text into a quantitative measure, highlighting critical reporting decisions at each stage. We emphasize the importance of clear and comprehensive validation in the process, allowing readers to critically assess both the methodology and the derived measure. To showcase our framework, we develop and validate a measure assessing the tone of questions directed at nominees during US Senate confirmation hearings. This study contributes to the growing literature promoting transparency in the application of machine learning methods.

1 Introduction

In the past decade, computer-assisted text analysis has evolved from a rarely used technology to an increasingly ubiquitous approach in political science. Access to massive volumes of digital text combined with advances in computing and algorithm designs make text analysis easier to implement and more relevant in research domains as diverse as authoritarian government censorship (Esberg, 2020), state repression (Gohdes, 2020), news exposure (Stier et al., 2022; Guess, 2021), political advertising (Fowler et al., 2021), congressional committees (Casas et al., 2020; Park, 2021), and more.

One common task is measuring important latent concepts embedded within text via supervised machine learning. Generically, we imagine a researcher who wishes to measure a predefined construct, such as negativity (Fowler et al., 2021), grandstanding (Park, 2021), or Islamophobia (Alrababa'h et al., 2021). The goal is to create a mapping from the written text to a valid quantitative measure. However, doing so requires multiple steps including data partitioning, label acquisition, text preprocessing, and model fitting. At each stage, researchers must make choices, many of which are consequential to the results. Thus, creating a measure can be conceptualized as a set of interconnected procedures. It is a *pipeline*, where the outputs of each stage are passed on to the next so that the final result reflects the cumulative decisions made at each point.

Although these methods are powerful and increasingly accurate, when combined with current standards for reporting, the complexity of this pipeline presents a problem. To begin, few of the decisions made at each stage are justified to readers and often they are not even described. This lack of transparency is particularly problematic given the sequential nature of the pipeline; information about one step of the process in isolation can be uninformative or even deceptive. Worse, standard practice provides readers with little objective evidence about the validity of the procedures or even the final outputs. Together, this means that far too often readers are presented with statistical analyses of text-based measures where the measurement procedure has not been adequately explained, evaluated, or validated.

To address this limitation, in this article we provide a conceptual framwork for understanding, reporting, and validating the text-to-measure pipeline. Like the *total survey error* approach in public opinion research (e.g., Althaus et al., 2022; Groves and Lyberg, 2010), our aim is to think holistically to identify when we might introduce errors into our pipeline and therefore what should be reported. We outline each stage and offer a concise discussion and checklist of key decisions for researchers to report.

Throughout, we emphasize the connection of supervised learning with the social science task of measurement (Grimmer et al., 2021; Ying et al., 2021). Thus, in addition to encouraging transparency, we emphasize the importance of robust validation, and identify three stages in the pipeline where it is possible to transparently assess performance for readers – making explicit the Grimmer and Stewart (2013a) adage to, "validate, validate, validate."

Throughout, our goal is not to criticise or target prior research. Nor do we intend to provide a definitive "best practices," since the text-as-data field is too diverse and methods are evolving too quickly to make this feasible. In any case, best practices for any particular measurement exercise are too context-dependent for any advice to be authoritative. Rather, this paper adds to the on-going efforts in the computational social science community to establish a frameworks for thinking about, reporting, and evaluating text-as-data methods (Ying et al., 2021; Kapoor et al., 2022), focusing specifically on the task of using supervised learning to measure latent concepts. While these issues have certainly been discussed before (c.f. Grimmer and Stewart, 2013a; Grimmer et al., 2022), our review of current practices in the field shows that the literature would benefit from an approachable and coherent explanation of the various steps involved, the key decision points for researchers, and a framework for making empirical results more credible and transparent to readers.

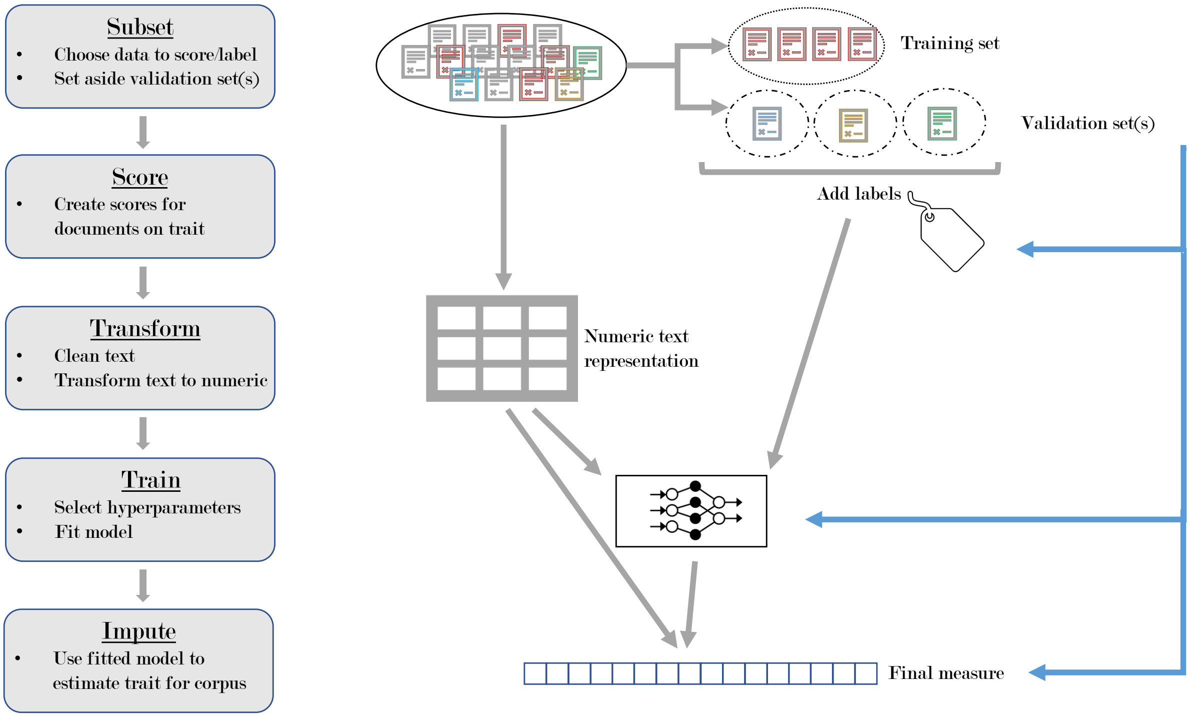
In the next section, we provide brief overview of the process of creating a new measure, beginning with choosing the cases to label and continuing through label acquisition, text representation, model fitting, and out-of-sample imputation. We then review current practices in the field in terms of whether and how these stages are reported and evaluated. We find that there is currently a great deal of heterogeneity in reporting practices and that robust validation at any stage of the pipeline is surprisingly rare, even in the field's top journals.

To address this, in Section 3 we propose a framework that summarizes the measurement pipeline and discuss key decisions that researchers may wish to report to readers. We emphasize that researchers have the opportunity to conduct validation exercises at (at least) three different stages: labeling, in-sample model fitting, and out-of-sample imputation. We argue that doing so will provide more transparency to readers and confidence in the validity of the final measure. In addition, it can help improve the quality of the final measures by identifying problems or errors in the pipeline. In Section 4, we then illustrate our framework with an application to questions asked by US senators to nominees during confirmation hearings from 1997 to 2019. Specifically, we build and validate a new measure of tone in over 89,000 statements made by senators in these hearings. We conclude with a brief discussion of how these recommendations might shift as the field more widely adopts alternative measurement strategies based on large language models.

2 A supervised learning pipeline for measurement

To limit the scope of this discussion, we make several assumptions about the research goal or "use case" we address. First, we suppose that the researcher has a predefined concept of interest, thus excluding exploratory techniques, such as unsupervised topic models. Second, the investigator's assumed goal is to create a valid measure at the *document* level, such that each speech, tweet, or article is scored. Third, we assume that the researcher wishes to use a supervised learning method. We discuss how our framework relates to recent innovations, such as zero-shot learning, in the discussion because although these methods are promising (Ornstein et al., 2022), they remain relatively novel and are still nonstandard in the field and to our knowledge have yet to appear in top political science journals.

Figure 1: Summary of the supervised learning pipeline


Subset
Choose data to score/label
Set aside validation set(s)
Score
Create scores for documents on trait
Transform
Clean text
Transform text to numeric
Train
Select hyperparameters
Fit model
Impute
Use fitted model to estimate trait for corpus
Training set
Validation set(s)
Add labels
Numeric text representation
Validate opportunities
Final measure

This figure visually depicts the text-to-measure pipeline for supervised machine learning including document subsetting, labelling, text transformation, model fitting, and imputation.

2.1 Pipeline overview

The goal in the text-to-measure pipeline is to train a model using a subset of the data so that values for the latent variable in the broader corpus can be imputed. A typical example is in Fowler et al. (2021, pg. 135), which states:

> "We had research assistants classify a training sample of the Facebook advertisements on issue and tone dimensions and then used these classifications ... as the basis for a supervised learning classification procedure ... . The fitted model from this process then produces predicted values for tone and issue content, which we use as our measure of these quantities for all ads in the data set."

A schematic for this exercise[1] is shown in Figure 1. First, we select a subset of the data to use to fit the supervised model. At this stage, we also select and hold out documents for downstream validation. Second, we label or score the document in the training set on the latent trait of interest. In the simplest case, this is done manually by research assistants as described in the quote above, although other approaches such as crowdsourcing or generative AI are also feasible. Third, we preprocess the text and extract a set of features that can be represented numerically. For example, we might remove punctuation and create a term-document matrix (See SI Section B for more details). More advanced preprocessing steps are possible, such as named entity recognition, which identifies important persons or places in the text (Fowler et al., 2021, e.g.,). Fourth, we fit some machine learning algorithm(s) to accurately predict the labels as a function of the inputs, typically requiring the selection of hyperparameters. Finally, we impute the latent trait with the fitted model for the entire corpus.

This approach is attractive for several reasons. Most importantly, supervised learning is much more cost-effective relative to labeling the entire corpus through a manual or crowd-sourced process. Although these approaches are possible for tasks like open-text responses

[1] Given the variety of methods and applications in the literature, an engaged reader can find exceptions to nearly every characterization we make in this section and the next. In some cases, we will try to note important ones. But there is an inherent trade-off in providing a maximalist summary of the complete literature and providing a concrete discussion of general issues involved. This paper adopts a "generalist" perspective, while freely admitting that we may not be doing justice to every paper in the literature.

(e.g., Bøggild et al., 2021), the size of contemporary text corpora can quickly go beyond the budgets of any research team. Fowler et al. (2021), for instance, analyzed over 400,000 Facebook ads. In our own example below, the corpus includes 89,279 statements.

In addition, supervised learning makes it relatively easy to impose a research question. We are not making discoveries within the corpus or interpreting outputs, but rather building a model designed to precisely estimate a predefined concept. This makes the output more interpretable for downstream tasks, such as theory testing, and can perform much better than dictionary methods (Grimmer and Stewart, 2013a).

The downside is that the end-to-end process can become elaborate; to some, it even may seem convoluted. Furthermore, at each step, researchers must make many decisions, often with little guidance. This adds many "researcher degrees of freedom" to the process and makes reporting results cumbersome. Therefore, researchers should consider the following questions. Which decisions need to be detailed to the readers? And, what evidence should we provide to give the reader confidence in the procedures and the validity of the measure?

