Human computation scaling for measuring meaningful latent traits in political texts [*]

Jacob M. Montgomery
Washington University in St. Louis

David Carlson
Washington University in St. Louis

October 15, 2016

ABSTRACT

Scholars are increasingly interested in measuring latent political concepts embedded in written or spoken records. After all, most important political behaviors and outcomes are encoded in language. However, current approaches of turning natural language into meaningful measures are sometimes unsatisfying, relying on either costly and unreliable human coding or automated methods for document classification that miss subtleties of language easily identified by human readers. In this paper, we develop and validate an innovative "human computation" method for encoding political texts that preserves much of the reliability of automated methods while leveraging the superior ability of humans to read and understand natural language. We validate the method with online movie reviews, open-ended survey responses, advertisements for U.S. Senate candidates, and State Department reports on human rights. The framework we present is quite general, and we provide software to help researchers interact easily with online workforces to extract meaningful measures from texts.

[*] We thank Burt Monroe, John Freeman, and Brandon Stewart for providing comments on a previous version of this paper. We are indebted to Ryden Butler, Dominic Jarkey, Jon Rogowski, Erin Rossiter, and Michelle Torres for their assistance with this project. We particularly wish to thank Matt Dickenson for his programming assistance. We also appreciate the assistance in the R package development from David Flasterstein, Joseph Ludmir, and Taishi Muraoka. We are grateful for the financial support provided by the Weidenbaum Center on the Economy, Government, and Public Policy. Finally, we wish to thank the partner-workers at Amazon's Mechanical Turk who make this research possible.

1 INTRODUCTION

Given the centrality of words to the political process, it is unsurprising that political scientists have always been deeply interested in studying and characterizing the content of spoken and written language. These records are, after all, often the most direct evidence we have about the true nature of political debates, the intentions of political actors, and the policy outcomes reached by political institutions. Indeed, political science in earlier eras rested heavily on the analysis of language originating from interviews (Key 1949; Kingdon 1973), participant observation (Fenno 1978), and more. In fact, in several areas of inquiry, a heavy focus on language persists. Projects such as the Comparative Manifesto Project (Budge 2001) and the Policy Agendas Projects (Baumgartner and Jones 1993) represent massive endeavors to characterize and explain the nature of important political texts and speeches.

Yet, considering the broad contours of standard practices in social science research over the past six decades, the systematic study of natural language has declined precipitously, largely in conjunction with the rise of statistically oriented forms of inquiry. Today, political scientists seeking to test theories largely ignore what is *said or written* and instead focus on more easily quantifiable *behaviors* (e.g., roll-call votes). While political science has advanced significantly relying on behavioral indicators, they do not encapsulate the wholeness of our political realities. On the contrary, few students of politics would deny that the discipline's near-exclusive reliance on easily quantifiable metrics impoverishes its engagement with the realities of politics in practice. It is as if Lincoln's ascendancy to the presidency can be captured by the mere fact of his inauguration rather than his appeal to the "better angels of our nature," the outcome of *Brown v. The Topeka Board of Education* is fully encapsulated as a vote total rather than an announcement that "Separate educational facilities are inherently unequal," or that Senator Obama's loss in the 2008 New Hampshire presidential primary can be summarized in the outcome rather than his proclamation that "Yes we can."

While the subsidiarity of language to other measures in recent decades likely has many explanations, certainly one contributing factor is the difficulty of turning naturally generated human language into quantitative measures that retain their substantive and theoretical value. Broadly speaking, researchers analyzing the content of written or spoken language have either used expert coders (e.g., trained research assistants) or one of several text-analysis algorithms derived from computer science (e.g., topic models). However, both expert coding and machine learning algorithms run afoul of the multi-valued and interpretative nature of human communication. The meaning of words is often subjective and interpretable only within a specific context, particularly in the political realm where the very definitions of terms are points of conflict and debate (e.g., "marriage"). So, for instance, the claim that humans have a "right to life" indicates freedom from state-sponsored violence in the context of Article 3 of the *Universal Declaration of Human Rights*, but is a justification for the violent annexation of territory in the context of 1920s German politics.[1] The result of the inherent contextual and subjective qualities of natural language means that quantifying the underlying meaning of any text is to some degree interpretive.

Why is the interpretive nature of natural language an obstacle to generating meaningful quantitative measures? It is extremely difficult to turn language into measures that are simultaneously reliable and valid indicators of important political concepts. Human coders are able to easily read a sentence and assign political meaning, but subtle differences in the underlying interpretations of a text lead to low levels of inter-coder reliability. Rigid coding rules focused on more objective aspects of texts are sometimes able to improve reliability, but at the cost of focusing on features that are intentionally less interpretive and, as a consequence, of less direct interest.

Recently, many scholars have turned to the analysis of political texts with automated techniques coming from computer science. Despite their promise, however, the outputs of these models may not reflect the underlying concepts of interest to researchers and can be difficult to interpret. As Grimmer and Stewart (2013, p. 271) note, "Automated text analysis methods can substantially reduce the costs and time of analyzing massive collections of political texts. When applied to any one problem, however, the output of the models may be misleading or simply wrong." Moreover, relatively little work to date has focused on measuring continuous latent traits. While there are several important exceptions (e.g., Laver, Benoit, and Garry 2003; Slapin and Proksch 2008), many prominent studies have instead focused on classifying documents into unordered categories (e.g., Quinn et al. 2010; Grimmer 2010).

In this paper, we draw on insights from the field of human computation to propose a method that combines the advantages of both the traditional content-analysis and automated approaches for turning language into data. We develop and validate a general-purpose framework for encoding natural language by dividing the larger undertaking into thousands of simple micro-tasks (viz., binary pairwise comparisons) that can be easily completed by a trained but non-expert online workforce. By sending thousands of these simple comparisons to online workers, we circumvent issues of unreliable human coding while capitalizing on the superior abilities of humans to understand context and assign meaning to language. We then statistically post-process the resulting data to construct valid and reliable estimates of latent traits within documents. Importantly, the output of our method is not a categorization, but a meaningful measure scored on a continuous scale.

After providing the details of our method, we evaluate it using texts from online movie reviews, open-ended survey responses, congressional advertisements, and State Department reports. In each case, we compare our estimates to existing measures of researcher-specified latent traits embedded within the texts to show that our measures are not only highly reliable, but also valid measures of the underlying latent traits of interest. The framework we present is quite general, and we provide

[1] "The first right in the world is the right to life, provided one has the strength for it" (Hitler 2013, p. 18).

software and practical guidelines to help researchers smoothly interact with online workforces.

2 THE CHALLENGE OF ENCODING NATURAL LANGUAGE

Until very recently, the most common approach to transforming language into data was to conduct some form of content analysis (e.g., Krippendorff 2013). This involves, first, dividing some text into units (e.g., paragraphs) to be analyzed, and, second, placing each unit into a category based on a coding rubric. While intuitive and conceptually straightforward, this approach suffers from two well-known weaknesses. First, it can be prohibitively expensive even for modestly sized collections of documents. Coding documents is often a tedious task, but one that demands a certain level of training and expertise to perform correctly. Finding, training, and compensating research assistants can be time-consuming and costly.

For instance, the most comprehensive effort to date that has assembled information on campaign strategy is provided by Druckman, Kifer, and Parkin (2009), who collected information from the websites of candidates for the Congress in the 10 days preceding the 2002, 2004, and 2006 general elections. But this effort itself reveals the arduous nature of this task using traditional methods. The authors are able to code *only* major-party Senate candidates and a random sample of 20% of House candidates.

Second, and more problematic, past experience shows that even highly trained and competent coders provide unreliable estimates when asked to code even modestly subjective topics. Mikhaylov, Laver, and Benoit (2012), for instance, investigate the inter-coder reliability of categorizations produced by expert coding of party manifestos. They find that the existing coding process "is prone to unacceptably high levels of unreliability " (p. 90). They conclude that "the propensity. . . for misclassification by human coders, even trained and experienced coders, suggests a need for a much simplified coding scheme" (p. 90).

This trade-off between the subtlety of the coding scheme and the reliability of the measure is well known. We might, for instance, wish to code the 'tone' of statements on candidate websites on a 100-point scale ranging from strongly negative to strongly positive. However, implementing such a scheme is simply beyond the capacity of human coders to execute reliably and is flatly impossible when multiple coders are involved (Krosnick 1999; Oishi et al. 2005). Thus, Druckman, Kifer, and Parkin (2009) coded candidate websites on characteristics easily identifiable to individual research assistants, such as whether the website provides a candidate biography, mentions her family, mentions her opponent, or includes polling results. In the end, despite requiring many hours of human labor to build, the dataset contains nothing about the tone, ideological content, or even the policy implications of candidate messages.

Given the expense and difficulty of manually coding language, it is unsurprising that political scientists have embraced advances in computer science that allow for automated coding. Recent work drawing on this family of methods has been used to study congressional speech (Monroe, Colaresi, and Quinn 2008), categorize open-ended survey responses (Roberts et al. 2014), understand the representational style of elected officials (Grimmer 2013), and more. Despite the obvious advantages of automated methods relative to relying on trained research assistants in terms of reliability and cost, it is important to understand their fundamental limitations. First, as most practitioners of text analysis acknowledge, the outputs of these models are dramatic simplifications of the underlying language. As Grimmer and Stewart (2013) note, "The complexity of language implies that all methods necessarily fail to provide an accurate account of the data-generating process used to produce texts" (p. 270). Virtually all text models applied in a political context are variants of "bag of words" methods that strip language of not only context but even word ordering, punctuation, and tense. This does not mean that text models are without value – far from it – but rather that they are not appropriate for capturing all aspects of language that may be of interest to researchers. More centrally, this reductive approach means that text models struggle to differentiate between statements that are trivial for even untrained human coders to distinguish.

