Inferring Concepts from Topics:
Towards Procedures for Validating Topics as Measures

Luwei Ying*
Washington University in St. Louis

Jacob M. Montgomery
Washington University in St. Louis

Brandon M. Stewart
Princeton University

November 6, 2020

Abstract

Topic models, as developed in computer science, are effective tools for exploring and summarizing large document collections. When applied in social science, they are commonly used for measurement, a fundamentally different task that requires careful validation to ensure that the measure actually captures the desired quantity of interest. In this paper, we extend an existing crowd-sourcing method for validating topic quality developed in the computer science literature (Chang et al., 2009) and create new evaluations for validating human-generated topic labels. We illustrate our method with a novel analysis of Facebook posts by US Senators and provide software and guidance for researchers wishing to validate their own topic models. While tailored, case-specific validation exercises will always be best, we aim to provide general, easy-to-use tools to validate topics as measures.

*Corresponding author, luwei.ying@wustl.edu

1 Introduction

Many concepts in social science are not directly observable. If we wish to study democracy, culture, affect, or ideology, we need to develop a measurement strategy. How do we take observable data and use it to infer some unobserved trait of interest? Methods for handling this problem have varied markedly over time and across application areas. Congress scholars developed a number of tools for using roll-call behavior to infer member ideology, (e.g. Poole and Rosenthal, 2000; Poole, 2005; Clinton, Jackman and Rivers, 2004), survey researchers rely on tools such as Guttman scaling or factor analysis to infer latent traits such as 'tolerance' from survey responses (Gibson and Bingham, 1982), while network scholars use block structures, latent space models to infer communities from network connections (Goldberg et al., 2013; Minhas, Hoff and Ward, 2019) and pairwise competition models (Renshon and Spirling, 2015).

Recently scholars have increasingly turned towards text-as-data methods as a way to derive measures from written text, supplementing a long tradition of manual content analysis with semi-automated techniques. Unsupervised probabilistic topic models have emerged as a particularly popular strategy for analysis since their introduction to political science in Quinn et al. (2010). Topic models are attractive because they both discover a set of themes in the text and annotate documents with these themes. Due to their ease-of-use and scalability, these models have quickly become a common way of measuring key explanatory and outcome variables. Recent examples include inferring, the categories of legislative speeches (Dietrich, Hayes and O'brien, 2019), the degree to which candidates discuss particularistic policies (Catalinac, 2016), how policies are defined and framed (Gilardi, Shipan and Wueest, forthcoming), how international organizations allocate regulatory effort (Pratt, 2018), the policy emphasis of media outlets (Barnes and Hicks, 2018) and engagement in team communication (Horowitz et al., 2019).

However, utilizing topic models as a tool for measurement deviates sharply with their

initial use case: language modeling and dimension reduction.[1] Scholars in this measurement tradition have accordingly emphasized the necessity of robust validation of model results before their use (Quinn et al., 2010; Grimmer, 2010), with Grimmer and Stewart (2013) naming a key principle for text methods, "validate, validate, validate." But at the same time, the literature on how to validate a topic model is extremely sparse. Early work was excruciatingly careful to validate the substantive meaning of the topics (e.g. Quinn et al., 2010; Grimmer, 2010) through many carefully constructed *application-specific* criteria.[2] These bespoke methods for validation are effective but appear in the literature with increasing rarity.[3] To date, few general-purpose procedures have been proposed for validating the results of topic models in a transparent way and there is nothing like a general standard for reporting.

This situation is unsatisfactory. On the one hand, we have the ability to induce new organizations of text from immense collections of documents that previous generations could neither have collected nor analyzed. On the other hand, the validity of the findings from these studies rests entirely on our confidence in the authors' qualitative interpretations of the model outputs, most of which cannot be (or at least are not) reported. In many cases, no validations are reported to readers beyond providing the most probable words for each topic and statements implying that the resulting topics appeared coherent and meaningful under inspection. If the discipline is to rely on topic modeling for rigorous scientific inference, it is time to establish a more rigorous set of standards for validation.

In this paper, we design and test a suite of validation exercises and provide tools to make them easy to run. We extend and improve the prior work of Chang et al. (2009) in computer science which proposes crowd-sourced tasks for model selection using human

[1] In the original article outlining latent Dirichlet Allocation (LDA), Blei, Ng and Jordan (2003) primarily focus on information retrieval, document classification and collaborative filtering applications.