2.2 Current practices for reporting and validation

To motivate our discussion in the following sections, we first review current practices in the field in terms of how these steps are reported and validated. To do so, we searched the 2020-2023 volumes of the *American Political Science Review* (APSR) and the *American Journal of Political Science* (AJPS) for all articles that used supervised learning of texts to create measures of latent concepts. In total, we identified fourteen observations and reviewed the article and the appendices.[2] The results are shown in Table 1.

First, most cases (10/13) provide at least some details of how a subset was selected for scoring.[3] However, in only three out of fourteen articles were any observations "held out" in advance for downstream validation.

[2] We focus only on steps that are reported. It is not possible to know all the validation procedures that the research team followed, and we believe that more was done to assess the quality of the measurement than is ultimately reported. However, it is instructive to review what is disclosed to readers.

[3] Anastasopoulos and Bertelli (2020) relied on a sample selection process justified in previous research but not elaborated in the text. We coded this observation as missing.

Table 1: A review of current reporting and validation practices applying the text-to-measure pipeline

	Yes	No	%
`Subset`			
Scoring subset explained	10	3	77%
Validation set held back	3	11	21%
`Label`			
Coding rules described	8	6	57%
Reliability or validity assessed	4	10	29%
`Transform`			
Text pre-processing described	10	4	71%
`Train`			
Model tuning explained	9	5	64%
Model validation	13	1	93%
Validation with held-out sample	2	12	14%
Validation beyond prediction	6	8	43%
`Predict`			
Validation of final measure	5	9	36%

Analysis of transparency in the text-to-measure pipeline for articles using supervised machine learning in the *APSR* and *AJPS* (2020-20223. The total number of articles changes since not all categories are relevant to all projects.

Second, we examine how document scores were assigned and evaluated. Only eight out of fourteen articles reported any details for labeling/scoring. In two other cases, the articles refer to existing coding schemes that are not explained in the article itself (Fowler et al., 2021; Wahman et al., 2021). It is even less common for articles to provide assessments for these scores. Only four out of fourteen (29%) evaluated the quality of their labels. Alrababa'h et al. (2021) report inter-coder reliability. Schub (2022) reports inter-coder reliability and provides two example posts to support the face validity of the labels. Only Zubek et al. (2021) and Emeriau (2023) report steps to validate the labels more extensively.

Next, it is common to describe text preprocessing steps. Ten out of fourteen articles provided at least some details. Model validation of some sort is almost universal (13/14), mostly appearing as some form of within-sample fit statistic (e.g., precision and recall for binary classifiers). However, other forms of model validation[4] were rare. The only clear example is Schub (2022) and Stier et al. (2022), which report highly predictive word stems.

[4] As we discuss below, this refers to inspecting parameters or other features of the model to validate that it is relying on text features that "make sense" for the given task and context.

In addition, assessment of fit almost always were conducted *within* the training sample.[5] This raises concerns because in many cases the machine learning model was selected or tuned based on performance within the training sample. For example, Casas et al. (2020) fits a large number of models and chooses the most accurate to create an ensemble. If models are tuned and evaluated on the same sample, this increases the risk of overfitting. However, in our review, only Hager and Hilbig (2020), Wahman et al. (2021), and Emeriau (2023) preserved cases for out-of-sample assessments. In addition, whether the tuning and testing were done on the same samples is difficult to assess because the tuning is explained or justified in only nine out of fourteen articles.[6]

Finally, and most critically, we examined whether the final scores – the measure to be used in downstream analyses – were validated. Stier et al. (2022), for instance, compares the final scores to a researcher-created metric to assess convergent validity. Anastasopoulos and Bertelli (2020) provides a validation by replicating basic findings from the existing literature. Gohdes (2020) provides the reader with a random selection of documents along with their predicted scores to assess face validity. Zubek et al. (2021) adopts a combination of these strategies.[7] Emeriau (2023) shows that the scores correlate with document metadata as expected by theory. However, in the remaining articles – nine out of fourteeen articles (64%) in the field's top journals – no validation of the final measure is provided despite the fact that they often serve as a main explanatory or dependent variable.[8]

In all, we found that current reporting standards for the text-to-measure pipeline are uneven at best and, in some ways, surprisingly sparse. While some steps are commonly reported, others are rare, including (surprisingly) assessing the quality of the labels and final measures. In other words, in many articles neither the model inputs nor the model outputs

[5] In some cases, articles report fit statistics but do not specify whether or not they are calculated within the sample. We assume here that these are in-sample statistics unless a test set is specified.

[6] However, two models relied on naive Bayes classifiers, which may require no tuning.

[7] It is interesting to note that both Anastasopoulos and Bertelli (2020) and Zubek et al. (2021) are research notes focused on developing and evaluating a new measure or measurement technique. Extensive validations are, perhaps, much more important where the measure itself is the major output.

[8] Out of the fourteen studies, the two not mentioned in this section are Park et al. (2020), Guess (2021), and Malesky et al. (2023).

are validated, leaving readers to judge the validity of the measure based only on in-sample predictive performance, which, in the absence of validated labels, does not actually speak to the quality of the measure.

Although this situation is worrying, it is important to recognize that there is likely more going on. We suspect that researchers are conducting additional analyses, but they are not included in the manuscript since there are few clear expectations about what *should be* reported. In fact, one potential problem may be pressure from reviewers and editors to remove critical information as extraneous. To address this, in the next section we articulate a framework for self-assessment and reporting of the text-to-measure pipeline.

3 Transparently reporting the pipeline

With the general schematic shown in Figure 1 in mind, we now discuss the most critical aspects of the text-to-measure pipeline that researchers should report. For convenience, we provide a checklist in Figure 2 that captures these key considerations.

First, researchers should report how they subset the overall corpus for labeling and validation, as this decision can significantly affect downstream outcomes. Ideally, readers should be able to assess the representativeness, size, and diversity of the labeled corpus. If the labeled set is unrepresentative, it can introduce subtle distortions. For instance, in our upcoming example, selecting only recent congressional documents might overemphasize temporal-specific features (e.g., "trump" or "vaccine") that are not dispositive for earlier periods. Similarly, Anastasopoulos and Bertelli (2020) notes that relying solely on major legislation texts could be consequential.

Random sampling is a simple strategy to ensure representativeness. However, researchers may choose to over-sample specific traits or use block sampling based on existing dictionaries for a balanced training set. In some cases, like Guess (2021), the sample is built based on available labels (news articles from specific web domains).

The optimal number of documents to score depends on various factors: sampling strat-

Figure 2: A checklist for the text-to-measure pipeline

1 Subset

- □ *Representativeness*: Is the training set representative of the larger corpus? If not, how might this bias the final measure?
- □ *Size*: Is the training set large enough for effective downstream learning?
- □ *Diversity*: Does the training set cover the full range of the latent dimension(s) of interest?

2 Label

- □ *Reliability*: Is the labeling procedure consistent with low error rates?
- □ *Validity 1*: Do the labels accurately represent the latent concept? Can convergent or face validity be assessed?

3 Transform

- □ *Feature reduction*: Is the input space dimensionality reduced enough to identify relevant features with the given training set?
- □ *Feature preservation*: Are critical textual features retained in the labeled set?

4 Model

- □ *Validity 2*: Do the most heavily weighted features align with the substantive context (face validity)? Does the model accurately predict within the training set (accuracy)? Does the model, fit on the training set, accurately predict a held-out validation set (regularization)?

5 Impute

- □ *Validity 3*: Is the final measure a valid representation of the latent concept? Can convergent, face, or predictive validity be assessed?

egy, distribution and complexity of the latent trait, text feature processing, and learning algorithms. Researchers can rely on downstream validation tasks, assuming that a valid final measure implies adequate sample size. Alternatively, they can use a "learning curve" approach, adding observations until out-of-sample prediction rates plateau (Mohr and van Rijn, 2022).

Researchers should also report the diversity of the labeled subset along key traits. This is crucial because any feature absent from the scored set cannot inform the learning algorithms about its mapping into the latent space. At a minimum, researchers should disclose the variation of the labeled set along the latent dimension of interest.

Second, researchers should detail their labeling process: who did the coding (e.g., researchers, trained RAs, or online workers), what rules were followed, and any steps taken to improve label quality (e.g., multiple coding). They should assess both reliability, meaning the consistency of labels (e.g., inter-coder agreement), and validity. Validity can be established through face validity (providing readers with specific documents and their predicted scores), convergent validity (showing that imputed values correspond with related measures as expected), or some other validation strategy (Grimmer and Stewart, 2013b).

Third, the textual input's feature space must be reduced to a manageable dimensionality. Standard steps include tokenization to create a term-document frequency (TDF) matrix (Grimmer and Stewart, 2013a), removing certain words, or stemming (Denny and Spirling, 2018). More dramatic transformations like word embeddings are also options (Mikolov et al., 2013; Rodriguez and Spirling, 2022). These steps are useful to report because insufficient reduction can leave important features missing, while over-reduction can eliminate critical information.

Fourth, researchers train models to predict labels for unlabeled documents. They should disclose model fitting procedures and justify hyperparameter choices. Model validation can involve assessing predictive accuracy within the scored set. However, for complex, tuned models, this risks overfitting. To mitigate this, researchers can evaluate predictive accuracy on a held-out validation set.

Additionally, researchers can validate model parameters or outputs for contextual sense. For example, Stier et al. (2022)'s model for identifying political news articles heavily weighted words like "trump," "president," "house," "mueller," "democrats," and "campaign," reassuring readers that predictions are driven by sensible features.

Finally, scores are imputed for all documents to create the final measure. The quality of this measure depends on all previous steps, sometimes in non-obvious ways. Therefore, multiple validation methods are recommended including convergent validity, face validity, and predictive validity. We provide explicit examples in the following section (see also Goet, 2019; Grimmer and Stewart, 2013a).

In summary, this section provides a framework for transparently reporting key decisions and validation steps in a supervised text-to-measure pipeline. By following the checklist in Figure 2, researchers ensure readers understand critical choices at each stage: subsetting the corpus, labeling training data, transforming textual features, modeling feature-label relationships, and imputing scores. This transparency enhances reproducibility and allows readers to critically evaluate the measure's validity and reliability. To illustrate these principles in practice, we now turn to a concrete example: measuring the tone of Senators' statements in 916 confirmation hearings. By walking through each pipeline step, we demonstrate how careful reporting can bolster confidence in the final measure and provide insights for future applications.

4 Application: Senate confirmation hearings

In this section, we turn to an example that illustrates our framework as well as a specific set of decisions to operationalize our recommendations. Again, we emphasize that we view these particular choices as appropriate for *our* application. In other contexts, other decisions may be preferable. However, we believe that this example will serve as a template that other researchers can build from, alter, and revise to suit their own purposes.

In this example, we measure the positive or negative tone of senators' statements in 916 US Senate confirmation or nomination committee hearing transcripts. These are from the 105th-115th Congresses and include both bureaucratic and judicial nominations. We collected these transcripts from the Government Publishing Office, Congress.gov, and ProQuest Congressional. Typically, a committee hearing transcript records statements made by both members of Congress and witnesses. We focus only on statements made by committee members.[9] In total, our corpus includes 89,279 member statements.

Our goal is to estimate the tone of the questions senators asked of nominees. An example

[9] The following types of procedural statements are not included: the initial and final statements of a hearing with eighty words or fewer; statements that come before the first witness's opening statement, which usually outlines the witnesses' backgrounds; statements with less than or equal to eighty words that include "come to order," "recognize," "expired," "yield," "adjourn," or "recess"; and statements with less than or equal to fifty words that include both "thank you" and "yield."

is the following statement:

> "If you are confirmed, would you be willing to hold a moratorium until you are sure that all of the local law enforcement officers have been properly trained and understand what the authority is under 287(g), which is not to stop people just based on the color of their skin? I am wondering whether or not you would be able to move in that direction."