A second limitation of many of the dominant text models in the literature is that they are focused on partitioning documents into distinct clusters or grouping (e.g., Hopkins and King 2010). While classification is certainly valuable for answering some questions, researchers are often more interested in scaling documents to extract measures of researcher-specified underlying traits.[2]

Scaling continuous latent traits in text is certainly not unknown in political science. The most prominent example is the `WordScore` model proposed by Laver, Benoit, and Garry (2003) for placing party manifestos on an ideological scale. Other notable examples include the `WordFish` model (Slapin and Proksch 2008) and dictionary-based methods (e.g., Owens and Wedeking 2012).[3] However, we argue that the general applicability of these methods across domains is limited. To begin with, there are concerns that – at least in some cases – these methods provide low-quality estimates. For example, Lowe and Benoit (2013) benchmark the `WordFish` method for scaling ideology in texts against expert human coders and found that in many cases the statistical methods were wildly inaccurate. Grimmer and Stewart (2013, p. 293) further show that the underlying meaning of `WordFish` scores changes radically depending on the content of the document set. Likewise, Budge and Pennings (2007) use `WordScores` to code speeches in the Irish Parliament to measure party support for the budget. While the automated methods placed Sinn Féin in the middle of the spectrum, all human coders were able to easily place them at the political extreme.

[2] While it is possible to partially equate unsupervised topic models with scaling models (e.g., Pang and Lee 2005), the process is likely to be frustrating and largely unproductive (Grimmer and Stewart 2013, p. 281).

[3] Other examples might include Lowe et al. (2011) or Jamal et al. (2015), depending on how one interprets the output. Another branch of recent research seeks to combine texts with ancillary information to provide unsupervised sentiment-scaling of speeches. Several recent papers, for instance, combine text with roll call votes to measure ideology (e.g., Lauderdale and Herzog 2014; Kim, Londregan, and Ratkovic 2014).

Further, a fundamental limitation of both supervised learning and dictionary-based approaches is that they assume the existence of a well-validated document set, something rarely available in political science. Supervised methods "learn" how specific word frequencies are associated with an underlying trait by estimating the relationship between words and measured traits in training datasets. Laver, Benoit, and Garry (2003), for instance, train their model using a set of expert-coded party manifestos. In the end, the validity of the resulting measure rests entirely on the quality of the training set. This brings us full circle to the difficulty of using human experts to reliably code complex documents or build dictionaries. Likewise, dictionary-based methods assume that the meaning (or valence) of a specific word is consistent with its assigned valence in the dictionary, which is itself developed within a specific research domain. However, "when dictionaries are created in one substantive area and then applied to another, serious errors can occur" (Grimmer and Stewart 2013, p. 274).

The method we propose relies neither on statistical methods nor human coding alone. The intuition is that we can combine the superior ability of humans to read and understand the meaning of natural language with the superior ability of computers to aggregate data into reliable measures of latent traits. As we demonstrate, this combination allows us to produce valid measures of latent concepts embedded in texts that better reflect the complexity of human communication. Further, the traits of interest can be specified in advance by the researcher, estimates exist on a continuous scale, and the measures are highly reliable.

Before presenting our method, it is important to note that the idea of using online workforces to code text is not unknown to political science. Henderson (2015), for example, uses online workers to guess the ideological origins of political ads. Honaker et al. (2013) outlines a method that is quite similar in basic structure to our own in that they rely on pairwise comparisons of statements. However, by far the most comprehensive study using online workers to encode political texts is Benoit et al. (Forthcoming), which presents a method for coding party manifestos.

The `SentimentIt` system we present below differs in that our aim is not to provide a method for encoding a specific corpus of texts (e.g., party manifestos), but rather to provide a general framework for creating reliable and valid measures of latent traits for a wide array of document sets. We show that our framework can be applied in areas ranging from online posts, to survey responses, to political advertisements, to reports from bureaucratic agencies. Further, our approach to quality control differs significantly in that we do not rely on "gold standard" tasks, which require researchers to regularly pay for coding tasks that are of no direct use. Instead, our method relies on a pairwise-comparisons framework that allows for seamless supervision of workers' outputs in the very process of collecting the data. Finally, the `SentimentIt` software we provide offers a suite of tools for researchers to smoothly set up tasks, manage workers, and analyze data, all within the increasingly common `R` computing environment.

3 THE SENTIMENTIT SYSTEM FOR HUMAN COMPUTATION TASKS

Human computation (HC) is the study of algorithms and procedures that integrate the innate abilities of humans with the developing capabilities of computer algorithms to together solve problems that neither humans nor computers can handle alone (Quinn and Bederson 2011). HC as a field was pioneered by Luis von Ahn, who, in summarizing his research, states:

> Although computers have advanced dramatically over the last 50 years, they still do not possess basic conceptual intelligence or perceptual capabilities that most humans take for granted. By leveraging human abilities ... I solve large-scale computational problems and collect data to teach computers basic human talents. To this end, I treat human brains as processors in a distributed system, each performing a small part of a massive computation (Von Ahn 2009, p. 418).

While a wide variety of methods fit under this umbrella,[4] HC methods have three basic components (Quinn and Bederson 2011). First, the researcher must create and organize a corpus of specific texts or images that need to be analyzed. Second, HC makes use of large online workforces to perform small evaluative tasks. For example, Von Ahn et al. (2008) used the `ReCAPTCHA` security protocol to have humans recognize words in non-digitized texts that computers could not confidently identify. Finally, human evaluations are aggregated in some fashion to create an output useful for the end-user. Typically, this is done using statistical post-processing, redundancy, or other methods that ensure that some degree of human error is removed.

3.1. *Design principles*

Following the principles laid out in Quinn and Bederson (2011), we designed our system, which we label `SentimentIt`, based on the following criteria. First, `SentimentIt` leverages the *human ability* to understand language and socially constructed political concepts. Importantly, while all of our tasks are to some degree subjective, we focus on the human ability to discern pre-defined characteristics embedded within text (e.g., positivity) rather than explicitly subjective characteristics (e.g., persuasiveness). That is, we endeavor to focus on characteristics that can be defined clearly and are therefore less subject to coder-specific biases.

Second, we designed the *task structure* to be cognitively appropriate for non-experts. Specifically, we ask workers to conduct pairwise comparisons of texts, simply indicating which text is more extreme along a single dimension of interest (e.g., “Which text is more positive?”). A significant body of research indicates that pairwise comparisons can reduce the cognitive burden for respondents, improve the reliability of responses, and eliminate problems such as differential item functioning and reference group effects that plague alternative question formats (Brady 1985;

[4] See Quinn and Bederson (2011) for further examples and a detailed discussion distinguishing HC from related concepts such as crowdsourcing and social computing.

King et al. 2004; Oishi et al. 2005). Note that this distinguishes SentimentIt from previous approaches for using online work forces to analyze text, which rely exclusively on Likert-format questions where texts are presented sequentially in isolation (e.g. Benoit et al. Forthcoming). In addition to the results below, we provide evidence that workers are able to reliably complete pairwise comparison tasks measuring a unidimensional concept in Appendix SI-5.

A third design principle is the *motivation* of workers. In this case, we relied on paid online workers recruited through Amazon's Mechanical Turk (AMT). In our examples below, we pay between \$0.04 and \$0.10 for each pairwise comparison.[5] While it is now common for researchers to use AMT workers as research subjects (Berinsky, Huber, and Lenz 2012), the majority of jobs posted at AMT are actually completing HITs (human intelligence tasks). AMT workers are generally well educated, part-time workers who are highly experienced at completing micro-tasks for small amounts of money – usually less than \$0.10 per task (Hitlin 2016). As we demonstrate below and in our online appendices, the result is a workforce capable of quickly providing high-quality data at a very low cost.

Our final design principle is *quality assurance* via training, redundancy, and statistical monitoring. To begin, in order to qualify for our micro-tasks, workers must complete an online training module that explains the task, provides detailed decision rules, and includes example HITs with discussion of difficult cases. As part of this training, workers must correctly complete a preset number of tasks before they are "certified" to participate.[6] Our training modules were built in the Qualtrics survey software, which can interact with SentimentIt via application program interface (API) calls. Details for setting up this interface are provided in Appendix SI-7. Evidence on the effect of training on worker quality is provided in Appendix SI-8.

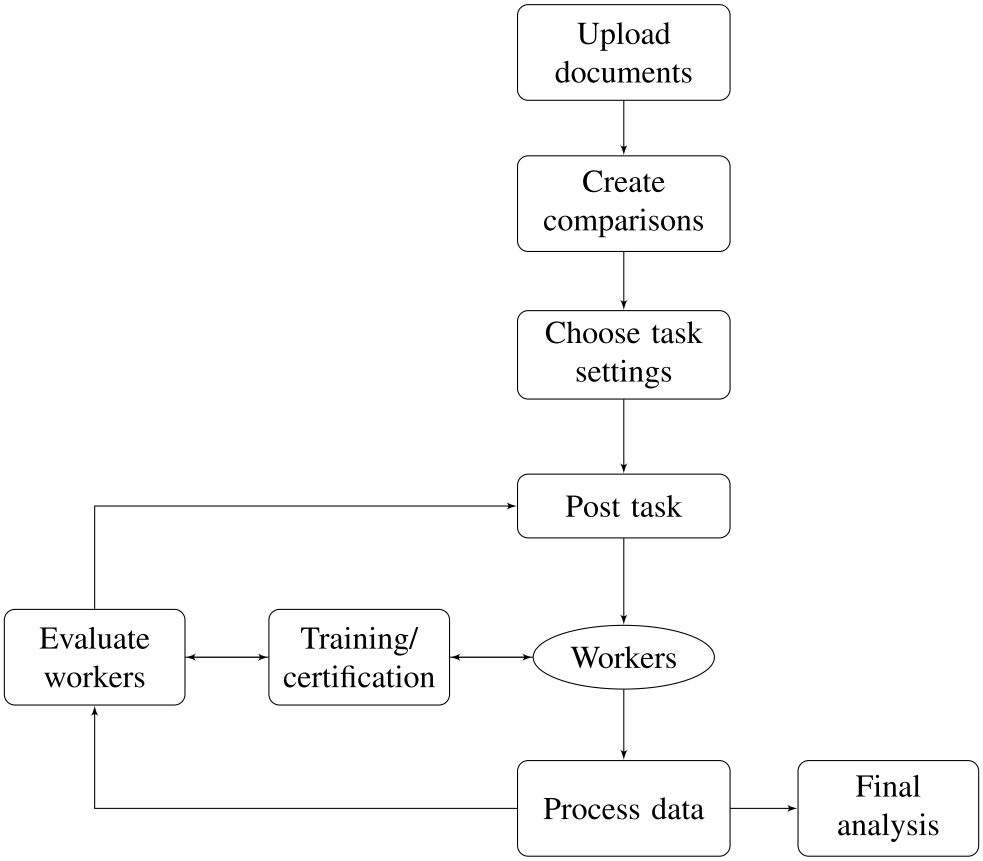
We also rely on redundancy. Each document is included in multiple pairwise comparisons. We found that including each document in 20 pairwise comparisons usually provides very high quality estimates, although deviations from this suggestion may be appropriate on a case-by-case basis. The concept of redundancy builds on extensive research illustrating that aggregated judgments by non-experts are often comparable to those provided by subject experts (e.g., Benoit et al. Forthcoming; Snow et al. 2008; Sheng, Provost, and Ipeirotis 2008).