[2] For example, Grimmer (2010) shows in an analysis of US Senate press releases that senators talk more frequently about issues related to committees they chair. This is an intuitive evaluation that the topic model is able to detect something we are *ex ante* confident is true, but it does not straightforwardly generalize to other settings.

[3] We speculate that this is a consequence of the increased routinization of text analysis methods. With early innovations there is both a greater need to demonstrate validity and more space in which to do so. As tools become a regular part of the toolkit, they become less a focus of the article and thus don't permit the space or time to do extensive validation.

judgment. Chang et al. (2009) evaluates whether word sets learned by a topic model appear semantically related. We then design our own novel validation tasks to assess label quality. Together we hope that these validation exercises can provide a ready-to-use approach to validation for the vast majority of topic model users in the social sciences.

We build upon prior work in three ways. First, we provide a novel task structure for evaluating semantic coherence that workers can complete with far higher accuracy rates and which provides outputs on a more intuitive scale. We show that our proposed validation, *random 4 word set intrusion*, is more reliable, provides clearer discrimination between models, and is more consonant with alternative qualitative and quantitative measures of model fit. In the main text we compare our task structure to those developed in Chang et al. (2009), and in the Appendix we compare this method to several alternative variations we also considered.

Second, previous methods provide no way to assess whether a topic captures the substantive quantity implied by the researchers' topic label. That is, it assesses whether workers can see a relationship between words but not whether a topic actually relates to the concept claimed by the researcher. We bridge this gap by proposing additional validation tasks for the concept labels themselves.

Finally, while the basic idea of crowdsourced topic validation are relatively straightforward, actually executing a validation involves a number of practical challenges that can be daunting for applied scholars. In addition to creating and posting tasks online, researchers must recruit, train, and monitor workers to ensure the validity of the final output. We therefore provide free-to-use software to ease the process of topic validation as well as guidance and practical advice on how to improve (and assess) the final output.

In the next section we discuss the use of topic models in social science, review the general problem of measurement validation, and consider standard approaches for topic model validation in both the social sciences and computer science. We then offer some principles for designing an off-the-shelf validation exercises using crowd-sourced coding from non-experts informed by our extensive testing of various task structures. Next, we describe and test three

tasks for evaluating semantic model fit and two tasks for label validation using topic models fit to Facebook posts from US Senators. Our results show that our recommended tasks are easy for workers to complete and provide discrimination across models and potential labels. We also show that the evaluations are reliable and can be completed quickly and cheaply with human workers on Amazon's Mechanical Turk. We conclude with a discussion of the limitations of this kind of crowd-sourced validation design and the need for more research on topic validation.

2 The use of topics and prior strategies of evaluation

As text-as-data methods have grown in popularity, researchers have increasingly used topic models to capture latent concepts.[4] This is not surprising since topic models are a method of data reduction that were specifically designed to make interpretation if not easy, then at least possible. In introducing latent Dirichlet allocation models, Blei, Ng and Jordan (2003, p. 993) state, "The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks."

In the social sciences, researchers quickly uncovered the potential of topic models for measuring explanatory and dependent variables. The last decade has witnessed important work in all sub-fields where topic models measure latent traits including: senators' home styles in press releases (Grimmer, 2013), freedom of expression in human rights reports (Bagozzi and Berliner, 2018), religion in political discourse (Blaydes, Grimmer and McQueen, 2018), styles of radical rhetoric (Karell and Freedman, 2019) and more. In these works, the topic models identified useful latent traits anticipated by the authors, but in other cases, the models directed scholars towards new conceptualizations (Airoldi and Bischof, 2016;

[4]We do not offer a complete enumeration of articles employing this practice, but examples are not difficult to find (e.g. Al-Saggaf, 2016; Bagozzi, 2015; Bagozzi and Berliner, 2018; Bauer et al., 2017; Hayden et al., 2017; Lucas et al., 2015; DiMaggio, Nag and Blei, 2013; Roberts et al., 2014; Ryoo and Bendle, 2017; Nowlin, 2016; Terman, 2017; Velden, Schumacher and Vis, 2018).

Grimmer and King, 2011; Velden, Schumacher and Vis, 2018).