4.1 Partitioning

We used a simple random sampling scheme to select 3,600 paragraphs for labeling and/or validity testing.[10] The paragraphs were chosen through the following procedure: statements containing more than 120 words and composed of multiple paragraphs were broken down into paragraphs; paragraphs that were too short (containing less than fifty words) were combined with adjacent paragraphs in the same statement; subsequently, among those containing fifty to 120 words,[11] 3,600 paragraphs were chosen at random. The 3,000 documents used to train our final model (Groups A and B described below) include statements from 781 unique hearings and 200 unique senators (45.53% Republican and 54.47% Democrat).

These 3,600 paragraphs were grouped into four subsets with 2,900 (Group A), 100 (Group B), 500 (Group C), and 100 (Group D) paragraphs, respectively. Table 2 summarizes our partitioning strategy. The first three sets (n = 3,500) were labeled by Amazon Mechanical Turk (MTurk) online workers using a common procedure described below. We hand-labeled 200 documents (Groups B and D) on a five-point scale for downstream validation.[12] The 500 paragraphs in Group C were labeled but held out from the model building, and the 100 documents in Group D were excluded from the entire procedure.

The partitioning strategy we propose is valuable in that it allows us to perform validations at three different stages. First, we can validate the crowdsourced labels by comparing them in

[10] The number of paragraphs was chosen based on the previous research (Park, 2021) that used the same labelling strategy we use and produced a measure with reasonably good validation metrics.

[11] These lengths were determined by reading subsets of speeches and considering the length where semi-trained workers could reliably summarize the tone.

[12] The following was our coding scheme: 1 = very negative; 2 = somewhat negative; 3 = neutral; 4 = somewhat positive; 5 = very positive.

Table 2: Partitioning of the labeled data for training and three forms of validation

	Group A (n=2900)	Group B (n=100)	Group C (n=500)	Group D (n=100)
MTurk label	✓	✓	✓	
Expert label		✓		✓
Held out from model building			✓	✓

Group B with the expert labels. Second, we can validate our model using a held-out sample by training the model on Groups A and B only, and comparing the model's prediction for Group C to its crowd-sourced labels. Third, we can validate our final measure using a held-out sample by comparing the model's prediction for Group D with our own coding of the documents. We explain these procedures in more detail in the following sections.

4.2 Labeling

We used Amazon Mechanical Turk (MTurk) workers to create a measure of tone based on pairwise comparisons (Carlson and Montgomery, 2017). Briefly, we randomly paired two documents in the sample and let workers choose the one that exhibited positivity more strongly. We created 25,000 random pairs to generate the same number of human intelligence tasks (HITs). Each document appeared in the pairwise comparisons with a similar frequency. We then used the Bradley-Terry model described in Carlson and Montgomery (2017) to estimate the latent position of each document on a continuous scale, which we refer to as the crowdsourced score of the documents. Figure B2 shows the distribution of the measure. We explain coding instructions and worker management in the SI Section B. Following (Carlson and Montgomery, 2017), we assess the reliability of the measure by comparing estimates using only the first half of the HITs to estimates generated from the second half of the HITs. These estimates are correlated at 0.655.

It can often be helpful to validate whether our labels capture our concept of interest, as low-quality labels can lead to a low-quality final measure. One option is to check the *face validity* of the measure. Table B2 shows ten randomly chosen paragraphs along with their estimated scores.[13] These generally confirm that more negative statements have a more negative score.[14]

Second, we also assess the *convergent validity* of our measure by comparing it with our expert coding. For 100 of these documents (Group B in Table 2) we hand-coded them on a five-point scale. The correlation coefficient between the two scores is 0.808 (see Figure B3), again suggesting that the labels are a valid measure.

4.3 Feature Representation

As we had no *a priori* reason to prefer any specific preprocessing approach, we created an ensemble of models based on term-document frequency representations (TDF), embeddings, and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019).

For the TDF, we remove punctuations and numbers, put words into lowercase, stem, remove stopwords,[15] include unigrams and bigrams, and remove infrequent words (appearing less than 5,000 times). To determine which of these preprocessing steps to choose, we used the Support Vector Machine (SVM) model in the `Kernlab` R package and chose the alternative in each preprocessing step showing the best prediction power as measured based on ten-fold cross validation in the training sample. The validation statistics we relied on are the out-of-fold RMSE and the correlation coefficient between the crowdsourced scores and the model prediction. The end result was a TDF representation with 37,359 words. Of these, 8,496 (22.75%) words are contained in the training set.

For embeddings, we use the `word2vec` model, which is an unsupervised learning model also known as distributed vector representation of words (Mikolov et al., 2013). This model assigns each word a numeric vector that captures the context in which the word appears in a corpus by measuring its relationship with the words surrounding it. We first took prepro-

[13] Note that random sampling of the documents is important to assure readers that examples have not been cherry picked.

[14] The sentiment measure in our example is a relatively straightforward concept. Measuring a more complicated concept may require additional pilot trials to optimize wordings in the instructions, training of the workers, sampling of the data, etc.

[15] The list of stopwords we used is provided in the SI C.

cessing steps, which includes removing punctuations, numbers, and stopwords, lowercasing, and stemming. To tune embeddings for our task, we varied four key parameters generating ninety-six different combinations: 400, 500 and 600 for the dimension of a word vector; 5, 7, 10 and 12 for the window size; 10, 15, 30 and 45 epochs; only unigrams and both uni- and bigrams. Among the ninety-six model specifications, the one with 400 dimensions, a window size of 5, and 15 epochs including only unigrams performed the best in our cross validation.

BERT is an unsupervised deep learning model that was pre-trained on a large corpus to capture representation of words and sentences and their semantic relationship. It can be fine-tuned with small data at the user's end (Devlin et al., 2019). As a tuning parameter, we varied epoch size and chose the one that resulted in the smallest loss.[16]

4.4 Fitting Learners

To fit our model, we combined Group A and Group B in Table 2 to create a training set of 3,000 statements.[17] We fit thirteen models in total and combined them using ensemble Bayesian model averaging (Montgomery et al., 2012). Table C4 provides a complete list of these learners and shows their predictive performance where prediction is calculated with a cross validation within our 3,000 document training set.[18]

Assessing the model itself is a second validation point. However, we have used the training data set to (i) determine the best approaches for text preprocessing, (ii) tune individual models, and (iii) create ensemble weights for the final model. Reusing the training data in this way – even with techniques like cross validation – risks overfitting.

Therefore, we used our model to predict the score on the 500 documents held back for model validation (Group C). We found that the RMSE for this set was 0.537. This compares favorably to the RMSE of the best single learning algorithm, 0.604, the Support

[16] For BERT, we used the English uncased preprocessing model version 3. For more details, see `https://tfhub.dev/tensorflow/bert_en_uncased_preprocess/3`

[17] Section B.3. in the SI demonstrates how the model's prediction performance changes as we change the size of the training set.

[18] In addition to BERT, we fit six learners on each of the TDF and word2vec based document matrices. These six learners are Support Vector Machine, Kernlab's Support Vector Machine, LASSO, Random Forest, Bayesian GLM, and Gradient Boosting Machine.

Vector Machine using the doc2vec matrix.[19] The Pearson's correlation coefficient between the crowdsourced scores and the ensemble scores is 0.757 (see Figure C6). In all, these results suggest that the ensemble learner can accurately, if imperfectly, predict crowdsourced labels from textual features.

In addition, we checked the most frequent word stems representing positive and negative tones obtained through the following procedure. First, we selected paragraphs whose scores fell in the top quartile (most positive statements) and the bottom quartile (most negative statements) of the entire corpus excluding the scored set[20] and extracted the 300 most frequent word stems from each side.[21] Then, we excluded 178 overlapping features to create a list of most frequent and exclusive stems.

We present the top 50 words from each category in Table C5 in the SI and find the following patterns: positive statements are featured with appreciations (e.g., "thank-much", "appreci"), greetings (e.g., "pleas", "honor"), endorsement (e.g., "proud", "friend") and, most importantly, compliments on nominees' ability (e.g., "experi", "energi", "distinguish", "leadership") or work-related experiences (e.g., "ambassador", "develop", "director", "manag", "univers"); in contrast, negative statements tend reference problems (e.g., "problem", "matter"), personal views (e.g., "agre", "opinion", "view"), money (e.g., "$", "money",), fact-checks (e.g., "correct", "report", "percent", "quote"), and rules (e.g., "constitut", "rule", "regul", "standard", "suprem-court"). Given the context, these results give us further confidence that the model is leveraging appropriate textual features to measure our latent trait of interest.

4.5 Final Output

Model validation provides evidence that our learner can accurately *predict* the labels, and we also took steps to validate the labels themselves. In some cases, this may be sufficient. However, we can also assess the validity of the final measure itself.

[19] RMSE for all component models ranged from 0.604 to 1.024.

[20] Another alternative would be to focus on the training set for this validation.

[21] We used the "topfeatures" function in the `quanteda` R package.

To assess the *convergent validity* of the measure, we turn to the 100 paragraphs we scored on a five-point scale but were held back from all of the previous steps (Group D). The correlation coefficient between the two measurements is 0.747 suggesting that the expert ratings and the ensemble predictions are picking up the same latent trait (see Figure C5).

To assess the *face validity* of the measure, Table 3 provides five examples of scored statements along with the five-point scale expert coding and the scores generated by the ensemble prediction. These were randomly selected to ensure that one document from each of the five-point categories was included. In these examples, both scores go hand in hand. The statements scoring lower tend to convey more negative tone than those scoring higher, which provides an excellent face validity of our ensemble prediction score.

Finally, we check predictive validity by analyzing whether members' speaking tone in Senate confirmation hearings correlates with institutional factors in theoretically expected directions. First, we expect that members' speaking tone should become more negative over time due to the increasing level of partisan polarization (McCarty, 2019) and intensified partisan conflict in Congress (Lee, 2016). Second, members' questions asked to appointees across party lines should be more negative than questions asked to appointees from their own party. Third, we expect that tone will be more negative in judicial hearings relative to nominations to the bureaucracy.[22]

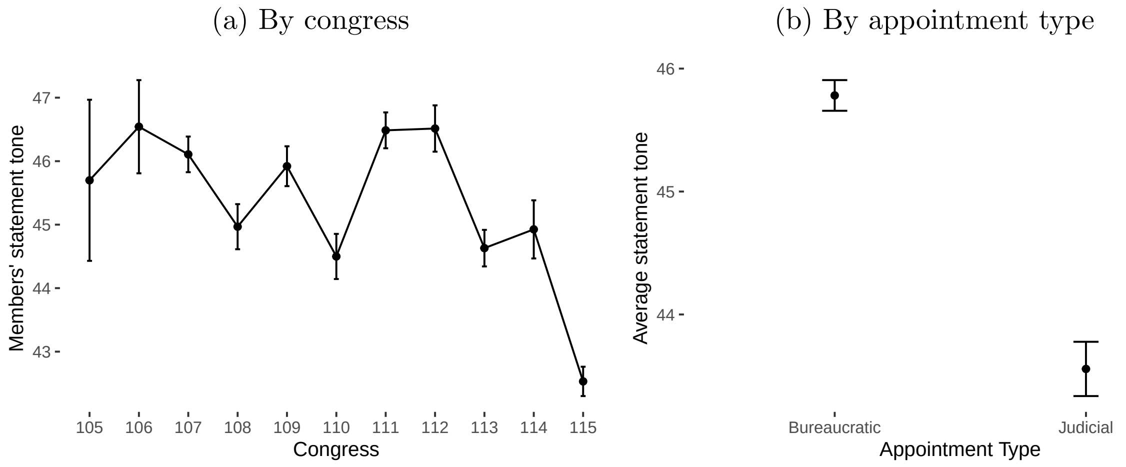
To facilitate the substantive interpretation of the changes in members' speaking tone, we rescaled the ensemble prediction scores to range from 0 to 100. It has a mean of 45.08 and a standard deviation of 15.58. Using this measure, we find support for most of our expectations. First, Figure 3a presents the average tone of the members in each congress with 95% confidence intervals. We can see a slight downward trend, especially in the most recent congress, which is the first two years of the Trump administration. However, the pattern is not as strong as expected.