Finally, our pairwise-comparison framework allows us to easily evaluate the data quality from individual workers as part of our statistical processing (discussed below). Rather than relying on attention filters (Berinsky, Margolis, and Sances 2014) or "gold standard" evaluations (Benoit et al. Forthcoming), we are able to assess the quality of the data from each worker on an ongoing basis as new data become available. SentimentIt provides easy-to-use tools for evaluating workers

[5] One limitation of AMT workers is that they are increasingly drawn exclusively from U.S. citizens, limiting its applicability to only English-language texts. However, the SentimentIt system is designed to be able to integrate with multiple online workforces, and we plan to extend it to allow researchers to post micro-tasks internationally.

[6] Illustrative text from one training model is shown in the online appendix.

Figure 1: SentimentIt workflow



Note: The workflow is straightforward and can be flexibly altered at any point by the researcher. Certifications can be required and crafted to meet the needs of the specific task, and the researcher can evaluate data at any point to revoke certifications of low-quality workers. As much or as little of this workflow as desired can be automated based on the researcher's preferences (see Appendix SI-10).

and removing certifications from workers providing low-quality data.

3.2. *Workflow*

The core functionality of the SentimentIt platform is a cloud-based web application that interacts smoothly with AMT to post jobs, certify workers, store responses, and generally reduce the difficulty for researchers wishing to utilize AMT workers. Researchers access the functionality of SentimentIt via application program interfaces (APIs) that can be called from any computing environment or platform (Python, Java, etc.). However, all of the functionality described below is fully integrated into our R package, making the process for researchers accustomed to the R language especially straightforward. The complete workflow is depicted in Figure 1.

First, we pre-process the textual data. This generally just involves ensuring our text is in a machine-readable format. However, where the documents are too long for simple comparison, we may choose to break the document into shorter, meaningful parts, such as paragraphs. The texts are then passed to `SentimentIt`. After the documents are in the system, we randomly pair the documents into a series of comparisons. For example, if we want 20 comparisons per document for 500 documents, we randomly create 5,000 unique comparisons ($500 \times 20/2 = 5,000$).[7] We then send paired document identification numbers and an associated question (e.g., "Which statement is more positive?") via API.

Once the comparisons are set up, they are ready to be sent to workers. At this stage, we can determine the task settings, dictating how much we want to pay workers and whether we require the worker to have a certification. We then send the comparisons out to the workers via an API call. In most cases, the complete universe of micro-tasks should not be posted simultaneously. We find that posting jobs in batches of 1,000 tasks allows us to keep track of how quickly the tasks are accomplished, and, importantly, gives us the opportunity to analyze the quality of the responses. If we determine that specific workers are providing poor data, we can revoke their certification and prevent them from further contaminating the data.

3.3. *Statistical modeling*

Once a sufficient number of the comparisons are complete, we can download the data via API. The data simply indicate which of the two documents was selected, the unique worker ID, and the time the task was completed. We then process the data using a random utility model, which creates document-level estimates along the dimension of interest (e.g., a measure of a document's positivity). Specifically, we model the probability that one document would be chosen over another, while estimating worker reliability given the choices made by that worker. Let i and j index documents in a comparison. Let k index the workers. The random utility model is specified as:

$$\Pr(y_{ijk} = j) = \frac{\exp(b_k(a_j - a_i)}{1 + \exp(b_k(a_j - a_i))} \tag{1}$$

The model is completed by specifying the following priors:

$$a_j \sim \mathcal{N}(0, 1) \qquad b_k \sim tr\mathcal{N}(0, \sigma^2) \qquad \sigma \sim tr\mathcal{N}(0, 3),$$

where $\mathcal{N}$ refers to the normal distribution, and $tr\mathcal{N}$ refers to the normal distribution truncated at zero to only support positive values.[8] We estimate the model using Hamiltonian Markov Chain

[7] Throughout the text, when we refer to random pairwise comparisons, we are randomly selecting comparisons from the set of all possible unique pairwise comparisons.

[8] Note that we set the variance term for the prior of a_i and σ at 1 and 3 respectively to identify the scale of the latent distribution.

Monte Carlo sampling using Stan (Carpenter et al. 2016).[9] In combination, this model produces posterior estimates for the documents' positions on the latent scale of interest (a_j) as well as the workers' reliability (b_k).

We can extend this model to allow a hierarchical structure. In our final application, we deal with large documents (State Department reports) that require simplification to create suitable micro-tasks. We therefore construct the pairwise comparisons using paragraphs rather than entire documents. To allow for a hierarchical structure of the data, let i and j still index paragraphs in a comparison and k index the worker. We now let m index the higher-level documents. The hierarchical random utility model is still specified as in Equation (1). However, the a estimates are now centered at a higher-level-document mean for document m, denoted θ_m. Letting M be the set of paragraphs contained in document m, the priors are,

$$a_j \sim \mathcal{N}(\theta_m, \sigma_m^2) \forall j \in M \qquad b_k \sim tr\mathcal{N}(0,1) \qquad \sigma_m \sim tr\mathcal{N}(0,.5) \qquad \theta_m \sim \mathcal{N}(0,1).$$

In both models,[10] the parameter estimates for the workers (b_k) give us an assessment of how well each worker performed. Intuitively, these estimates become lower for workers whose choices do not reflect how the documents are understood by other workers. If we estimate a worker as being a (low) outlier, we can ban the worker from future tasks. We find that revoking qualifications from these workers modestly improves the validity of our final estimates relative to benchmarks (see the Online Appendix for additional discussion of worker quality). Therefore, we recommend periodically running the statistical model and removing all workers who are obvious outliers. After disqualifying problematic workers (if any), we can post further tasks and repeat as necessary.

Our `R` package can do as little or as much of this in an automated fashion as the researcher wants (see Appendix SI-10). The package can take in text and post the documents to `SentimentIt`, create comparisons, post tasks, check if the tasks are completed, download the data, test for worker outliers, ban unwanted workers, and repeat the process until all of the desired data have been collected and analyzed. If, instead, the researcher wishes to have more control of every stage, each step specified above can be controlled manually through `R` or via calls to the `SentimentIt` API.

[9] The code for the complete Stan models used in our applications is shown in Appendix SI-3. Tuning parameters are set automatically through Stan.

[10] While there are many alternative ways of modeling this data, in our experience the resulting document-level estimates are largely invariant to these choices. For instance, Table SI-1 in Appendix SI-4 shows the correlations between the a_i estimates used in our applications below are correlated (Person's r) at $0.95 - 0.98$ with the arithmetic mean coder choice (where a document is coded as zero when it is not chosen and as one otherwise). Thus, while we feel that the model above is useful – particularly for identifying low-quality coders – the conclusions we draw below are robust to specific modeling choices and priors.

4 APPLICATIONS

Having described the details of our method, in this section we evaluate `SentimentIt` using texts from an online forum, a survey, political ads, and formal reports from a bureaucratic agency. In each exercise, we demonstrate that our estimates are valid measures of underlying latent traits of interest. To do this, we compare our estimates to relevant benchmarks created either by humans or automated methods. We show that our estimates correlate highly with these benchmarks and argue that where `SentimentIt` scores disagree with benchmarks, `SentimentIt` is usually better at capturing underlying latent traits. Further, for each example we demonstrate that the measures are reliable and replicable. Following the same procedure using the same settings in `SentimentIt` results in estimates that are highly correlated (Pearson's $r \geq 0.86$ in all cases).

4.1. *Movie reviews*

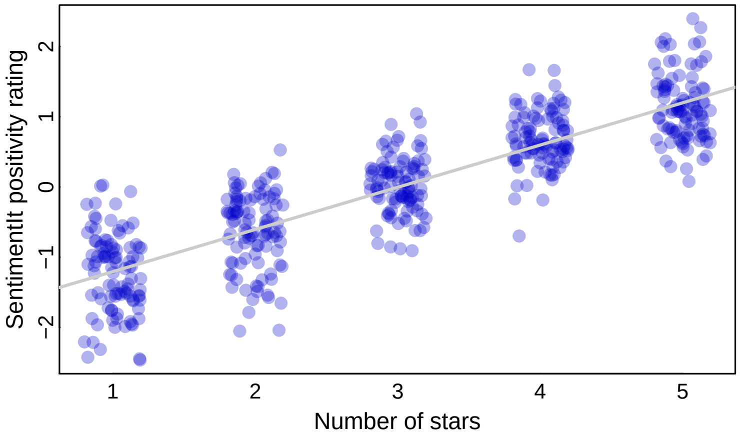
In our first application, we evaluate texts from an online movie review forum. Specifically, we selected 500 user-contributed movie reviews from the Rotten Tomatoes website along with their associated star ratings. Rotten Tomatoes allows users to rate movies on a five-star scale, and we selected 100 reviews from each category. We then apply the `SentimentIt` system to measure the positivity of each review using the text alone. We compare our measure to the star ratings chosen by the reviews' authors, which serve as a benchmark for validation (Pang and Lee 2005). While valuable as a benchmark, it is important to remember that star ratings are discrete and may have different meanings across individuals. Therefore, our first goal is to demonstrate that our estimates are strongly correlated with the star ratings. However, we also wish to show that where disagreement exists, the `SentimentIt` platform often provides a superior numerical summary of the tone in the underlying text.