This trend is promising in that this approach opens up important new lines of inquiry — especially in the context of the explosion of new textual data sources online. At the same time it is worrying in the sense that we may be running ahead of ourselves. Do these topics measure what they are supposed to measure? How would we know? We lack an established standard for affirming that a topic measures a particular concept.[5]

2.1 The problem of topic validation

The strength and weakness of topic models is that topics are simultaneously learned and assigned to documents. Thus, the researchers must, first, infer whether or not there are *any* coherent topics, second, place a conceptual label on those topics, and only *then* assess whether that concept is measured well. In this more open-ended process the potential for creative interpretation is vastly expanded — with all of the advantages and disadvantages that brings. These concerns are magnified because people are extremely good at seeing patterns even where none exist (see e.g. Kalish, Griffiths and Lewandowsky, 2007).

Validity is a concern for any measurement in the social sciences, but the *post hoc* nature of topic model interpretation makes it especially pressing. A useful analogy here is with confirmatory (CFA) and exploratory factor analysis (EFA) in a survey setting. When assessing the validity of, for example, a newly proposed survey scale via CFA, we at least know that the survey items were developed with a specific concept in mind and can impose structure onto the model. Our problem is simply assessing whether the concept was measured well relative to our pre-determined target. To establish validity, researchers need to show that items load strongly on the underlying factor as expected and that the proposed model fits the data adequately.

[5]These issues are complicated by the explicitly confirmatory, hypothesis-testing style of most quantitative work in the social sciences and the relative undervaluing of exploratory or descriptive work. That is, the problem isn't with using text analysis to measure properties of text but rather that in published work we erode the difference between confirming an *ex-ante* hypothesis and a data-driven discovery. See, for example, the discussion in Egami et al. (2018).

Conceptually, CFA is closer to supervised learning, where the analyst designs a coding scheme (documenting it in a codebook). She then codes a large sample of documents according to that coding scheme, trains the learner, and then the algorithm annotates the remainder of the corpus. In this setting, the analyst has designed the measurement in order to fit with their particular argument and our primary validity concern is that the algorithm can annotate the documents with sufficiently high accuracy that the error is negligible.[6] Because we designed the measurement device, there is substantially more control over exactly what the latent categories are capturing. And if we believe in the originally proposed measurement scheme, the validity of the measurement is justified by high-predictive accuracy.

In contrast, topic models are more analogous to exploratory factor analysis (EFA), a method that has itself undergone significant criticism in the past.[7] Like EFA, topic models make specific assumptions about how the data is generated to estimate latent traits from the observed data. These latent topics must be interpreted and linked to specific concepts by examination of the documents and model outputs (e.g. high probability words). However, the interpretation and adequacy of the various "factors" or "topics" are not justified by the model fitting process—those motivating assumptions were simply *conveniences* not structural assumptions about the world to which we are committed. Instead, our confidence in the topics as measures comes from the validation exercises that come *after* the model is fit (Grimmer and Stewart, 2013). The nature of this interpretation makes it difficult for researchers to both complete this task in a replicable fashion or to justify their decisions and conclusions in a way that can be assessed by readers. In the language of King, Keohane and Verba (1994), the procedure is not public.

[6]This glosses over a number of problems common to supervised learning in practice. Notably, we have the same concerns about discovery and estimation being done on the same document set if the coding scheme is developed with the same corpus where the coding is done (see e.g. Egami et al., 2018) for a formalization. In practice, it also isn't immediately clear what to do with non-trivial classification error although proposals based on resampling (Stewart and Zhukov, 2009; Benoit, Laver and Mikhaylov, 2009) and analytic corrections are available.

[7]For a particularly sharp, if perhaps over-broad, criticism of EFA, see Armstrong (1967).

2.2 Validation practices for topic models

Since the procedure of turning a topic into a measure is not fully public, the need for validating the resulting numerical summary becomes even more important. Here we borrow from extant work in the measurement literature to classify validation exercises into three inter-related categories (Adcock and Collier, 2001; Bollen, 2014):

1. Content validity: Does the measure include indicators we would expect it to include?
2. Criterion validity: Does the measure relate to external events in ways we would expect?
3. Construct validity: Does the measure capture what it claims to measure?
    - Convergent validity: If various measures theoretically should be related, is this expectation met?
    - Divergent validity: If various measures theoretically should *not* be related, is this expectation met?