[22] First, there are fewer seats for judicial nominations and they are lifetime appointments making them more consequential. Second, the president wields greater influence over bureaucratic agencies, as they are the head of the bureaucratic branch; the president's influence over the judiciary, the third branch, via nominations may convey greater political implications.

Table 3: Five random sample statements

Statement	Expert coding	Ensemble prediction
Madam soon-to-be Secretary, I would be very upset if you didn't disagree with the Secretary of the Interior who is leaving in some respects and on some issues. As a matter of fact, if you choose to be as mellow about the way you feel about some of his decisions, I might not vote for you. Who knows. I mean, you ought to honestly tell us that many of the things he has done and that he put on the books of this country are not exactly what George Bush for President wants, but you will comply with the law and hopefully you can make some changes.	1	-0.331
November of last year, she took her life. Her sister, Dana Lee, found her. And during testimony, Dana said that she felt Jami would still be alive had there been trained mental health professionals available near the Spirit Lake Reservation to diagnose the needs. But Jami didn't receive those services. Her death was tragic and unnecessary.	2	0.139
I appreciate that. The National Institute for Minority Health and Health Disparities funded a program in Maryland, in Baltimore, to show disparities, and that has been extremely helpful. And I would just encourage you to look at that institute as a real, valuable resource to you to carry out that commitment. The Affordable Care Act also increased dramatically the funding for Qualified Health Centers that allow access to care in minority communities. Are you committed to maintaining the support for Qualified Health Centers?	3	0.335
Madame Chair, if there's one thing I know about water, it takes a special kind of leader to reach consensus with such disparate groups and come back for more. This is the kind of leadership we need at Interior. It's the kind of leadership that our colleague Senator Salazar has provided there at Interior. It's the kind of leadership that Anne Castle offers.	4	0.426
Thank you, Chairman Dodd. Welcome, Mr. Chairman. We all know Chairman Bernanke's academic accomplishments prior to joining the Board of Governors, first as a member and then as its Chairman. He was and remains one of our Nation's leading scholars on the Great Depression. I believe that his expertise in this area has served him well during our current crisis.	5	1.088

Figure 3: Changes in members' tone


(a) By congress

(b) By appointment type

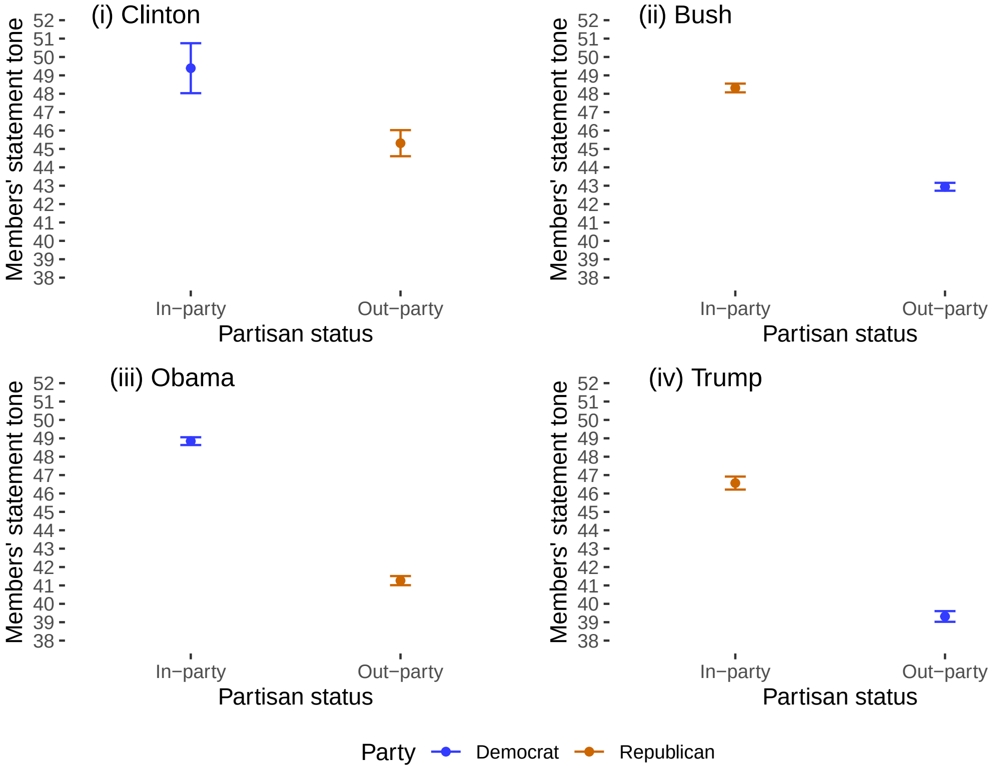
In graph (a), members' tone averaged by congress; in graph (b), the average tone of statements made by members who participated in both bureaucratic and judicial confirmation hearings at any point of time included in the current analysis measured for each appointment type. In both graphs, 95% confidence interval is around for each estimate.

Second, we analyze whether the same member tends to be tougher on judicial nominees than on bureaucratic nominees. We include only the statements made by those who participated in confirmation hearings for both types of appointments. This totals 79,158 statements made by 161 senators. Then, we compare the average tone of member statements made in bureaucratic nomination hearings and those from judicial nomination hearings. These statistics are presented in Figure 3b with 95% confidence intervals. As expected, senators tend to speak more negatively to judicial appointees (43.778) than they would to bureaucratic nominees (45.906).

Third, we calculate the average tone for Democrats and Republicans, respectively, for each presidential administration.[23] Figure 4 shows that senators are generally more negative toward those nominated by presidents from the opposite party. The partisan gap is visible in all four administrations. It is interesting to note that, first, the partisan gap grows over time, and second, the tones of both parties become negative from an administration to the next. Both findings are consistent with the intensified partisan conflict reported in the literature

[23] Independent senators were excluded from this analysis. These cases include Jim Jefferson from the 107th to 109th Congresses and Bernie Sanders from the 110th to 115th Congresses.

Figure 4: Tone by party and administration



Points are the average statement tone for Democrats and Republicans, respectively for four presidential administrations.

(Lee, 2016).

5 Alternative approaches

One potential question is whether taking these steps actually matters to the quality of the measure. In part, this question is tangential to our argument, in that our goal is not merely to build superior measures but to communicate the quality of the measure to readers both transparently and robustly. We wish to provide a framework that instills *confidence* in the procedure. We believe that doing so will also help scholars develop superior measures, but that is not the direct goal.

Nonetheless, it is informative to consider the consequences of following alternative strategies. To do this, in this section, we compare our measure to three alternatives: (i) an off-the-shelf dictionary method for measuring sentiment, (ii) a pre-trained BERT-based sentiment classifier, and (iv) a customized machine learning classifier built with our own data but in which we took various shortcuts.

First, one of the most preliminary approaches to measure sentiment is a dictionary method. Among many, we use the dictionary developed by Jockers (2017) for two reasons: first, it assigns a continuous score based on a word-embedding assisted dictionary, which facilitates the comparison to our measurement that is also continuous; second, this R package, `syuzhet`, which introduces this dictionary, is one of the most widely used sentiment analysis tool for R (Kim, 2022).

Second, as another alternative measurement, we use a pre-trained sentiment classifier. Unlike the dictionary method, it applies different meanings of the same word based on the context when assessing the sentiment of a text. `pysentimiento`, a Python package by Pérez et al. (2021), provides a convenient off-the-shelf application of a pre-trained BERT-based sentiment classifier.

Last, we construct another version of our measurement to demonstrate what happens when some of the steps that we proposed are bypassed. To simulate the case where the human-coded labels are not validated enough, we add random noise to our labels, which renders them inferior to our original labels.[24] In addition, we use only one simplest learner, support vector machine, without parameter tuning to simulate the case where the optimization and validation of learners were not done properly.

Figure 5 presents a correlation matrix comparing our measure and alternative measures proposed above to the five-point scale expert coding, which serves as a gold standard measurement of the latent trait in this study. All four continuous measurements are rescaled to run from 0 to 100 to facilitate comparisons. The panels on the diagonal show the distribution of each measurement. The panels below the diagonal show scatter plots for each pair of measurements with a fitted line; those above present the Pearson's correlation coefficient between each pair.

[24] Inferior labels may result from various reasons, such as unclear instructions to human coders, inattentive decisions made by human-coders, not enough human-coded data, etc. The noise was generated from a normal distribution of random numbers with zero mean and standard deviation of 0.05, which is the size that makes the noise large enough but barely changes the range of our labels.

Table 4: Face Validity of Alternative Measures

Expert Coding	Ours	Dictionary	Pre-trained	Simulation
1	38.719	23.696	33.470	30.984
2	47.968	23.870	7.598	31.217
3	53.562	32.387	99.103	34.638
4	54.957	25.319	97.554	33.217
5	73.041	28.216	99.555	43.464

The first row of the figure displays correlation coefficients between the expert coding and each of the sentiment measures. The correlation is 0.743 for our crowdsourced measurement; 0.399 for the dictionary-based measurement; 0.69 for the pre-trained BERT-based classifier; and 0.574 for the simulated measurement using our own data. These results show that our measure is most closely related to the expert coding implying that our customized machine learning classifier better captures the latent trait of interest than any alternative measurements.