For this application, we created 40 random pairwise comparisons per document, resulting in 10,000 comparisons. We required participating workers to complete a qualification, and paid them $0.04 per task. We sent tasks to AMT in batches of 1,000 or 500 and analyzed the results between each batch to identify low-quality workers. Throughout the experiment we banned only two workers out of 126.

Figure 2 shows our estimates plotted against the number of stars assigned by the author.[11] The figure shows a very clear trend: as the stars increase so do our estimates (Pearson's $r = 0.87$). There is only a modest level of overlap between categories, with zero one-star ratings scoring higher than any five-star ratings. Further, Table 1 shows seven reviews where our positivity measure disagreed most with the user-provided star ratings. Our readings of these texts is that where `SentimentIt` estimates disagree with the stars, the star ratings seem to have been assigned in

[11] For the purposes of exposition, in the main text we focus exclusively on the point estimates (posterior means) for each document. However, the model also produces posterior measures of uncertainty for each estimate, which we discuss in Appendix SI-11.

Figure 2: `SentimentIt` movie review positivity estimates on number of stars



Note: As the number of user-provided stars increases, so do the estimates of positivity from `SentimentIt`. There is little overlap between estimates even as proximate as two stars.

a manner not supported by the text. That is, we believe that our measure of the underlying sentiment more accurately reflects how a *standard* reader would translate the language into stars. For instance, a review that reads (in part), "Almost plotless, but with moments that stick to your soul like a coating of grime," is language more consistent with a slightly negative review rather than the four-star rating assigned by the author.

How many comparisons are needed to generate valid estimates of latent traits embedded in texts? To provide an answer to this question, we estimated positivity measures using the first 10 pairwise comparisons (per document), the first 20 comparisons, the first 30, and then the complete dataset.[12] Figure 3 shows the correlation between these estimates of positivity and the user-provided stars. After 10 comparisons per document, the correlation is 0.84. After 20, the correlation is 0.86. After 30 and 40, the correlations are both about 0.87. Further, the point estimates after analyzing only 20 comparisons are virtually identical to the point estimates after analyzing 40 ($r = 0.99$). There is a very mild gain in precision, with the mean standard deviation of the estimates decreasing from 0.26 to 0.22 as we move from 20 to 40 comparisons. Thus, in this experiment there is little benefit to adding tasks beyond 20 comparisons.

Turning to the question of reliability, Figure 4 plots the point estimates generated by analyzing *only* the first 20 comparisons and those estimated using *only* the last 20 comparisons. The correlation between these estimates is 0.92. This is strong evidence that the `SentimentIt` measures are highly reliable, and that 20 comparisons is an adequate number to minimize costs while maximizing reliability.

Finally, in Appendix SI-1, we provide evidence from additional analyses of online movie re-

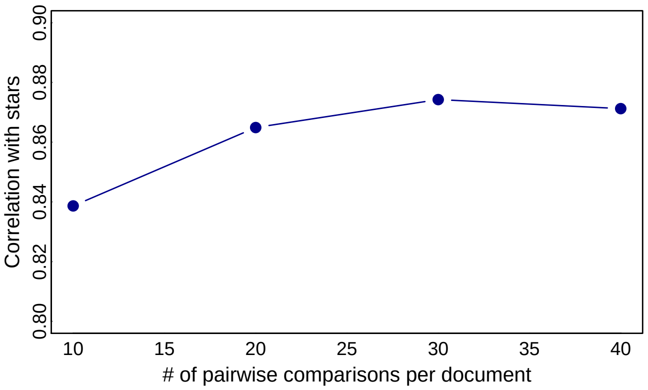
[12] Pairwise comparisons were constructed in blocks of 10 to make this analysis feasible.

Table 1: Reviews with largest disagreement between `SentimentIt` scores and star ratings

Review text	Positivity score	Stars
"I really enjoyed the original with Brooke Shields and yes it is on my movies that are very hard to find list, however deemed this sequel just ok. Somethings they should just leave as they are."	0.03	1
"Eddie Murphy. What happened to your taste in movie roles? This was the stupidest movie..not funny at all. Just stupid."	-2.05	2
"A film that quite possibly showcased Monroe playing herself, especially late in her career. Nell was unstable and not able to handle the pressures of the outside world (hence the metaphor of being a sort of live-in nanny, cut off from the rest of the world but protected as well). Monroe's's late-career personal troubles and demons were well documented and attested by her suicide. Even though this film was early on in her career it offers a chilling and eerie view of what's to come for the legendary actress. They say the best of the best actors and actresses draw upon their very souls to come up with their startling performances. I think Monroe was pulling from her very core in this film. Worth a viewing and analysis."	1.04	3
"I think Buster Keaton is one of the more inconsistent actors from the silent film era. I really didn't like The general but I adored Sherlock Jr. This one I would say is ok. Buster and the woman who rejected to marry him accidentally both end up on a ship at sea alone. In this journey they encounter a storm, cannibals, and a scary painting of sailer. Now there were some nice laughs in here but at the same time for a film only an hour long I shouldn't have been bored as much as I did."	-0.91	3
"Residents of an institution escape and wreck the grounds with childlike acts of vandalism and petty cruelty. The entire cast is composed of dwarfs. Almost plotless, but with moments that stick to your soul like a coating of grime (tiny Hombre laughing at the struggling camel may haunt your nightmares for years to come). Animal lovers beware."	-0.70	4
"The best sequel and best boxing film since Rocky. Stallone is superb and now I know why he was applauded and Michael B. Jordan is amazing as man trying to respect his legacy but find his own path to victory. Also good is Thomson as Jordan's singer love interest. Great directing, writing and cinematography. Was cheering in my seat and left the theatre feeling satisfied."	1.67	4
"This is a story of an unjust system, looking only to protect its own neck. This is the outrage of onlookers and commentators who cannot stand the ridiculous logic behind taking a man's life away when he did not commit any wrong. Most of all, this is one young man's brave fight to show the world that he can be beaten down, but not beaten. The story is reminiscent of Mumia Abu Jamal's own plight, except the circumstances of Paco's innocence are more blatant and apparent."	0.08	5

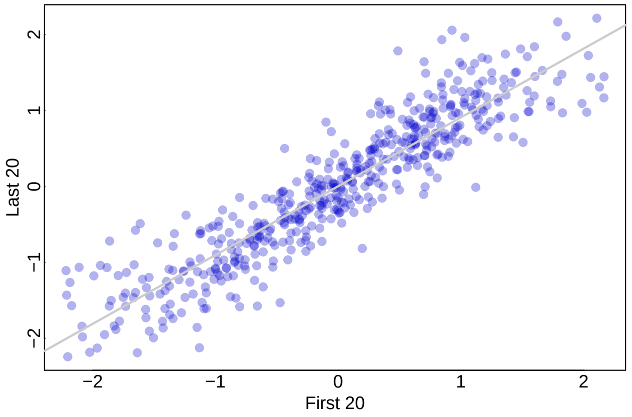
Note: It is evident that when `SentimentIt` estimates movie reviews in an unexpected way given the star ratings, the star ratings seem to be given in a way that differs from average ratings.

Figure 3: Correlations between SentimentIt estimates and movie review stars



Note: Analyzing data after 10, 20, 30, and 40 comparisons shows that after 20 comparisons, the increase in correlations between SentimentIt estimates of positivity and movie review stars diminishes significantly.

Figure 4: Reliability of SentimentIt measures of positivity in online movie reviews



Note: We analyzed the first 20 and last 20 pairwise comparisons estimating the positivity of movie reviews separately. This plot shows the high correlation (0.92) and therefore high reliability between runs.

views that further illustrate these points. Importantly, in this application, data was collected over a month apart using mostly different workers, yet the reliability estimates are virtually the same. In Appendix SI-2, we benchmark SentimentIt against an advanced supervised learning method (Socher et al. 2013) using a similar dataset to demonstrate that our approach surpasses even the best available automated methods for supervised sentiment analysis.

4.2. *Open-ended survey responses*

In our next application, we analyze a corpus of open-ended survey responses about immigration collected by Gadarian and Albertson (2014). This example is useful because these statements have been analyzed using both human content coding and an automated structural topic model (STM). Gadarian and Albertson (2014) used two trained research assistants to code each statement as indicating no (0), some (1), or extreme (2) levels of fear, anxiety, or worry towards immigrants or immigration. Coders were given specific instructions to distinguish between statements that indicated anger versus those that indicated fear, anxiety, or worry. They averaged the two coders' evaluations to generate a 2-point scale for each response. Roberts et al. (2014) subsequently analyzed these same statements using a structural topic model (STM), and identified one topic as indicating fearfulness. Thus, we are able to assess SentimentIt relative to two alternatives.

We analyze the survey responses with 40 pairwise comparisons per document in two separate experiments of 20 comparisons. We required a certification and paid $0.04 per task. Workers were instructed to choose which statement indicated the greater degree of fear, anxiety, or worry towards immigrants.[13] Each experiment was completed in about three hours. These experiments were approximately one week apart. In total, 20 workers participated and we banned two.

The left panel of Figure 5 compares the SentimentIt estimates relative to the mean expert-coder rating. Figure 6 compares the correlations between SentimentIt, each expert coder, the mean of the expert codes, and the STM topic probabilities. The figures show that as the expert coding increases, the SentimentIt measure does as well. The SentimentIt scores are correlated with the mean expert rating at $r = 0.78$, a modestly strong correlation. Indeed, the SentimentIt measures correlate more highly with each expert coder than the individual coder scores correlate with each other.