We can use this measurement perspective to evaluate common practices of validation in the applied social scientific topic modeling literature.[8] In our reading, current practices are strongest with respect to content validity. Most research based on topic models provide word clouds or top words that allow us to assess (if imperfectly) whether or not expected words are correlated with the concept as we might expect. If a topic is supposed to represent the European debt crisis, it is comforting to see that top words for the topic include word stems like: "eurozone", "bank", "crisi", "currenc", and "greec" (Barnes and Hicks, 2018). However, it is important to recognize that this is far from perfect or complete. Many articles utilizing topics as measures provide *only* word clouds or top words. However, this evidence is rarely clear-cut. So, for instance, we also see in the Euro/Debt crisis topic words like "year", "last", "auster", and "deficit". The first two words are at best ambiguous and the last two seem more associated with other topic labels (*Austerity Trade-Offs* and *Macro/Fiscal*) in the article (Barnes and Hicks, 2018).

[8]Note this isn't a novel perspective. In one of the earliest applications of topic modeling in the social sciences, Quinn et al. (2010) frame the exercise through the lens of measurement theory similarly to how we have above.

Some scholars show that topic frequencies vary as expected in the face of external events (e.g. Quinn et al., 2010), a practice that be conceptualized as establishing criterion validity. Thus, Greene and Cross (2017) show that the topics prevalances in speeches by members of the European Parliament respond in expected ways to exogenous external events such as the collapse of the Lehman Brothers bank (p. 88). This is useful, but do researchers start with external events and confirm expected trends, or do they observe trends and then identify external events that offer plausible explanations? Moreover, given how context-dependent this process is, it is not available for all models or even all topics in the same model.

Finally, some scholars validate topics by using research assistants to categorize documents based on pre-specified coding schemes and comparing the results (e.g. Grimmer and Stewart, 2013) or developing new coding schemes to evaluate a newly discovered concept (Grimmer and King, 2011). This practice aims to test construct validity since measures intended to capture the same underlying concept should be highly correlated.

Validation for topic models in computer science focuses much more heavily on predictive accuracy, usually assessed through some measure of held-out log likelihood (Wallach et al., 2009). This provides a sense of whether or not the model is over-fitting but provides little direct evidence that is capturing something of interest for making a particular argument.

Pushing beyond predictive accuracy in computer science, Mimno et al. (2011) and Newman et al. (2010) introduced surrogate measures for coherent and interpretable topics based on point-wise mutual information. These measures reward topics which have high probability words which frequently occur together in documents. Mimno et al. (2011) show that this metric, which they call "semantic coherence," correlates well with expert human judgments in a task analyzing grants submitted to the National Institute of Health. Roberts et al. (2014) extend this idea by suggesting that the model should jointly maximize exclusivity of words to individual topics as well as semantic coherence.

While maximizing model fit through the held-out log likelihood or surrogate criteria like semantic coherence may have correlated well with human judgment in the past, it is difficult

to know how this criterion will generalize to future settings. Some researchers have moved to involve external coders to obtain indicators of topic model qualities at the aggregated level. Airoldi and Bischof (2016) and Newman et al. (2010), for example, have directly asked people to rate the coherence of the learned topics. Others have focused on finding better ways communicate the results of topic models graphically so readers might judge for themselves (Chuang, Manning and Heer, 2012; Chaney and Blei, 2012; Sievert and Shirley, 2014; Freeman et al., 2015). The challenge is that expert judgment does not scale and it is not obvious how to make use of these methods with non-expert judges.

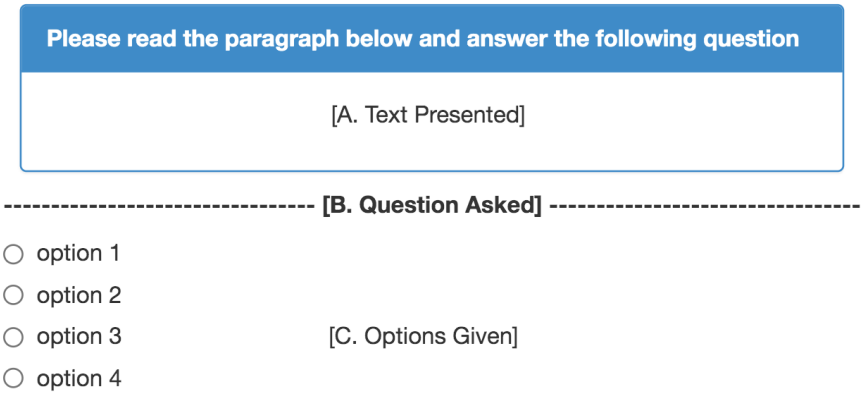
In short, there is no clear-cut, general procedure for topic model validation.