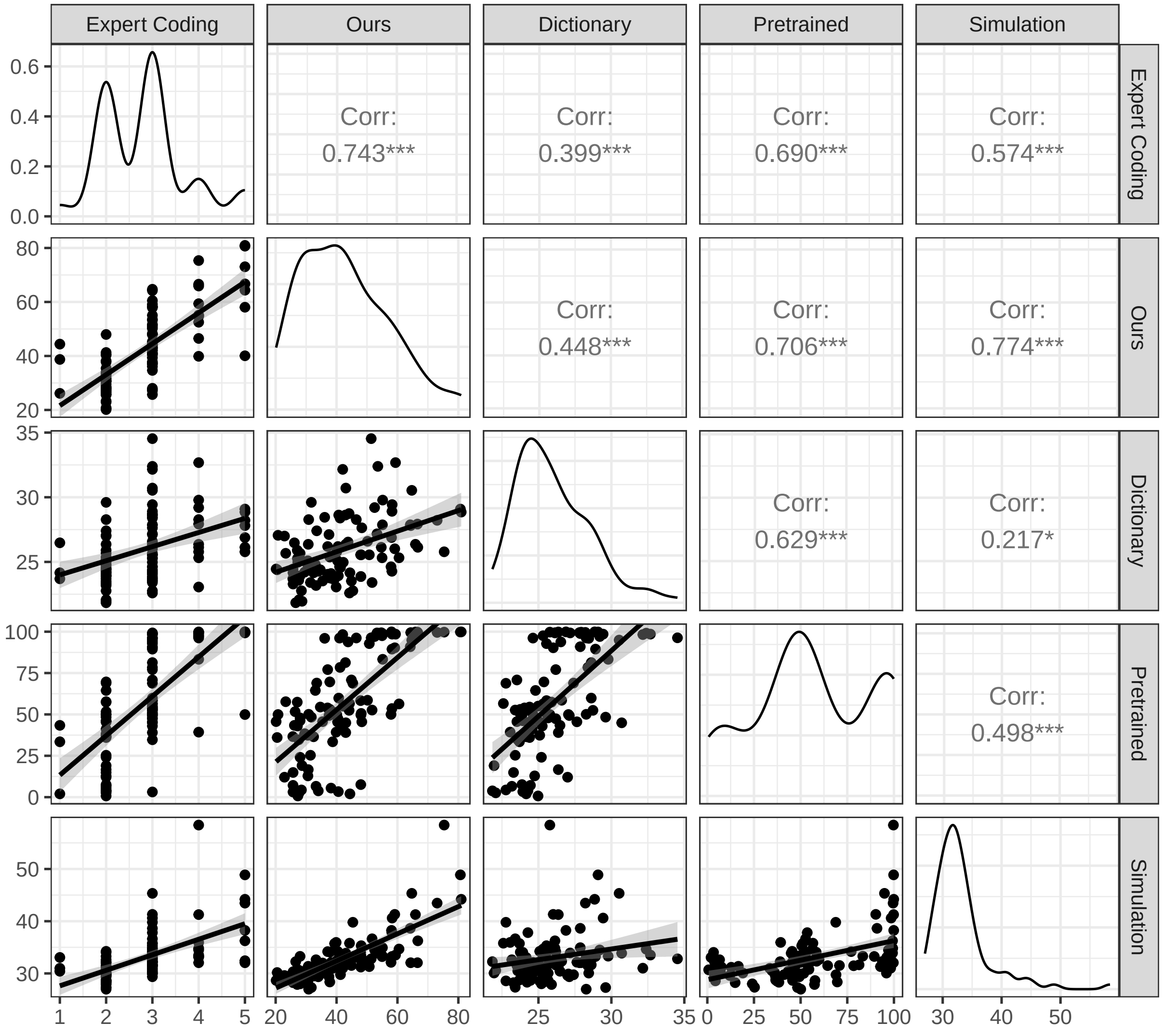
In addition, we compare face validity of these measurements. Table 4 extends Table 3 by including the predictions of the other three alternative measures for the five randomly selected paragraphs in the table. Here again, while our ensemble predictions are well aligned with the five-point scale expert coding, the predictions by the three other measures do not exactly match the ordering of the expert coding. Dictionary and simulation based measures predict the third paragraph more positively than the fourth paragraph, and the dictionary method predicts it is more positive even more than the fifth paragraph. The prediction of the pre-trained model has the most disconnection with the expert coding.

Overall, these experiments suggest that the measurement procedure we propose not only provides stronger validity of the measurement and more transparency of the measurement procedure but also can lead to a better measurement.

6 Conclusion

Social scientists are increasingly using supervised learning methods to study important concepts latent within large data sets of text, audio, video, and multimodal data. Most

Figure 5: Correlation among the sentiment measurements



frequently, the goal of these analyses is to construct a measurement that captures theoretically important concepts. Traditionally, checking the validity of a new measurement has been important in social science research. For example, survey researchers often extensively validate proposed measures, sometimes in stand-alone publications. However, in our review of recent articles using supervised learning, we find inconsistent standards in terms of the parts of the pipeline that are explained to readers and surprisingly little effort to validate the measure beyond within-sample prediction.

To address this concern, this article provides a framework for conceptualizing the process, highlighting points where important decisions might be disclosed to the reader, and multiple opportunities for validation. While not every measurement process needs to follow all of these steps, our hope is that this can serve as a point of departure for scholars developing their own measure using supervised machine learning.

For our example, we analyzed twenty-two years of US Senate confirmation or nomination hearing transcripts for federal offices. We measured the sentiment of senators' questioning tone on a continuous scale. We demonstrate the steps of our measurement framework, beginning with subsetting the data, followed by text preprocessing, labeling, transforming the data, fitting a model, and finally imputing measurements for the entire corpus. At each stage, we suggested several checkpoints focusing on the accuracy and validity of the measurement. This includes performing validations for the labels, the model, and the final output. We argue that the result is a more trustworthy and transparent measure that can be used to answer important substantive questions more credibly.

Our framework can be easily extended to analyze other types of data. Importantly, while our specific example involves a continuous measure, the general guidance we provide should work equally well for categorical outcomes. The main difference would simply be to choose alternative measures of predictive performance that are more appropriate to a categorical outcome. In addition, scholars are increasingly interested in non-text data, such as images, videos, audio, etc. Here too, researchers face important questions about how to choose subsets, acquire labels, preprocess inputs, train models, and validate the results. While the specific steps will differ for these research domains, we feel that the general underlying issues are the same and require careful thought and transparency from research teams.

Before concluding, it is worth considering how the above recommendations might apply with the advent of large language models (LLMs) that enable zero-shot and few-shot learning. These approaches promise to reduce or eliminate the need for task-specific labeled data, potentially streamlining the text-to-measure pipeline. Yet, many of our recommendations remain relevant. Researchers using LLMs should still consider (and report) the representativeness of their text corpus, the validity of their prompts or few-shot examples, and the need to validate the final measures. Moreover, the "black-box""nature of LLMs may necessitate even *more* rigorous validation to ensure that generated labels or measures align with the intended latent concepts. Alternatively, researchers may need to consider innovative methods to understand the textual features the models are responding to by, for example, experimentally manipulating inputs. Finally, for extremely large corpuses LLMs may still be too cost prohibitive, opening up opportunities for hybride approaches wherein LLMs are used to build training datsets for downstream supervised learning. In this case, nearly all of our above transparency recommendations would apply. Nonethless, as these methods become more prevalent in political science, developing frameworks tailored to their unique challenges will be crucial to maintain the field's commitment to transparency and empirical rigor.

References

Alrababa'h, A., W. Marble, S. Mousa, and A. A. Siegel (2021). Can exposure to celebrities reduce prejudice? the effect of mohamed salah on islamophobic behaviors and attitudes. *American Political Science Review 115*(4), 1111–1128.

Althaus, S., B. Peyton, and D. Shalmon (2022). A total error approach for validating event data. *American Behavioral Scientist 66(5).*

Anastasopoulos, L. J. and A. M. Bertelli (2020). Understanding delegation through machine learning: A method and application to the european union. *American Political Science Review 114*(1), 291–301.

Bøggild, T., L. Aarøe, and M. B. Petersen (2021). Citizens as complicits: Distrust in politicians and biased social dissemination of political information. *American Political Science Review 115*(1), 269–285.

Carlson, D. and J. M. Montgomery (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review 111*(4), 835–843.

Casas, A., M. J. Denny, and J. Wilkerson (2020). More effective than we thought: Accounting for legislative hitchhikers reveals a more inclusive and productive lawmaking process. *American Journal of Political Science 64*(1), 5–18.

Denny, M. J. and A. Spirling (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis 26*(2), 168–189.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1.*

Emeriau, M. (2023). Learning to be unbiased: Evidence from the french asylum office. *American Journal of Political Science 67*(4), 1117–1133.

Esberg, J. (2020). Censorship as reward: Evidence from pop culture censorship in chile. *American Political Science Review 114*(3), 821–836.

Fowler, E. F., M. M. Franz, G. J. Martin, Z. Peskowitz, and T. N. Ridout (2021). Political advertising online and offline. *American Political Science Review 115*(1), 130–149.

Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the uk house of commons, 1811–2015. *Political Analysis 27(4).*

Gohdes, A. R. (2020). Repression technology: Internet accessibility and state violence. *American Journal of Political Science 64*(3), 488–503.

Grimmer, J., M. E. Roberts, and B. M. Stewart (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science 24.*

Grimmer, J., M. E. Roberts, and B. M. Stewart (2022). *Text as data: A new framework for machine learning and the social sciences.* Princeton University Press.

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

Grimmer, J. and B. M. Stewart (2013b). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Groves, R. M. and L. Lyberg (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly 74*(5), 849–879.

Guess, A. M. (2021). (almost) everything in moderation: New evidence on americans' online media diets. *American Journal of Political Science 65*(4), 1007–1022.

Hager, A. and H. Hilbig (2020). Does public opinion affect political speech? *American Journal of Political Science 64*(4), 921–937.

Jockers, M. (2017). Introduction to the syuzhet package.

Kapoor, S., E. Cantrell, K. Peng, T. H. Pham, C. A. Bail, O. E. Gundersen, J. M. Hofman, J. Hullman, M. A. Lones, M. M. Malik, P. Nanayakkara, R. A. Poldrack, I. D. Raji, M. Roberts, M. J. Salganik, M. Serra-Garcia, B. M. Stewart, G. Vandewiele, and A. Narayanan (2022). reforms: Reporting standards for machine learning based science. *arXiv preprint arXiv:2201.12150*.

Kim, H. (2022). Sentiment analysis: Limits and progress of the syuzhet package and its lexicons. *Digital Humanities Quarterly 16(2).*

Lee, F. E. (2016). *Insecure Majorities: Congress and The Perpetual Campaign.* University of Chicago Press.

Malesky, E. J., J. D. Todd, and A. Tran (2023). Can elections motivate responsiveness in a single-party regime? experimental evidence from vietnam. *American Political Science Review 117*(2), 497–517.

McCarty, N. (2019). *Polarization: What Everyone Needs to Know.* Oxford University Press.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Mohr, F. and J. N. van Rijn (2022). Learning curves for decision making in supervised machine learning–a survey. *arXiv preprint arXiv:2201.12150*.

Montgomery, J. M., F. M. Hollenbach, and M. D. Ward (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis 20*(3), 271–291.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott (2022). How to train your stochastic parrot: Large language models for political texts. Technical report, Working Paper.

Park, B., K. Greene, and M. Colaresi (2020). Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects. *American Political Science Review 114*(3), 888–910.

Park, J. Y. (2021). When do politicians grandstand? measuring message politics in committee hearings. *The Journal of Politics 83*(1), 214–228.

Pérez, J. M., J. C. Giudici, and F. Luque (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv:2106.09462*.

Rodriguez, P. L. and A. Spirling (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics 84*(1), 101–115.

Schub, R. (2022). Informing the leader: Bureaucracies and international crises. *American Political Science Review 116*(4), 1460–1476.

Stier, S., F. Mangold, M. Scharkow, and J. Breuer (2022). Post post-broadcast democracy? news exposure in the age of online intermediaries. *American Political Science Review 116*(2), 768–774.

Wahman, M., N. Frantzeskakis, and T. M. Yildirim (2021). From thin to thick representation: How a female president shapes female parliamentary behavior. *American Political Science Review 115*(2), 360–378.

Ying, L., J. M. Montgomery, and B. M. Stewart (2021). Topics, concepts, and measurement: a crowdsourced procedure for validating topics as measures. *Political Analysis*, 1–20.

Zubek, R., A. Dasgupta, and D. Doyle (2021). Measuring the significance of policy outputs with positive unlabeled learning. *American Political Science Review 115*(1), 339–346.

Online Supplemental Information

For "Towards a framework for creating trustworthy measures with supervised machine learning"

Ju Yeon Park
The Ohio State University

Jacob M. Montgomery
Washington University in St. Louis

Contents

A Pre-processing texts

Table A1 summarizes the preprocessing steps used for each type of document matrix analyzed. We used different packages to preprocess documents for word2vec and for the other two. However, we made sure that each procedure was identical (e.g., the same list of stopwords) if it was applied to both. For the word2vec approach, we tested including only unigrams vs. both unigrams and bigrams as part of the grid search of the optimal model specification, and the model only with unigrams was chosen.