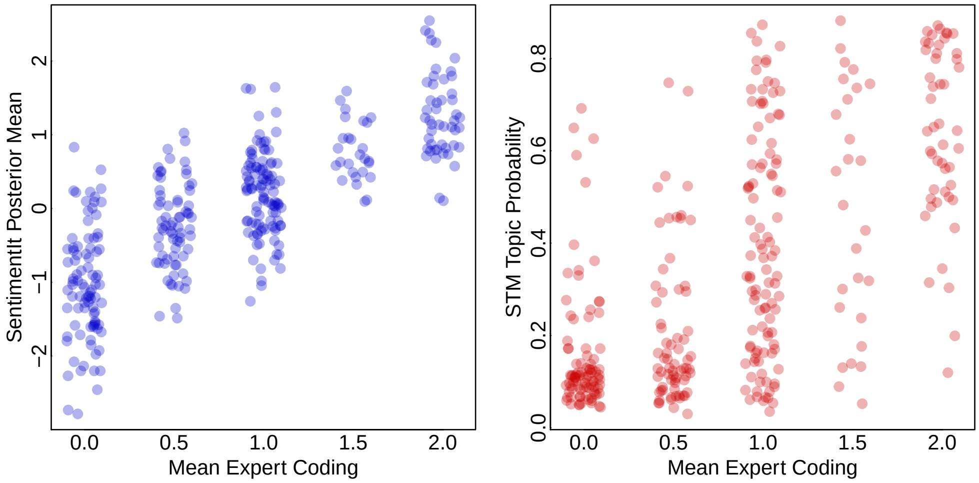
One advantage of the SentimentIt approach is that it provides scores on a continuous metric and can reveal important variation within the broad categories identified by the expert coders.

[13] The full prompt was:

> Which of these two statements expresses more fear, anxiety, or worry about the negative impact of immigrants or immigration on America? Remember, we are not interested in whether the writer dislikes immigrants, wants them to go home, resents them, or blames them. We are only interested in whether the writer is expressing fear, anxiety, or worry.

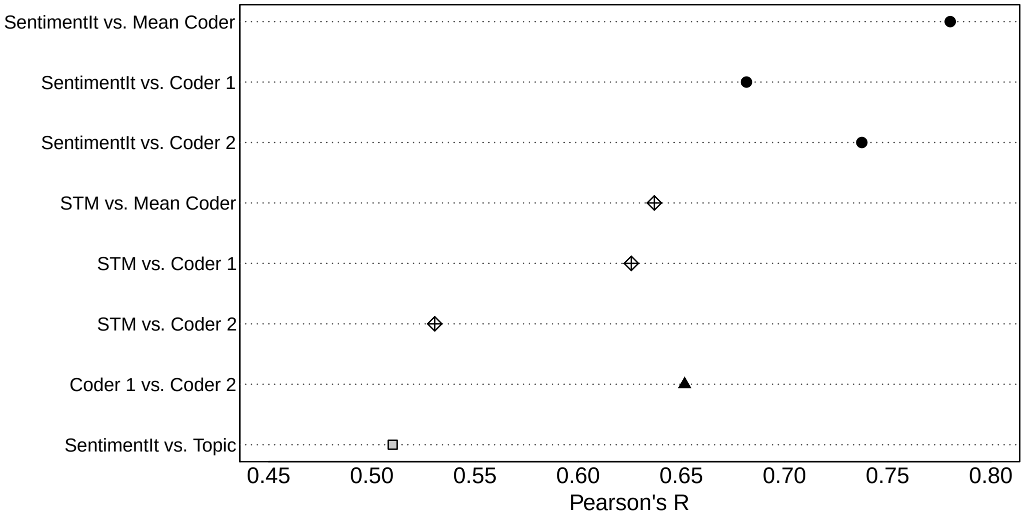
This prompt, as well as the required training, were designed in consultation with the original authors to replicate their coding scheme.

Figure 5: Fearfulness estimates from `SentimentIt` and STM on expert coding



Note: The estimates from `SentimentIt` increase as do the human codes. The STM estimates, however, tend to pool towards zero and perform particularly poorly at higher human codes.

Figure 6: Correlations between different sources of fearfulness estimates



Note: The correlations between `SentimentIt` and the human coders are notably higher than the correlations between STM and the coders. The human coders are also poorly correlated.

Table 2: Variation within categories uncovered by SentimentIt

Survey response	Expert coders	SentimentIt
when i think of immigration i think of people who enter this country legally, who go through the proper immigration process, no matter how long it takes. i think of people who are willing to learn the english language, make an honest living, honor our country and pledge allegiance to our flag. those who come to america by any other means, who sneak in here and file false paperwork, who think they have the right to drive and have a license, who manage to obtain false ssn's don't deserve to be here, and our borders need to be much, much more secure.	None (0)	0.83
i am fine with immigration.	None (0)	−2.79
illegal immigrants coming into our country.illegal immigrants being a artificially low wage earner which takes away jobs that could go to americans.illegal immigrants working here but not paying their fair share of taxes since it is in large part an underground ”cash based” society	Some (1)	1.64
all the illegals coming in to the country.	Some (1)	−1.26
job losses, threat of terrorism, property value decreasing, drug trafficing, becoming citizens and going on welfare	Extreme (2)	2.58
jobs crime taxes	Extreme (2)	0.11

Note: The table shows responses that were unanimously coded by the expert coders as belonging in categories 0, 1, or 2. The SentimentIt approach reveals significant variation within categories that is largely consistent with a close reading of the actual texts.

Table 2 shows statements that were unanimously categorized by the expert coders (scored 0, 1, or 2). The statements listed are those which SentimentIt identified as being the most fearful or least fearful within each category. These examples illustrate that there is significant variation within categories (as identified by the expert coders) that is successfully identified by the SentimentIt approach. For instance, the system correctly identifies that a statement where a respondent indicates concerns that immigrants being “artificially low wage earner takes away jobs that could go to Americans” exhibits more fear than a statement that mentions less specific concerns about “all the illegals coming in to the country.” Likewise, it correctly identifies statements like, “i am fine with immigration,” as indicating far less anxiety than those that express concern about illegal immigrants who “think they have the right to drive” along with a preference for legal immigrants who “honor our country and pledge allegiance to our flag.” In general, a close reading of the texts and the associated scores show that the SentimentIt system is not perfect, but it assigns scores to texts that broadly reflect the level of fear indicated in the text—in many cases more so than the scores assigned by the expert coders.

Table 3: Anomolous topic probability assignments among unanimously coded documents

Expert coders	Statement	STM prob. fearful	`SentimentIt`
None (0)	"nothing"	0.65	-2.20
None (0)	"they do not pay taxes"	0.69	-0.40
Extreme (2)	"too many non speaking english americans. too many people that do not stand with what we were founded on. too much trouble wanting their ways and not becoming americanized"	0.12	0.74
Extreme (2)	"changing the basic makeup of the united states. creating a population with more socialistic values. mexico is largely an economic failure, so the immigrants from there may tend to have the values that created that failure."	0.20	0.81

Note: When the structural topic model misclassifies, the topic model is more likely to be in error than the expert coders.

Topic models are not intended to provide measures of concepts within texts as defined *ex ante* by researchers. Nonetheless, it is instructive to compare the performance of `SentimentIt` to more traditional automated approaches for encoding texts. The right panel of Figure 5 shows how the mean coder rating compares to the probability that a document is assigned to the "fearfulness" topic by the STM model (Roberts et al. 2014). In general, the STM does not capture the sense of anxiety or fearfulness as accurately. Specifically, the STM model tends to pool towards the low end, and the predictive validity of the STM model is particularly poor at capturing moderate and high levels of fearfulness. In the end, STM probabilities are only correlated at 0.64 with the mean expert coding.

This degree to which unsupervised methods often imperfectly capture researcher-specified concepts is further illustrated by examining specific texts. Table 3 provides examples where the topic model most strongly disagrees with the expert coders. As can be seen, the topic model estimates several benign responses as fearful, such as "nothing" and "illegallanguage" (sic). The bottom entries are statements that the automated method identified as non-fearful where the expert coders unanimously identified them as extremely fearful statements. We see that the topic model categorizes some responses as low that actually exhibit strongly fearful emotions mentioning worries such as "changing the basic makeup of the United States."

Finally, we turn to the question of reliability. Note that the two expert coders provide estimates that are correlated at only 0.65. Further, although in a trivial sense the topic model is perfectly reliable,[14] in practice the model estimates are a subjective choice of the researcher since

[14] Starting the algorithm using the same random seed will necessarily lead to the same estimates.

the topics identified by the algorithm change significantly depending on small changes in initial conditions. So, for instance, Roberts et al. (2014) ran 50 topic models to analyze this dataset and chose one "based on exclusivity and semantic coherence criterion" (p. 1073).[15] In comparison, the `SentimentIt` procedure is transparent, reliable, and does not rest on the judgment of the researcher. Approximately one week after our initial data collection, we exactly duplicated our process and ran 20 more comparisons with the same settings and analyzed the resulting data separately. The correlation between these two `SentimentIt` analyses is a very impressive $r = 0.86$.