2.3 Using the wisdom of the crowds

In an agenda-setting piece of work, Chang et al. (2009) introduced a set of crowd-sourced tasks for evaluating topic models.[9] The core idea is to transform the validation task into a set of two games which, if they can be completed with high accuracy, would imply the quality of the evaluated topic model.

The common structure for both tasks is shown in Figure 1. In each, a specific question

Figure 1: A Diagram for the Common Structure of Crowd-Sourced Validation Tasks



[9]This has been followed up in Lau et al. (2011) and Lund et al. (2019). In political science, Lowe and Benoit (2013) used an innovative crowd-sourcing task design for assessing the validity of a scaling measure which can address other unsupervised text analysis designs like those based on Wordfish (Slapin and Proksch, 2008) or more advanced approaches (Spirling, 2012).

(B) is presented to the coders and they must choose from several options (C). For some tasks, additional information is provided above the question (A).

The first task in Chang et al. (2009), *Word Intrusion* (WI), is designed to detect topics which are semantically cohesive in the sense that they identify a well-defined concept. Workers are presented with five words[10] and asked to identify the "most irrelevant" word — the intruder. Four words are chosen randomly from the highest probability words related to a topic and the remaining "intruder" is chosen from high probability words from a different topic (so it is not distinguishable from rarity alone).

The second task, what we call *Top 8 Word Set Intrusion* (T8WSI), is more complex.[11] The coder is presented with an actual document (or snippet from the document) and asked to identify a word set that does not belong. The coder must choose from four word sets where each set is the eight highest probability words for a topic, three of the word sets represent the high probability topics for the shown document, and one of the word sets is from a low probability topic (that is, a topic not associated with the the document).

In their study, Chang et al. (2009) find that the models with the best performance are not consistently those with the best model fit statistics (such as held-out log likelihood). This crowd-sourced approach to validation has become well-known, now accumulating over 1,900 Google Scholar citations, and spawning attempts at automation (Lau, Newman and Baldwin, 2014).[12] In the language of measurement models, we think of these tasks as testing convergent validity (words constructing the same concept are supposed to correlate) and discriminant validity (words constructing the different concepts are supposed to diverge) of the proposed latent concepts.

This crowd-sourced design is a way of showing that the topics make sense on their face but social scientists also need to establish whether the topics correspond to stated latent

[10]Chang et al. (2009) presented tasks with six words, but we found that five words resulted in more reliable coding.

[11]Chang et al. (2009) call this Topic Intrusion but we have given it a more descriptive name.

[12]While the work has been heavily cited, it isn't clear that it is heavily used. It is often cited for the observation that expert judgment does not always correlate with model fit.

concepts. If we wish to use the outputs from topic models as a measure of some concept, it is not enough to show that "there is a *there* there." We must also show that the conceptual labels are themselves valid. In the next sections, therefore, we improve and extend these evaluations for measurement validation.

3 Designing and assessing an off-the-shelf evaluation

In this paper, we pursue the goal of designing an off-the-shelf evaluation design for topic models. We develop two classes of designs: one oriented towards model selection which extends the intrusion tasks of Chang et al. (2009) to evaluate the semantic coherence of a given topic model, and a second oriented towards validating that a given topic corresponds to its label. Before we present our method, however, in Section 3.1 we offer some basic design principles and in Section 3.2 we describe the data that we use to assess the designs.

3.1 Principles

In developing the task, we want designs which are (1) generalizable, (2) discriminative, (3) easy-to-use, and (4) reliable. In this paper we use the structural topic model (Roberts et al., 2013; Roberts, Stewart and Airoldi, 2016) to produce our topic models, but we want to find designs that are *generalizable* to a wide variety of mixed-membership and single-membership topic models. The designs we present below should work for any mixed-membership topic model that uses a multinomial distribution over words to represent a topic and several will also work for single-membership models. The generalizable principle also reflects our desire to have evaluations that work in a variety of different substantive settings, with different size document collections, document lengths, and numbers of topics.