Table A1: Preprocessing steps for each of the three document matrices

Preprocessing steps	Word2vec	TDF	BERT
Remove punctuation	✓	✓	✓
Remove numbers	✓	✓	✓
Lowercase	✓	✓	✓
Stem	✓	✓	✓
Remove stopwords	✓	✓	✓
Include bigrams		✓	
Remove infrequent word stems		✓	

B Labeling procedure

This section provides further details about how we labeled paragraphs in the training set using Amazon Mechanical Turk online workers. When using crowdsourced workers, it is critical that the training and question construction carefully consider the latent trait. In this case, workers were given background information on the specific context as well as specific coding rules. These rules were developed iteratively based on preliminary crowdsourced results as well as our own hand coding of the documents discussed below. The training module included six “practice HITs” where coding choices were explained in detail with a specific focus on edge cases. Workers then completed a ten-HIT test to qualify for the tasks. We paid workers $0.08 per HIT for a total of 25,000 HITs and an approximate cost of $2,400 (including fees to Amazon). The complete coding instructions are as follows.

B.1 Coding instructions for MTurk workers

If you finish this training module with a passing score, you will be qualified to complete HITs posted by the requester SentimentIt with the title Compare Confirmation Hearing Statements.

Task: You will be presented two paragraphs from the senators’ statements during congressional confirmation hearings. Your task is to choose the paragraph that is relatively more positive or less negative.

To give you some background knowledge, the congressional Senate committees hold confirmation hearings to approve government nominees (e.g. the Secretary of State, Deputy Secretary, Attorney General, Circuit Judge, etc.). A confirmation hearing proceeds as follows: It starts with the committee chair’s and several other committee members’ opening statements followed by other congressmen’s statements recommending a nominee. Then, the nominee gives an opening statement. Finally, the chair proceeds to a Q&A session where committee members ask questions of a nominee. Long statements are broken down to paragraphs to facilitate this labeling task. Thus, some paragraphs you will compare can be part of a longer statement.

A statement is a positive statement if it does one of the following:

1. Speaking in support of a nominee (e.g., Praising her personality, qualification for the position, her career achievements, etc.)

2. Giving a positive description of a situation (e.g., Describing success of a policy implementation,

Examples

E.g.) Praising a nominee’s qualification for a position:

“She is smart, she is tough, she is hard-working and independent. She is a prosecutor’s prosecutor, and her qualifications are beyond reproach.”

“Under your leadership, the Enoch Pratt Free Library has flourished and serves as an indispensable beacon of higher learning and civic engagement for the City of Baltimore and the entire State of Maryland, and it is no surprise, given your four-decade career of success and exemplary work in the library sciences.”

E.g.) Giving a positive description of a situation:

“The people of North Carolina are very pleased about the results of the Base Realignment and Closure (BRAC) Commission, and the Army and Marine Corps “Grow the Force” initiative. Both Fort Bragg and Camp LeJeune are slated to receive a large influx of personnel. The Fort Bragg and Pope Air Force Base BRAC Regional Task Force are ultimately expecting total gains of about 40,000 military and civilian personnel in and around the city of Fayetteville. I think that those changes are ultimately going to be a great thing for the military and the State of North Carolina.”

A statement is a negative statement if it is one of the following:

1. Speaking against the nomination of a candidate (e.g., Disqualifying her previous job performance, misconduct, etc.)

2. Giving a negative description of a situation (e.g., Denouncing a policy and a related government agency, expressing concerns about a current policy-relevant situation, etc.)

3. Grilling a nominee to give her a hard time

E.g.) Disqualifying a nominee’s candidacy

“Mr. Holbrooke’s nomination was announced by the President on June 17, 1998, but it was not forwarded to the Foreign Relations Committee until February 10, 1999, almost 1 year later. The delay was caused by an 8-month-long criminal investigation of Mr. Holbrooke by the Department of Justice for alleged violations of U.S. ethics-in-government while he worked for Credit Suisse First Boston, and they always have initials after this, CSFB.”

E.g.) Denouncing a related government agency:

“The Department has also failed to hold another Government agency accountable: the Internal Revenue Service. We watched with dismay as that powerful agency was weaponized and turned against individual citizens. And why? What exactly did these fellow citizens do to make their Government target them? They had the courage to get engaged and speak out in defense of faith, freedom, and our Constitution.”

E.g.) Grilling a nominee to give him a hard time

"Your duty as Attorney General is not to defend the President and his policies; your duty is your oath, to defend the Constitution. So my first question, with that oath in mind, I ask you, do you believe that the President has the legal authority to unilaterally defer deportations in a blanket manner for millions of individuals in the country illegally and grant them permits and other benefits, regardless of what the U.S. Constitution or–immigrations laws say?"

"I would like to know how to reconcile these two statements. If what the President said was accurate, then why in the world would the FBI be conducting an ongoing criminal investigation? A rhetorical question: Would the FBI investigation be just for show? I would like–I am going to take Director Comey at his word. So if there is an ongoing criminal investigation at the FBI, then how could it be possible–be appropriate for the President to reach a conclusion about the facts before Director Comey?"

Warning: Some texts can be neutral so that hard to make a choice, but try your best. Here are some example statements where you might encounter this situation.

1. Procedural remarks

2. Describing a situation or explaining facts in a neutral tone (e.g. Nominee's career history)

3. None of these mentioned above

E.g.) Procedural remarks:

"Before I turn to our opening statements, I want to go over a couple of housekeeping items and explain how we are going to proceed. Senator Leahy and I will give our opening statements. Then I will call on Senators Schumer and Gillibrand to introduce the nominee."

E.g.) Nominee's career history in relatively neutral tone

"Ms. Lago currently serves as Assistant Secretary at the Treasury Department. There, she works to improve global market access for American goods and services. Prior to joining Treasury, Ms. Lago held a number of positions promoting economic development in State and local governments and in the private sector. She also served as the head of the Office of International Development at the SEC."

E.g.) Describing a situation in relatively neutral tone

"Our argument was, if we walk off the playing field, you bet, that invites a race to the bottom. But having the United States on the playing field is a force for driving standards up, and that is why your position is so important, and our priority is those strong, enforceable rules on labor and environment. In your position, you are going to play a key role in ensuring that each of these priorities is realized.”

E.g.) Explaining facts in neutral tone

“As we all know, the nominee, Dr. Carla Hayden, is the President’s nominee to be the 14th Librarian of Congress. Her successor, Dr. Billington, served ably for 28 years. Senator Schumer and I worked last year together to establish a term for this job, and so Dr. Hayden is the first person to be appointed for a specific term. That term would be 10 years.”

E.g.) None of these mentioned above

“My first question is just out of historical curiosity. Could you tell us a little about Enoch Pratt? His name is everywhere, the Enoch Pratt Library, and they said he was a merchant.”

Therefore, consider that statements can be placed onto a continuum of which one extreme end is positive statements and the other extreme end is negative statements. In the middle of the two ends, statements that are neither the two can be located. For each HIT, you will see two speech extracts. Your task is to read both and select which of the two statements is more positive or less negative. That is,

If statement A is...	If statement B is...	Then, choose
Positive	Negative	Statement A
Positive	Neutral or neither	Statement A
Neutral or neither	Negative	Statement A
Positive	Positive	The one that is relatively **more** positive
Negative	Negative	The one that is relatively **less** negative

It is important that you read each statement carefully, and that you judge each by the standards listed above and the information in the text. In comparing the two paragraphs, DO NOT make your judgments on your own knowledge of a person or a policy in question, on statements from previous HITs in this exercise, or on definitions of opinions different to those listed above. Note that the length of a statement is irrelevant to and does not cue the type of the statement.

Your choices will be evaluated after you complete each HIT. Low-quality workers, whose answers significantly deviate from other workers,’ may not be invited to participate in our future studies. Therefore, read each statement carefully. Skimming or reading quickly will result in low-quality evaluations. (Note: Training module which includes five practice questions and ten test questions are available here).

B.2 Labeling results

We originally planned each document to appear twenty times in the pairwise comparisons generating 35,000 HITs. However, we later decided to field only 25,000 HITs because we dropped a large number of HITs completed by workers who produced low quality work (see the next paragraph for this definition). Thus, each paragraph appeared less than twenty times but appeared with similar frequency in the 25,000 HITs.

Throughout the data collection process, we closely monitored worker-quality scores, which is produced as part of the output of a model that measures labels for each paragraph at the same time. The workers scoring less than 1 were considered to be producing low quality work (see Carlson and Montgomery, 2017). Then, we removed qualifications for workers whose score fell below the threshold of 1 for their worker-quality score. The distribution of worker scores are shown in Figure B1. In total, 182 workers participated in the coding and twenty-three were banned.

Figure B1: Distribution of the worker quality

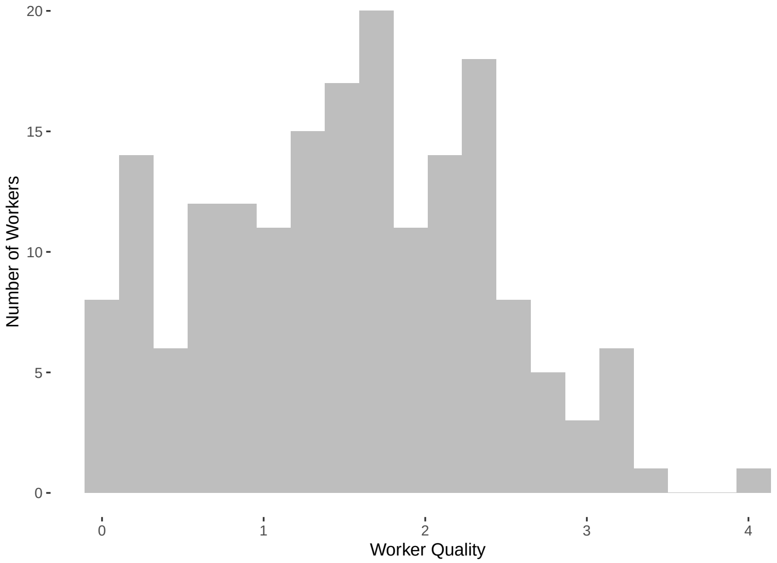


Figure B2 presents the distribution of the scores assigned to the 3,500 paragraphs that online workers labeled. In order to validate the crowdsourced labels, we classified a random sample of 100 paragraphs that were labeled into five-point scale categories: very negative, somewhat negative, neutral, somewhat positive, and very neutral. Figure B3 shows the correlation between the crowdsourced score and our five-point classification. The Pearson's correlation coefficient between the two is 0.808.