4.3. *Congressional advertisements*

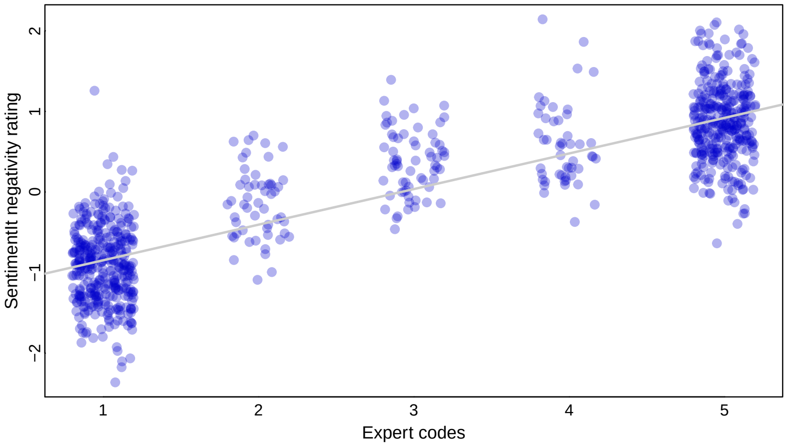
For several election cycles, the University of Wisconsin Advertising Project (WiscAds) collected and coded political ads in the United States and released their data for political science research. In this application, we focus on the "tone" of ads, or their level of negativity, which is a measure widely used in political science research (e.g., Freedman and Goldstein 1999; Goldstein and Freedman 2002). In the WiscAd dataset, ad tone is determined by expert coders who categorized ads as either promoting a single candidate, contrasting two candidates, or attacking a candidate. If the ad is contrasting, it is further categorized as either being more aimed at promoting than attacking, more attacking than promoting, or equally attacking and promoting. The result is a five-point scale of negativity ranging from one (positive) to five (attack). We apply `SentimentIt` to analyze all televised ads for the U.S. Senate in 2008 ($n = 942$) and compare our estimates to this five-point negativity scale. This allows us to again validate `SentimentIt` against a meaningful benchmark. Further, this example illustrates how a tedious expert-coding task can be more easily accomplished via the automated `SentimentIt` approach while actually improving the reliability and validity of the measure.

For each ad, we created 20 pairwise comparisons for a total of 9,420 tasks. Workers were required to complete an extensive training module and were paid $0.06 per comparison. Workers were instructed to select the ad that was "most negative towards the candidate(s) mentioned, or least positive about the candidate(s) mentioned." In all, 123 workers participated in the task and none were banned.

Figure 7 shows our estimates plotted against the five-point negativity scale. The `SentimentIt` scores are correlated with the expert codings at $r = 0.85$. This is a very high correlation, but there are some disagreements. Table 4 shows a few of the greater disagreements between the `SenitmentIt` measures and those generated by the expert coders. The first example is coded as positive by the expert coders, but the ad is negative in tone and appears to be a strongly negative contrasting ad that was mis-coded. The second ad is coded as contrast (2 = more positive than attack), but `SentimentIt` estimates it as having almost no negative content. In fact, the

[15] A technical solution for the sensitivity of structural topic models to starting values is addressed in subsequent work (Roberts, Stewart, and Tingley 2016).

Figure 7: `SentimentIt` negativity rating relative to five-point expert codes



Note: Using the five-point scale of human codes, we can see that as human codes increase, so do our estimates. In general, the differentiation between codes is evident. The largest disagreements are happening in the middle categories, as discussed in the text.

ad does not mention an opposing candidate or party and is another clear mis-coding. The third ad is coded as contrasting by expert coders because it mentions both candidates by name. However, the ad mostly consists of strong and negative language, causing our estimate of negativity to be quite high (1.39). Finally, the last example is coded as an attack by human coders because it does not provide positive information about a named candidate. However, the ad only mentions the target of the attack (Collins) in one sentence and otherwise conveys positive information about the Employee Free Choice Act. These last examples illustrate that the strict coding rules designed to improve the reliability of content analysis can obscure the underlying latent trait of interest.

To determine the reliability of our measure, we repeated the exercise 35 days later with 20 more comparisons on a random sample of half the ads ($n = 467$) using the same procedure as before. In all, 52 workers participated in this exercise. We revoked the certification from only one worker. The correlation between the two runs is 0.90.

4.4. *Measuring torture using human rights reports*

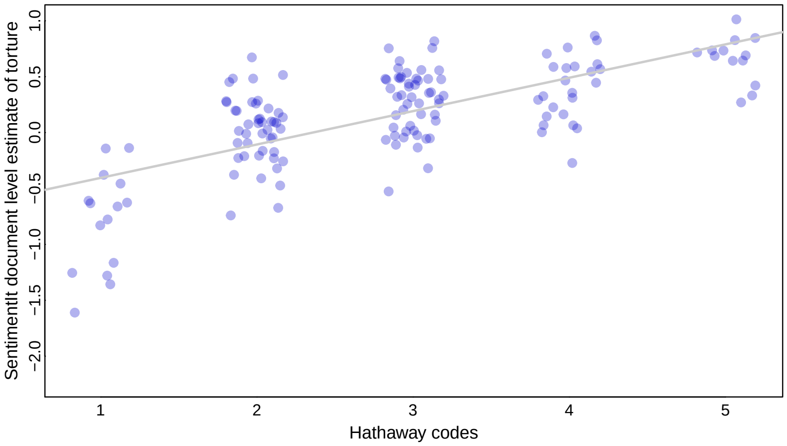
In our final application, we demonstrate how `SentimentIt` can be generalized to larger documents. Specifically, we turn to the sobering task of coding the section entitled, "Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment," from all of the U.S. State Department Human Rights Reports issued annually for nearly every country in the world (Fariss et al. 2015). In this application, we use reports from 1999 to create a continuous scale indicating the amount of torture conveyed in the report. Hathaway (2002) codes these documents by hand on a five-point scale, with higher values indicating more entrenched, brutal, or frequent torture. Using this measure, Hathaway (2002) argues that ratifying human rights treaties is associated with greater degrees of torture (see also Neumayer 2005). We emphasize that our aim is not to capture

Table 4: Example ads where SentimentIt disagreed with expert coders

Advertisement	SentimentIt	Experts
[Priscilla Lord Faris]: "Early in this campaign I believed that Al Franken could defeat Norm Coleman. But, no matter how many millions he spends it is clear that his history of pornography, degrading women and minorities, and his questionable financial transactions will continue to be the focus of blistering Republican attack ads. I represent real Minnesota values as a mother, a teacher, a volunteer, and an advocate. I'm Priscilla Lord Faris, I approve this message, and ask for your vote September 9th."	1.26	Positive (1)
[Steve Novick]: "I'm Steve Novick and I approve this message." [John Kitzhaber]: "I'm John Kitzhaber and I approve of Steve Novick. Negative politics as usual or something different? Steve Novick is not a typical politician and he's not running a typical campaign. Steve is standing up for principle and that's why Oregon Democrats are standing up for Steve. Oregon teachers are supporting Steve, so are papers across the state. And I think Steve Novick is the only candidate we can count on for real healthcare reform. Steve Novick, the cure for politics as usual."	-1.09	Contrast (2)
[Announcers]: This isn't complicated. Roger Wicker serves with honor and integrity. Ronnie Musgrove. His ethics? Shameful. Roger Wicker. Supported by Thad Cochran, the VFW, the NRA. Ronnie Musgrove. Supported by pro-abortion, pro-gay marriage groups. Roger Wicker. Never voted for a pay raise, always supports Social Security. Ronnie Musgrove. Failed governor, lost jobs, the beef plant scandal and now he's lying about Roger Wicker. This isn't complicated. Roger Wicker. [Roger Wicker]: "I'm Roger Wicker, and I approve this message."	1.39	Contrast (3)
[Announcer]: CEO's salaries and benefits are getting fatter and fatter... while workers face soaring gas prices, foreclosures, and rising healthcare costs. The Employee Free Choice Act gives workers the freedom to form a union so they can earn better wages, retirement security, and healthcare coverage. Call Senator Susan Collins tell her to support the Employee Free Choice Act and stop siding with wealthy CEO's over working families. American Rights At Work is responsible for the content of this advertising.	-0.64	Attack (5)

Note: The first two examples show errors in expert coding, as the ads were put in the wrong category. In the latter two, it is evident that when SentimentIt estimates differ from human codes, expert coding scheme's reliance on strict coding rules mischaracterizes the overall tone within the ads.

Figure 8: SentimentIt document-level estimates of torture on Hathaway codes



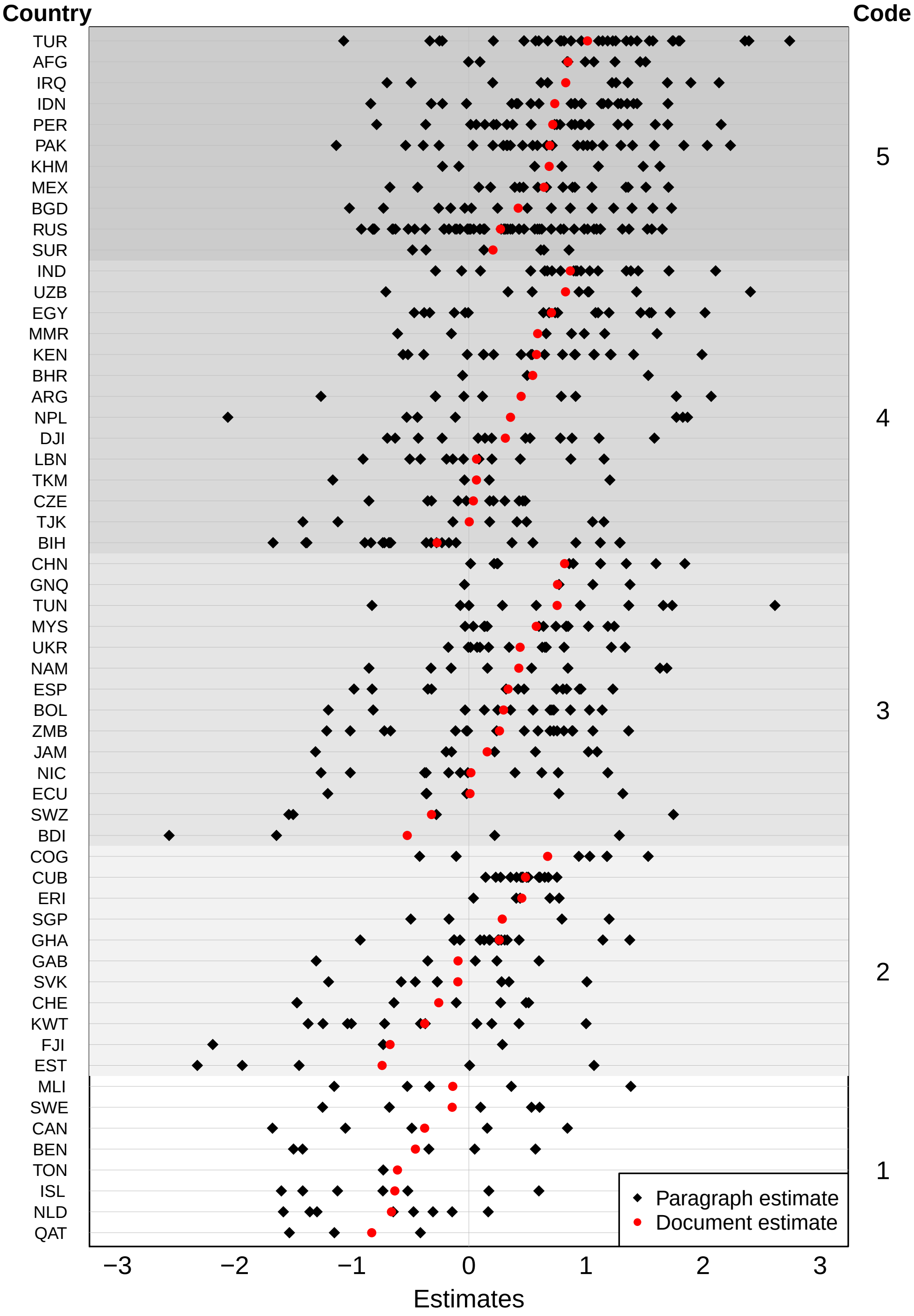
the actual amount or severity of torture in each country – a task well beyond the scope of this article – but only to measure the concept of torture as it is expressed in these reports.