The tasks also need to be *discriminative*. Our task structures follow the general design of Chang et al. (2009) which have the form of games. These games need to be of medium difficulty because if they are too easy or hard, ceiling and floor effects (respectively) will

hinder our ability to discriminate across different models. Indeed, our proposed extensions to the Chang et al. (2009) designs are motivated by a desire to make the tasks more possible to do and thus more discriminative. Better discrimination in turn leads to better information about model selection.

To be deployed in practice, the tasks need to be *easy-to-use*. All the tasks are designed to be completed by Mechanical Turk workers quickly. Along with the paper we will provide implementations using our R package so that the tasks can be run quickly and cheaply.[13]

Finally, we demonstrate that our tasks are *reliable*. Despite the fact that these tasks involve inherently subjective decisions, we show across a variety of tasks and topics that the results are surprisingly stable under replication.

We divide the validation tasks into two parts. Model selection (Section 4) establishes that the model is semantically coherent in the sense of the prior work by Chang et al. (2009). A model should pass this task if the topics are immediately recognizable to a human evaluator as distinct concepts. Label validation (Section 5) goes further to assess whether or not a set of analyst-provided labels correspond with the contents of the model and the documents. We envision the model selection tasks as being most useful at early stages when making choices among competing models and the label validation tasks as helping to verify that the topic model is measuring what it is intending to measure (as expressed by the label).

3.2 Empirical Illustration

To illustrate and assess our method, we rely on topic models fit to a novel dataset relevant to political science. The corpus comes from US senators' Facebook pages from the 115th Congress.[14] We scraped every individual post from April 2018 back to when each page was initially created.[15] For text pre-processing, we removed all numbers, punctuation marks,

[13]The R package, `validateIt`, is being prepared for submission to the *Comprehensive R Archive Network* (CRAN). For a detailed user manual, see the Appendix.

[14]This was the Senate as composed in July 2017. Three Senators did not have public Facebook pages.

[15]The end date occurs when Facebook implemented changes to their Graph API. The earliest date of a post is September 2007.

and stopwords in the SMART stopword list. Additionally, we made a customized stopword list with state names (full or partial), state abbreviations, and the words such as "sen" and "senator" which are ubiquitous in senators' public pages. We converted all words to lower cases but did not stem them. Finally, we removed those posts in non-English, about life events (e.g. "XX added a life event."), and those shorter than 10 words. We fit five structural topic models using the remaining 163,642 documents:[16]

1. a model with 10 topics limited to one EM iteration (which keeps the model from properly converging);
2. a model with 10 topics;
3. a model with 50 topics;
4. a model with 100 topics, and;
5. a model with 500 topics.

The first model, which is a topic model that has not been allowed to converge, is a baseline that we use to assess whether or not the tasks can clearly identify a flawed model. Note that even this model appears reasonable on first glance because of the initialization procedure in STM.[17] The other four models provide different feasible options for analyzing this corpus that we might want to consider in practice and we have no *ex ante* preference between them. Additional information about these topic model fits are provided in the rest of the paper and some additional information is listed in the Appendix.

4 Model selection using coherence evaluations

We considered three task structures for evaluating the semantic coherence of topic model fits. Topic models that perform well on these tasks are those where the topics pick out sharply

[16] We randomly select 10% of the documents (16,364) and held out 50% of the tokens in these documents so later we will be able to compare the results from our methods with held-out log likelihood.

[17] By the default the `stm` package (Roberts, Stewart and Tingley, 2019) uses a spectral method of moments (Arora et al., 2013) initialization strategy. Roberts, Stewart and Tingley (2016) show that it is a highly effective initialization strategy, but Arora et al. (2013) show that it has strong performance in its own right. Thus this is a relatively strong baseline that could have proven challenging to detect.

defined topics which are distinctive from each other. The task structures are summarized in Table 1, where the column names correspond to the annotated slots in Figure 1. The first two, the Word Intrusion (WI) and the Top 8 Word Set Intrusion (T8WSI) tasks are slight alterations from the original methods from Chang et al. (2009) described above. We combine mass for words with the same root to the most frequent form and then randomly draw words based on their mass. This operation is equivalent to stemming in the pre-processing stage, but showing online workers the words in their complete form and in their most common form in each topic.