Figure B2: Distribution of the crowdsourced scores

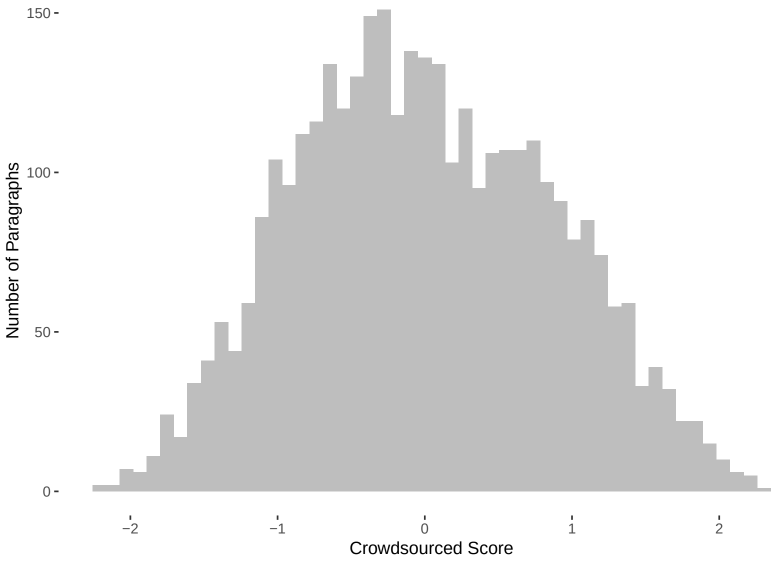
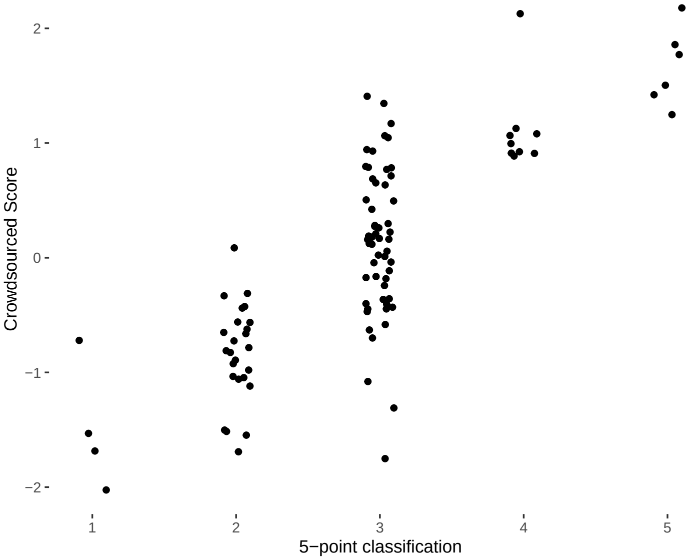


Figure B3: Correlation between the crowdsourced score and expert's five-point classification



The Pearson's correlation coefficient between the two measures is 0.821.

To provide a face validity of the crowdsourced labels, we offer ten paragraphs sampled from the 3,500 labeled paragraphs with their labels (See Table B2). We used random block sampling by selecting one paragraph from the paragraphs scoring below -2; two paragraphs ranging from -2 to -1; two from -1 to 0; two from 0 to 1; two from 1 to 2; and one from those scoring over 2.

Table B2: Ten Sample Paragraphs Labeled by Workers

Score	Information...	Paragraph
-2.03	Edward Kennedy; Nomination of James W. Holsinger, JR., M.D., PH.D.; Committee on Health, Education, Labor, and Pensions; 110th Congress	"Many concerns were raised by a paper that Dr. Holsinger wrote in 1991 on homosexuality for a study committee at the Methodist Church. Dr. Holsinger wrote this paper from his perspective as a medical doctor and he drew on his medical training to analyze the scientific studies he cited. Yet as I read it, it cherry-picks the science and is widely disputed scientifically. For example, I recently received a letter from nine doctors, highly respected in their fields, including one of the authors of the papers cited by Dr. Holsinger's paper of 1991, denouncing that paper as unscientific, biased and incredibly poor scholarship."
-1.29	Richard Durbin; Nomination of Charles W. Pickering, Sr. to be Circuit Judge for the Fifth Circuit; Committee on the Judiciary; 107th Congress	"Let me ask you to fast-forward to a more recent date because this is history; it goes back many, many years. And the year was 1994 and it involved a cross burning case which I am sure you expected to be questioned on. This was a case which was described to us as a very sad and tragic situation, as I read it."
-1.17	John Cornyn; Nomination of Janice R. Brown, of California, to be Circuit Judge for the District of Columbia; Committee on the Judiciary; 108th Congress	"Because of the clear terms of Proposition 209, the United States Supreme Court recently noted that in California racial preferences in admissions are prohibited by State law. Do Justice Brown's critics also disagree with Justice O'Connor who authored the opinion or Justices Stevens, Souter, Ginsburg and Breyer, who joined her? All Justice Brown did was her job. She authored a majority opinion for a unanimous Supreme Court, in forcing the clear terms of Proposition 209. Indeed, every single judge involved in the case at the trial court, the Court of Appeals, and the Supreme Court agreed with her. They agreed that the challenged San Jose program violated the will of the voters as expressed in Proposition 209."
-0.86	John Rockefeller; Nomination of Hon. Gary F. Locke to be Secretary of the U.S. Department of Commerce; Committee on Commerce, Science, and Transportation; 111th Congress	"So there are some who actually think that it won't happen. It won't finish on time, which will feed into my next question, the DTV. But I am interested in the level of your confidence and what your program is to make sure that the census is taken. There is a lot of conflict about the census, and it is a very harmful conflict to the fabric of our Nation. And so, it must be done properly, and it rests on your watch."

-0.18	Maria Cantwell; Anticipated Nomination of Steven Terner Mnuchin; Committee on Finance; 115th Congress	“Well, I think, to me, I think this election was a lot about the frustration of the American people on the implosion of our economy and the fact that they have not recovered. And I think that the President-elect, whether he directly or not meant to–I was pleased that his party adopted coming up with a very bright line separating commercial from investment banking.”
0.05	Charles Grassley; Confirmation Hearing on the Nomination of Hon. Loretta E. Lynch to Be Attorney General of the United States; Committee on the Judiciary; 114th Congress	“Before you seat yourself, would you take an oath, please? Would you raise your hand? And I will give the oath. Do you affirm that the testimony you are about to give before the Committee will be the truth, the whole truth, and nothing but the truth, so help you God?”
0.77	Ron Wyden; Nomination of Lieutenant General Paul M. Nakasone, U.S. Army, to Be Director of the National Security Agency and Chief of the Central Security Service; Committee on Intelligence; 115th Congress	“Thank you very much, Mr. Chairman. Mr. Chairman and colleagues, just a quick comment before we go to our nominee. The nomination of Gina Haspel to head the CIA comes at an especially momentous time. Senator Heinrich and I have asked that certain aspects of her background be declassified so that the American people can see what sort of person might head the agency at a particularly important time. I’ll just wrap up this point by saying I hope members will support what Senator Heinrich and I are calling for with respect to declassification.”
1.38	Barbera Boxer; The Nomination of Lieutenant General Robert L. Van Antwerp, Jr., to be Chief of Engineers and Commanding General of the United States Army Corps of Engineers; Committee on Environment and Public Works; 110th Congress	“I know. It is very exciting, isn’t it? I think the one thing that pulls the country together is people everywhere want the same thing. They want a good quality of life with their families, in good communities, solid, and not to have to worry about things that they really don’t have control over, which is where we come in.”

1.53	John Barrasso; Hearing on the Nominations of Michael Dourson, Matthew Leopold, David Ross, and William Wehrum to be Assistant Administrators of the Environmental Protection Agency, and Jeffery Baran to be a Member of the Nuclear Regulatory Commission; Committee on Environment and Public Works; 115th Congress	"Mr. Ross has a total of two decades of environmental, legal, and consulting experience in both the public and private sectors. His nomination has elicited bipartisan praise and support within my home State of Wyoming. Dave Freudenthal, the former Democrat Governor of Wyoming, has said "Mr. Ross's private practice experience in DC, combined with his service in two State environmental protection agencies, make him uniquely qualified to implement America's nuanced structure of Federal and State environmental protection.""
2.22	Orrin Hatch; Confirmation Hearing on the Nomination of John. P. Walters to be Director of the Office of National Drug Control Policy	"I would also like to praise General McCaffrey and his efforts as well. He worked very hard and did many good things. John Walters' career in public service has prepared him well for this office. Like you, Mr. Chairman, he has worked tirelessly over the last two decades helping to formulate and improve comprehensive policies designed to keep drugs away from our children. Also like you, he has truly unparalleled knowledge and experience in all facets of drug control policy."

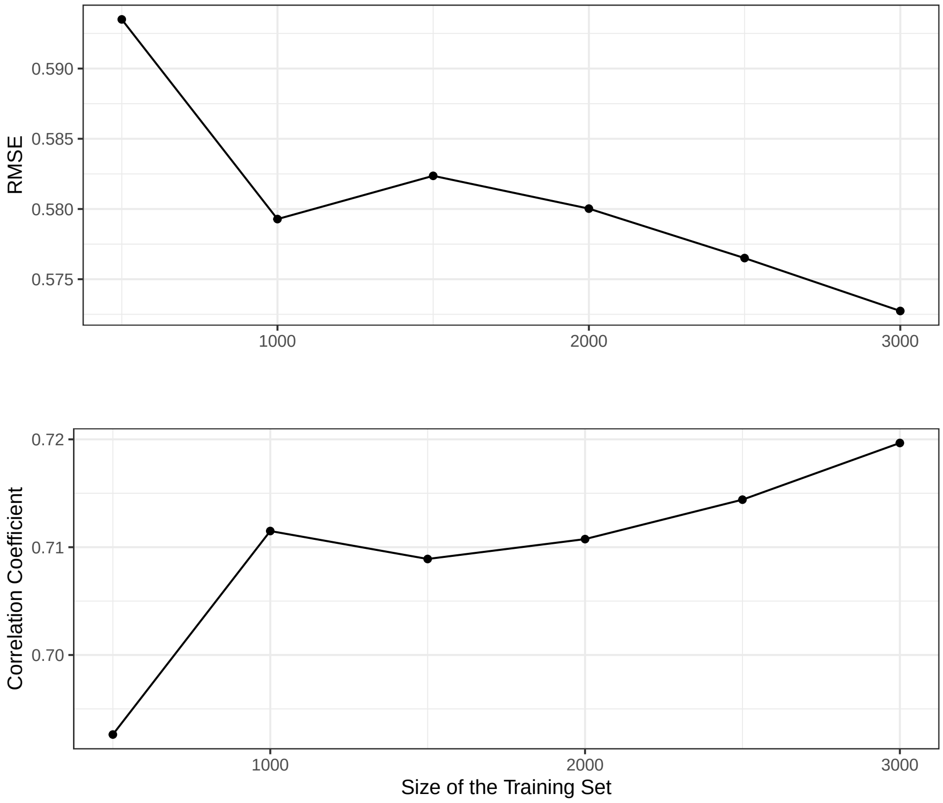
B.3 Determining the Size of the Training Set

What is the right size of training data? There is no single answer to this question. Various elements related to one's research question will factor into this question, such as the complexity of the latent trait one wants to measure, the labeling scheme - whether it is a binary, categorical, or continuous label, the length of the text units labeled, and so on. However, using our concept of measurement, this section aims to demonstrate how varying the size of the training data can change the predictive performance of a machine learning model. While our analysis used 3,000 labeled paragraphs to train various machine learning models, here we reduce the size of the training data to 500 by randomly selecting these cases from Groups A and B, and increase it by 500 to 2,500. In this simulation, we train Kernlab's Support Vector Machine (KSVM), which is one of the machine learning models we use in this paper, and predict the 500 labeled paragraphs in Group C to cross-validate the model.

Figure B4 presents the Root Mean Squared Errors (RMSE) and the correlation coefficients comparing the human-coded labels and the model predictions for these 500 paragraphs. The result shows that both validation metrics improve significantly as the training set increases from 500 to 1,000, and gradually improve as the size increases.[25] Thus, the larger the training set, the better the learners are at predicting labels, but

[25] Even though we are using only a portion of our training set, note that the labels are still based on

Figure B4: Varying the Size of the Training Set



the marginal benefit of increasing the training set may diminish as one increases the size of the training set. While these metrics are specific to our data and measurement strategy, the general implication would apply broadly to other research designs.

the pairwise comparison of all 3,500 paragraphs and are thus more precise than the labels we would have obtained from the data of a smaller size. For this reason, the actual changes in the model prediction would have been more dramatic than what is shown in this section.