Since many of these reports are quite lengthy and would be difficult for even an expert to read and meaningfully compare, we divided the 182 documents into 1,652 paragraphs. We created 20 comparisons per paragraph. Since reading and understanding these documents is far more challenging than the examples above, we paid workers $0.10 per task and required an extensive certification training (see Appendix SI-6). We asked workers to indicate which paragraph suggested more torture, which was defined in detail during the training. Our definition and coding rules were designed to approximate those provided by Hathaway.

We posted the tasks in batches of 1,000. The process of gathering the data took approximately three days, and 81 workers completed tasks (we banned none). To analyze the data, we adjusted our statistical model to allow for a hierarchical structure (paragraphs within documents) as described in Section 3. The resulting document-level estimates are correlated with Hathaway's coding at 0.69. Figure 8 plots the SentimentIt estimates of torture to the Hathaway code. In Figure 9, we show a subset of the document- and paragraph-level estimates generated by our procedure.

Broadly speaking, our measure is consonant with Hathaway's coding. For instance, all of our estimates for the countries coded as one (no torture) by Hathaway fall within the first quantile of our estimates. Yet, there are significant disagreements. For instance, among those cases coded as a one, our estimates are noticeably higher for specific cases. Our estimate for Mali (-0.14) is the highest among all countries coded as one in Hathaway's scheme. Two paragraphs drive this result: one explains at length the harsh prison conditions of the country (estimated at 0.36) and the other explicitly states Amnesty International reported that "security forces had tortured ... in order to extract confessions" (estimated at 1.38). On the other hand, we estimate some countries coded as two slightly lower than expected. We estimate Estonia lower than any other country in category two (-0.74). The only evidence of torture from this document is a paragraph stating, "police used excessive force and verbal abuse during the arrest and questioning of suspects," and "[p]unishment

Figure 9: `SentimentIt` document- and paragraph-level estimates of randomly selected countries and countries estimated differently than Hathaway



Note: Document-level estimates are shown in red, and paragraph-level estimates are shown in black. There is a clear trend that as Hathaway's codes increase, so do ours. The notable exceptions are explicitly included and discussed in the text. Where our estimates deviate from Hathaway's codes, there is strong reason to believe estimates from `SentimentIt` are outperforming expert codings.

cells ... continued to be used" (estimated at 1.07).

The Republic of the Congo is coded as a two by Hathaway, but our estimate (0.67) is higher than any other document coded as a two. The report states explicitly that police and security forces were using beatings, rape, unwarranted strip searches, and unlawful imprisonment to solicit confessions, impose power, and punish. Our estimate for Cuba is also relatively high (0.48), which is also coded as a two by Hathaway. The Cuba document is a lengthy account of many acts of violence by police, unwarranted detention, acts of indirect violence against those who do not support the government, and harsh prison conditions.

Burundi is coded as a three in Hathaway, but we estimate Burundi to be on the lower end (-0.53). The first paragraph reads "members of the security forces continued to torture and otherwise abuse persons. In one such case, [Amnesty International] reported that members of the security forces were believed to have withheld food from detainees and beaten one of them severely. There were no known prosecutions of members of the security forces for these abuses" (estimated at 1.29). This is a clear indicator of torture, but it is the only mention of torture in the document, making it not obviously distinguishable from other reports coded as two in the Hathaway scheme. On the other hand, China (0.82), Tunisia (0.75), and Equatorial Guinea (0.76) are all categorized as threes, but we estimate rather high degrees of torture for these countries. The documents for all three contain multiple paragraphs detailing severe torture including beatings, administering electric shocks, hanging prisoners and suspects by their wrists, and torturing detainees to death.

Bosnia and Herzegovina is coded as a four, but we actually estimate it at slightly lower than the mean amount of torture (-0.27). The document has clear instances of violence, but overwhelmingly the violence described is not at all related to police or government acts of torture. Some of the paragraphs are actually positive, for example: "[i]nternational community representatives were given widespread and for the most part unhindered access to detention facilities and prisoners in the RS as well as in the Federation" (estimated at -1.39). Though this was clearly a tumultuous time for Bosnia, very little of the document speaks to torture. The same is true for Turkmenistan (0.06), Tajikistan (0.003), Lebanon (0.07), and the Czech Republic (0.04). The overall tone of all these documents is not very negative, at least with regard to torture.[16]

Based on these and other examples, we believe that the `SentimentIt` coding more faithfully reflects the level of torture in these documents than the original human-coded measures. However, to further assess the validity of these estimates, we compare our measure with three other well-known measures of torture drawn (in part) from this same document set. Specifically, we analyze the State Department variable of the Political Terror Scale (PTS) data on human rights viola-

[16] Space does not a permit a fuller discussion of all the ways in which our coding disagrees with the Hathaway scheme. However, the full document set along with our estimates will be included in the replication archive for this article at the time of publication.

Table 5: Standardized OLS coefficients for SentimentIt and Hathaway measures regressed on torture measures and treaty ratification

	Dependent variable:							
	ITT		PTS		CIRI		Torture Convention	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hathaway	**0.724**		**0.263**		**0.275**		0.252	
	(0.210)		(0.079)		(0.064)		(0.523)	
SentimentIt		**0.964**		**0.298**		**0.275**		**1.296**
		(0.205)		(0.078)		(0.064)		(0.517)
Observations	93	94	106	108	103	105	106	108
R^2	0.382	0.442	0.654	0.663	0.511	0.510	0.343	0.367

Note: Bolded coefficients are significant at the $p \leq 0.05$ level. Additional controls for per capita GDP, population, population growth, trade, foreign aid, GDP growth, civil war, level of democracy, and nation durability not shown. SentimentIt and Hathaway codes are standardized for comparability. CIRI increases as the degree of torture decreases, and is reverse coded in the analyses.

tions (Gibney et al. 2015), the Ill-Treatment and Torture (ITT) data (Conrad and Moore 2012), and the torture variable from The CIRI Human Rights Dataset (Cingranelli, Richards, and Clay 2014).[17] We also test whether ratification of a torture convention is positively correlated with the amount of torture in a country as hypothesized by Hathaway. Hathaway (2002) finds that signing an anti-torture convention is positively but unreliably associated with torture. For each variable, we conduct separate regressions and calculate standardized regression coefficients. We maintain all controls of the original analysis. Our only deviation is that we look at only one year while the original analysis included 15 years of reports.

In almost all cases, we find that both the Hathaway and SentimentIt scores are positively related to the other measures ($p \leq 0.05$). The only exception is that the relationship between level of torture and ratifying an anti-torture convention is estimated unreliably using Hathaway's coding. However, although both measures are correlated with the alternative torture measures, both the standardized coefficients and R^2 values indicate that the SentimentIt measure is more highly related to these other variables. The only exception is the CIRI measure, where the two approaches are essentially indistinguishable in terms of their predictive validity.

Finally, to test for reliability of our measures we separately estimate the (roughly)[18] first 10 comparisons to the last 10 comparisons for each paragraph. The correlation between paragraph estimates is 0.77. The correlation between document estimates is 0.88. Considering we are only comparing the estimates between two rounds of 10 comparisons, we consider this to be strong

[17] Additional information on these measures is provided in Appendix SI-9.

[18] Because the number of paragraphs was not divisible by 1,000, our last batch only had a few hundred comparisons, necessitating the division into halves be slightly less than perfect.

evidence indicating the `SentimentIt` platform is generating highly reliable measures.[19]

5 CONCLUSION

Human language is central to the processes and outcomes of politics. However, this rich source of data is often overlooked in favor of more easily quantified behaviors. Many studies that do utilize texts and speeches rely on hand coding by trained experts, which is time-consuming, costly, and often unreliable. More recently, scholars have turned to automated text analysis algorithms that, while very promising, can perform poorly when uncovering the underlying latent sentiment of political texts and are primarily aimed at classification rather than measuring continuous traits.