Table 1: Task Structures for Coherence Evaluations

	A. Text Presented	B. Question Asked	C. Options Given
WI	NA	Please read the five words below, and choose one that is most IRRELEVANT to the other four.	Four words mass-based selected from the top twenty high-probability words of one topic and one word (the intruder) mass-based selected from the top twenty high-probability words of another topic
T8WSI	A randomly selected document	After reading the above passage, please click on the set of words below that is most UNRELATED to passage.	Three word sets (each containing the top eight high-probability words) from the top three high-probability topics and one word set (the intruder) from another topic
R4WSI	NA	Please click on the word set below that is most UNRELATED to the other three.	Three word sets (each containing four mass-based selected words) from the top twenty high-probability words of one topic and one word set (the intruder) mass-based selected from the top twenty high-probability words of another topic

4.1 Novel task structure

We also designed and tested one additional task to address concerns about the Chang et al. (2009) tasks. Initial work indicated that coders found the WI and T8WSI tasks to be so difficult that the results could be uninformative. Further, we worried that the T8WSI results were too sensitive to the specific words included in the top 8 word sets for each topic, making

the results somewhat arbitrary and again less informative.[18] Our aim was to produce a task structure that would make the task easier for coders to complete and would be less sensitive to the words that happen to fall in the top eight in order to increase our ability to discriminate between models. This new task is summarized in the final rows of Table 1.[19]

This *Random 4 Word Set Intrusion* (R4WSI) task asks workers to identify the intruder word set among four different word sets. Structurally, this task is similar to the WI task. However, we provide more words in each option by randomly choosing four words from the top 20 words of the topic, based on their mass and preventing repetition. Interacting with more words also makes RSWSI advantageous over T8WSI because for some topics the highest eight words were hard to interpret while their semantic relationship was perfectly clear when looking at 12 of the top 20 words. The end result, we argue, is coder decisions that are more informative about the quality of the underlying model leading to better discrimination.[20]

4.2 Results

We tested these three task structures using workers with "master" certifications from Amazon's Mechanical Turk (AMT) from March to July, 2020. To qualify to complete tasks, workers had to complete an online training module.[21] These modules were designed to explain the task, provide some background about the document set, and walk workers through examples to ensure they understood the goal. Workers were paid \$0.04/task for WI, \$0.08/task for T8WSI, and \$0.06 for R4WSI.[22]

For each task structure we posted 500 tasks for all five models. These tasks were often completed in hours. To assess the consistency of task structures, we then posted these *exact same tasks* again. To monitor the quality of the work, we randomly mixed in a gold-standard

[18] As we will discuss in Section 6 this also makes the task particularly sensitive to the pre-processing choices making it difficult to adjudicate between different specifications (see also Denny and Spirling (2018)).

[19] In all tasks but T8WSI, we ensure that each topic from a given model is represented equally.

[20] An additional advantage is that this structure is valid for topic models that assign documents to only one topic, making it more generalizable.

[21] Additional details are in the Appendix.

[22] We note that these payment rates are actually somewhat higher than most tasks available on AMT for the amount of time they take.

HIT whose answer is less ambiguous every ten HITs.[23] Thus, in total workers completed 16,500 tasks.

Figure 2 presents the results. All task structures easily identified the non-converged baseline model as the worst, which provides a check that the test has the ability to identify a model known to be a relatively poor fit. All of them are able to identify over-fitting as the 500-topic model appears to be worse than the 100-topic model in all task structures, our novel task structure penalizes over-fitting to the most extent. The estimated held-out log likelihood for Models 1-5, respectively, are -8.315913, -7.981415, -7.767319, -7.70538, and -7.983551. This order ranking is consistent with that suggested by R4WSI, but not WI and T8WSI.