C Learner selection and their prediction performance

First, we fit a BERT model predicting our continuous label. In addition, for embeddings and the TDF, we fit multiple models suitable for a continuous response variable.[26] The models we use are Support Vector Machine (SVM), Kernlab's Support Vector Machine (KSVM), Lasso, Random Forest, Bayesian Generalized Linear Model (Bayes GLM), and Gradient Boosting Machine (GBM). These models were fit using the `SuperLearner` R package (van der Laan et al., 2007).[27]

We varied key tuning parameters for each model and chose the setting that performed best based on a ten-fold cross validation within the training sample.[28] The set of parameters that we tried for each learner is presented in Table C3.

Table C3: Parameters for machine learning algorithms

Model	Tuning parameter	Values
SVM	cost	2^{-5}, 2^{-3}, 1, 2^{3}, 2^{5}
Kernlab's SVM	epsilon	0.001, 0.01, 0.1, 0.5, 1
	cost	1, 5, 10
Lasso	nlambda	200
Random Forest	num.trees	100, 500, 1000
	mtry	0.5, 1, and 2 times of $floor(ncol(x_train)/3$
	min.node.size	5, 10
GBM	max.depth	3, 6
	nrounds	100, 200, 300
	eta	0.1, 0.01

Note: This table presents various values of tuning parameters we considered for each learner. We chose the best combination of parameter values for each model based on the RSME from out-of-sample predictions. For Bayes GLM, we used the default parameters for the model as defined in the `SuperLearner` R package.

Through an ensemble Bayesian model averaging technique using the `EBMAforecast` R package (Montgomery et al., 2012), the models that received non-zero weights are as follows: BERT, SVM, KSVM and GBM models using the doc2vec matrix; and the GBM model using the TDF approach. The weights that the ensemble model assigned to each model as well as their fit statistics are shown in Table C4.

[26] For embeddings and the TDF, we removed stopwords. The stopwords used in this study include the default English stopwords from the `quanteda` R package. In addition, we included the following stopwords to the list: Im, youre, hes, shes, its, were, theyre, ive, youve, weve, theyve, id, youd, hed, shed, wed, theyd, ill, youll, hell, shell, well, theyll, isnt, arent, wasnt, werent, hasnt, havent, hadnt, doesnt, dont, didnt, wont, wouldnt, shant, shouldnt, cant, cannot, couldnt, mustnt, lets, thats, whos, whats, heres, theres, whens, wheres, whys, hows, aint, and ain't.

[27] For Random Forest, we used the "ranger" model, which is a fast implementation of Random Forest (See Breiman, L. (2001). Random forests. Machine learning 45:5-32.). For Gradient Boosting Machine (GBM, we used the "xgboost" model which is a variant of GBM. (See **`https://xgboost.readthedocs.io/en/latest/index.html`** for more information.)

[28] The folds were set up in advance and the same partitioning strategy used across all learning models to facilitate more accurate comparisons.

Table C4: Machine learning algorithms

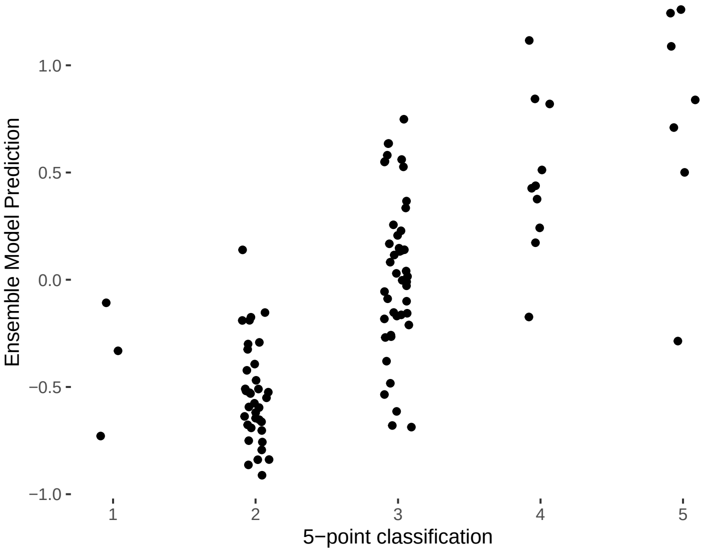
Doc. Matrix	Model	Tuning parameter	RMSE	Correlation	Weight
BERT	Dropout and Dense	epoch=20	0.606	0.721	0.433
Word2vec	SVM	Cost=1	0.607	0.719	0.050
Word2vec	Kernlab's SVM	Epsilon=0.5, C=1	0.604	0.722	0.324
Word2vec	Lasso	Number of lambdas=200	0.611	0.714	0.001
Word2vec	Random Forest	Number of trees=1000 Minimum node size=5	0.649	0.695	0
Word2vec	Bayesian GLM		0.637	0.691	0
Word2vec	GBM	Maximum depth=5 Shrinkage=0.01 Number of trees=2000	0.623	0.701	0.057
TDF	SVM	Cost=32	0.654	0.662	0
TDF	Kernlab's SVM	Epsilon=0.1, C=1	0.642	0.678	0
TDF	Lasso	Number of lambdas=200	0.664	0.652	0
TDF	Random Forest	Number of trees=1000 Minimum node size=5	0.679	0.674	0
TDF	Bayesian GLM		1.024	0.463	0
TDF	GBM	Maximum depth=5 Shrinkage=0.01 Number of trees=3000	0.646	0.673	0.135

The best set of tuning parameters were chosen from multiple combinations for each model through a grid search. RMSE and Pearson's correlation coefficients were calculated with a cross validation within our 3,000 document training set. Weights were the assigned to each model from the ensemble Bayesian model averaging procedure, and only six models received non-zero weights and contributed to the final model predictions.

Table C5: Most frequent and exclusive terms for positive (top quartile) vs. negative (bottom quartile) documents

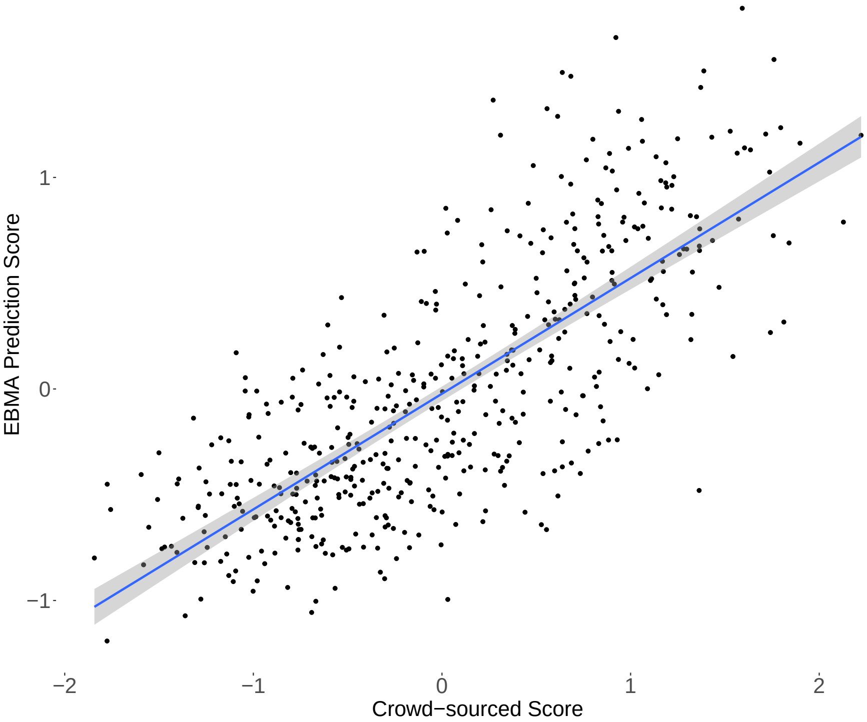
Positive	Negative
serv, famili, thanks-much, forward, welcom, appreci, thank-mr, introduc, look-forward, experi, pleas, ms, district, assit, colleagu, opportun, open, challeng, univers, dr, manag, school, director, join, career, friend, communiti, thank-senat, distinguish, develop, bring, morn, strong, willing, ambassador, leadership, honor, testimoni, energi, board, want-thank, congratul, repres, proud, associ, next, governor, econom, chief, three	whether, view, decis, report, problem, agre, rule, mean, chang, someth, suprem, inform, suprem-court, $, constitut, happen, polit, percent, clear, seem, reason, author, requir, find, opinion, actual, anoth, money, involv, matter, power, correct, rais, regard, action, regul, effect, fair, standard, major, still, enforc, quot, exampl, specif, fund, increase, review, compani

Figure C5: Correlation between the ensemble prediction and five-point expert classification



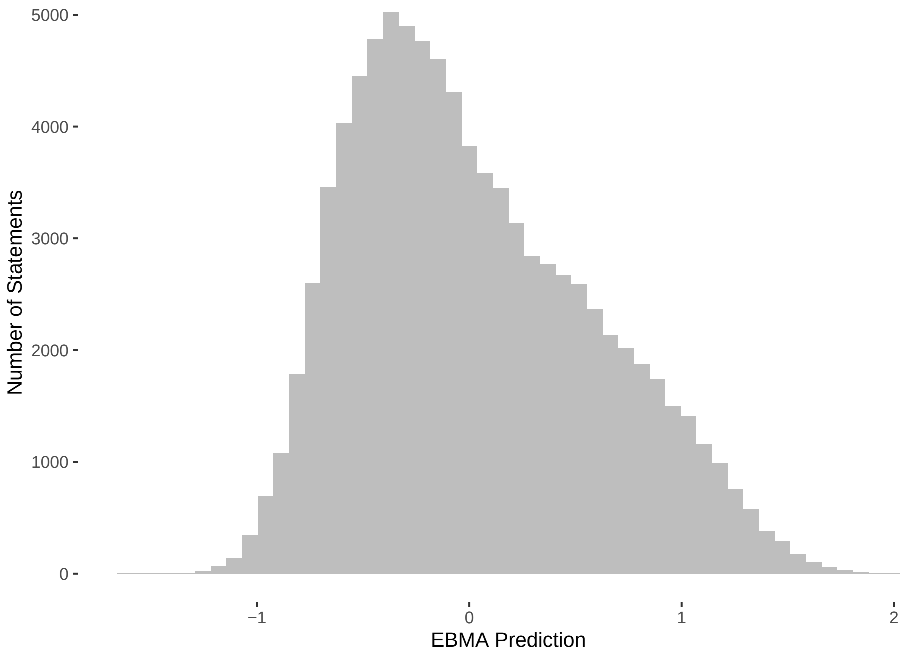
Comparison of scores predicted from the ensemble to expert coding (1-5 scale) of 100 documents held out from model labeling and training. Pearsons's correlation is 0.747.

Figure C6: Correlation between the crowdsourced score and ensemble prediction



Predicted versus observed crowdsourced labels for the validation set (n=500) held out during model building and ensemble calibration. RMSE for this sample is 0.538 and Pearson's correlation is 0.756.

Figure C7: Distribution of the EBMA prediction



References

Carlson, D. and J. M. Montgomery (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review 111*(4), 835–843.

Montgomery, J. M., F. M. Hollenbach, and M. D. Ward (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis 20*(3), 271–291.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical Applications of Genetics and Molecular Biology 6(25).*