In this paper, we propose a novel human computation framework approach to analyzing political texts and measuring latent traits that provides the reliability of automated methods but leverages the superior abilities of humans to read and understand natural language. Specifically, we rely on having an online workforce complete pairwise comparisons of texts. By having the workers indicate which of two documents is more extreme along a dimension of interest (e.g., positivity), including documents in multiple pairwise comparisons, training and monitoring workers, and statistically post-processing the worker evaluations, we are able to reliably produce valid estimates of latent traits within texts. Further, we provide software that can automate as little or as much of the process as the researcher desires and provide details about the workflow and steps needed to replicate our approach.

To demonstrate the benefits of the `SentimentIt` system, we apply it to measure positivity in movie reviews, fearfulness in open-ended survey responses about immigration, negativity of political ads, and levels of torture indicated in U.S. State Department human rights reports. In each application, we use meaningful benchmarks to show that estimates produced by `SentimentIt` are reliable and valid measures of underlying latent qualities. We believe that these results show that our approach can be fruitfully applied to the analysis of natural language in a wide variety of applications across subfields in political science and beyond.

Though the method we propose is quite powerful, it is not entirely unproblematic. First, although the examples above show the versatility of the method, there may be unknown limits to its applicability. Asking coders to evaluate the "tone" of a political ad may differ in kind from asking them to evaluate the quality of the legal reasoning in court cases. Moreover, the tasks we designed were specifically aimed at evaluating aspects of a text that are somewhat objective. Thus, bias in the pool of workers – even if widely shared – is unlikely to contaminate the data. However, asking workers to draw more deeply on their own judgment to evaluate, say, the "persuasiveness" of a specific text, may require a more representative workforce. Whether the `SentimentIt` system

[19] Further, this level of reliability is actually higher than the expert coding procedure reported by Hathaway (2002, p. 1971) (Cohen's κ=0.8).

can be successfully applied in these situations requires further investigation.

However, the most obvious limitation is that, though the system is much cheaper than human coding, it is not free. If we are coding medium-sized document sets, this is a relatively trivial problem. For example, our second application involving open-ended survey response cost approximately $130 and a few hours to complete. Indeed, we believe that SentimentIt may generally be more cost effective than relying on expert coders in almost all cases. However, the method is clearly infeasible with large document sets containing tens of thousands of documents and is much more expensive than relying on unsupervised automated methods.

Nonetheless, integrating the evaluations of online workforces into the process of encoding text is not limited to the exact procedure we discussed above. Moving forward, we plan to use this platform to handle larger document sets by further melding together the power of human workers and computer algorithms. First, we can use SentimentIt to build supervised learners by providing high-quality training sets. Second, we plan to extend the method to be more dynamic, combining supervised machine learning algorithms with pairwise comparisons from online workforces dynamically to allow sophisticated algorithms to *themselves* post comparisons to gain more information about specific documents. Finally, the platform could also be used as a verification of existing automated methods. That is, when researchers use a text-analysis algorithm, they can analyze some documents to assess the validity of the results (Grimmer and King 2011).

5 References

Baumgartner, Frank R., and Bryan D. Jones. 1993. *Agendas and instability in American politics*. Chicago: University of Chicago Press.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. Forthcoming. "Crowd-sourced text analysis: Reproducible and agile production of political data." *American Political Science Review* .

Berinsky, Adam J., Gergory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 329–50.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3): 739–753.

Brady, Henry E. 1985. "The perils of survey research: Inter-personally incomparable responses." *Political Methodology* 11(3/4): 269–291.

Budge, Ian. 2001. *Mapping policy preferences: Estimates for parties, electors, and governments, 1945-1998*. Vol. 1 Oxford University Press.

Budge, Ian, and Paul Pennings. 2007. "Do they work? Validating computerised word frequency estimates against policy series." *Electoral Studies* 26(1): 121–129.

Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael A. Betancourt, Michael Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2016. "Stan: A probabilistic programming language." *Journal of Statistical Software* .

Cingranelli, David L, David L Richards, and K Chad Clay. 2014. "The CIRI human rights dataset." *CIRI Human Rights Data Project* 6.

Conrad, Courtenay R., and Will H. Moore. 2012. "Ill-Treatment and Torture (ITT) Dataset." `http://www.politicalscience.uncc.edu/cconra16/UNCC/ITT_Data_Collection.html`.

Druckman, James N., Martin J. Kifer, and Michael Parkin. 2009. "Campaign communications in US congressional elections." *American Political Science Review* 103(3): 343–366.

Fariss, Christopher J, Fridolin J Linder, Zachary M Jones, Charles D Crabtree, Megan A Biek, Ana-Sophia M Ross, Taranamol Kaur, and Michael Tsai. 2015. "Human rights texts: converting human rights primary source documents into data." *PloS one* 10(9): e0138935.

Fenno, Richard F. 1978. *Home style: House members in their districts*. Boston: Little, Brown.

Freedman, Paul, and Ken Goldstein. 1999. "Measuring media exposure and the effects of negative campaign ads." *American Journal of Political Science* pp. 1189–1208.

Gadarian, Shana Kushner, and Bethany Albertson. 2014. "Anxiety, immigration, and the search for information." *Political Psychology* 35(2): 133–164.

Gibney, Mark, Linda Cornett, Reed Wood, Peter Haschke, and Daniel Arnon. 2015. "The Political

Terror Scale 1976-2015.”.

Goldstein, Ken, and Paul Freedman. 2002. “Campaign advertising and voter turnout: New evidence for a stimulation effect.” *Journal of Politics* 64(3): 721–740.

Grimmer, Justin. 2010. “A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases.” *Political Analysis* 18(1): 1–35.

Grimmer, Justin. 2013. “Appropriators not position takers: The distorting effects of electoral incentives on congressional representation.” *American Journal of Political Science* 57(3): 624–642.

Grimmer, Justin, and Brandon M Stewart. 2013. “Text as data: The promise and pitfalls of automatic content analysis methods for political texts.” *Political Analysis* p. mps028.

Grimmer, Justin, and Gary King. 2011. “General purpose computer-assisted clustering and conceptualization.” *Proceedings of the National Academy of Sciences* 108(7): 2643–2650.

Hathaway, Oona A. 2002. “Do human rights treaties make a difference?” *The Yale Law Journal* 111(8): 1935–2042.

Henderson, John A. 2015. “Using experiments to improve ideal point estimation in text with an application to political ads.” Unpublished manuscript.

Hitler, Adolf. 2013. *Hitler’s second book: The unpublished sequel to Mein Kampf*. New York, NY: Enigma Books.

Hitlin, Paul. 2016. *Research in the crowdsourcing age, a case study*. `www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/`: Pew Research Center.

Honaker, James, Michael Berkman, Chris Ojeda, and Eric Plutzer. 2013. “Sorting algorithms for qualitative data to recover latent dimensions with crowdsourced judgments: Measuring state policies for welfare eligibility under TANF.” Paper at the 2013 meeting of the Society for Political Methodology.

Hopkins, Daniel J, and Gary King. 2010. “A method of automated nonparametric content analysis for social science.” *American Journal of Political Science* 54(1): 229–247.

Jamal, Amaney A., Robert O. Keohane, David Romney, and Dustin Tingley. 2015. “Anti-Americanism and anti-interventionism in Arabic Twitter discourses.” *Perspectives on Politics* 13(1): 55–73.

Key, V.O. 1949. *Southern politics in state and nation*. New York: A. Knopf.

Kim, In Song, John Londregan, and Marc Ratkovic. 2014. Voting, speechmaking, and the dimensions of conflict in the US Senate. In *Annual Meeting of the Midwest Political Science Association*.

King, Gary, Christopher JL Murray, Joshua A Salomon, and Ajay Tandon. 2004. “Enhancing the validity and cross-cultural comparability of measurement in survey research.” *American Political*

Science Review 98(01): 191–207.

Kingdon, John W. 1973. *Congressmen's voting decisions*. New York: Harper & Row.

Krippendorff, Klaus. 2013. *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage Publications.

Krosnick, Jon A. 1999. "Survey research." *Annual Review of Psychology* 50: 537–67.

Lauderdale, Benjamin, and Alexander Herzog. 2014. "Measuring political positions from legislative debate texts on heterogeneous topics." Working Paper.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(02): 311–331.

Lowe, Will, and Kenneth Benoit. 2013. "Validating estimates of latent traits from textual data using human judgment as a benchmark." *Political Analysis* 21(3): 267–297.

Lowe, Will, Ken Benoit, Slava Mihaylov, and M. Laver. 2011. "Scaling policy preferences from coded political texts." *Legislative Studies Quarterly* 36(1): 123–155.

Mikhaylov, Slava, Michael Laver, and Kenneth R Benoit. 2012. "Coder reliability and misclassification in the human coding of party manifestos." *Political Analysis* 20(1): 78–91.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4): 372–403.

Neumayer, Eric. 2005. "Do international human rights treaties improve respect for human rights?" *Journal of Conflict Resolution* 49(6): 925–953.

Oishi, Shigehiro, Jungwon Hahn, Ulrich Schimmack, Phanikiran Radhakrishan, Vivian Dzokoto, and Stephen Ahadi. 2005. "The measurement of values across cultures: A pairwise comparison approach." *Journal of Research in Personality* 39(2): 299–305.

Owens, Ryan J, and Justin Wedeking. 2012. "Predicting drift on politically insulated institutions: A study of ideological drift on the United States Supreme Court." *The Journal of Politics* 74(02): 487–500.

Pang, Bo, and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics pp. 115–124.

Quinn, Alexander J, and Benjamin B Bederson. 2011. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM pp. 1403–1412.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1): 209–228.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2016. "Navigating the local modes of big data." *Computational Social Science* p. 51.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4): 1064–1082.

Sheng, Victor S, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM pp. 614–622.

Slapin, Jonathan B, and Sven-Oliver Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52(3): 705–722.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics pp. 254–263.

Socher, Richard, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Vol. 1631 Citeseer p. 1642.

Von Ahn, Luis. 2009. "Human computation". In *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE*. IEEE pp. 418–419.

Von Ahn, Luis, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. "reCAPTCHA: Human-based character recognition via web security measures." *Science* 321(5895): 1465–1468.