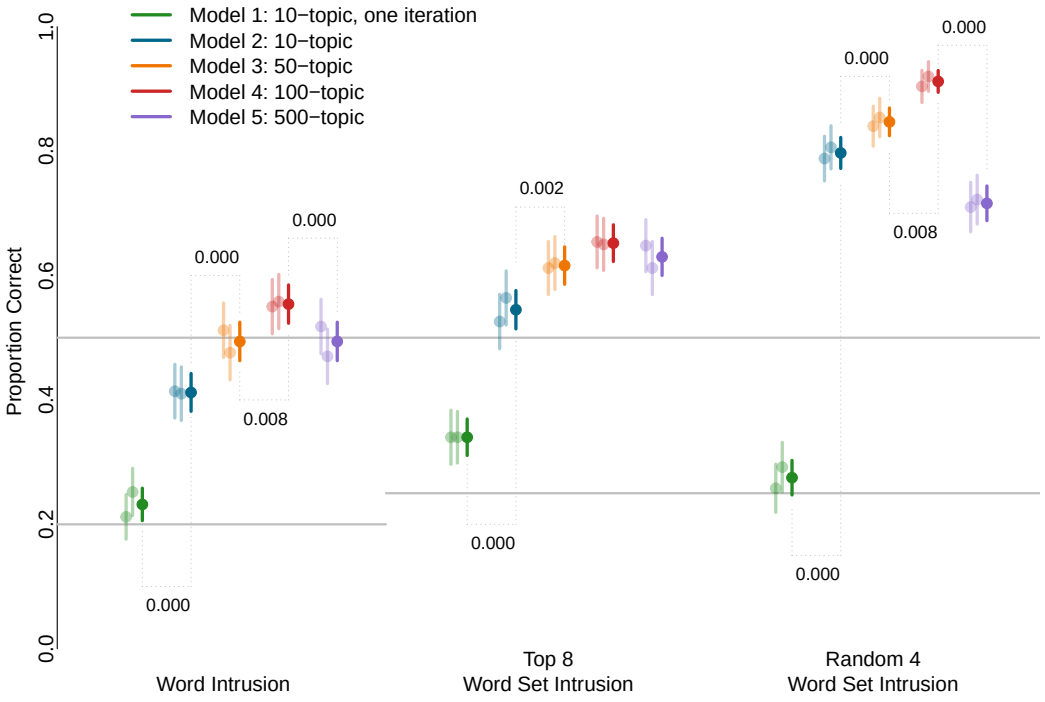
5 Label Validation

Using our semantic coherence evaluations above, we selected Model 4 (estimated with 100 topics) as having the best fit. In practice a researcher would now need to place conceptual labels on the topics, essentially committing to what they are claiming the topics each measure. This process is inherently qualitative, requiring researchers to consider top words, read representative documents, and understand the general context of the document set. Once the researcher assigns a label to a topic, e.g. "Income Inequality", the reader needs a way of assessing whether or not the topic is measuring the concept implied by the label. Unfortunately the best assessment, carefully reading many documents, is not easy to translate into an article and so in this section we try to design alternative checks on label quality. None of the previous methods (such as those in Chang et al. (2009)) use human generated labels and thus they are unable to assess this component of measurement validity.[24]

[23] We suppressed the qualification of workers who have missed more than 2 gold-standard HITs or who have done a relatively large number of HITs of a specific task structure. This operation has no negative impact on their Mturk records. We have rejected and replaced work from two workers (267 HITs in total) who missed more than 4 HITs each.

[24] Nielsen (2017) provides an excellent example of how to build confidence in a set of text-based measurements by triangulating many pieces of evidence. However, this is only possible as he is able to devote large

Figure 2: Results for Coherence Evaluations



Note: The 95% confidence intervals are presented. The two transparent bars represent two identical trails (500 HITs each). The solid bar represents the pooled result (1000 HITs). When two models yield significantly different results, the p-value is noted. (Significance tests are difference in proportions as calculated by the `prop.test` function in R.) No identical trails (two transparent bars) are significantly different from each other.

Here we specify two use-cases, one asking "Can we usefully distinguish two broad categories of discussion?" and the other "Can we make more fine-grained distinctions between related categories?" Correspondingly, we provide one design focused only on major "bread and butter" domestic policy areas, e.g. Economy, healthcare, education, and a second design where we mixed in a second broad category of labels related to international/military topics.

Ideally, our task will allow us to discriminate between good and bad labels. To assess whether or not our task structure is working then, we need an example of what "good" and "bad" labels are. To address this, members of our research team, independently, label each of the 100 topics. Each of us carefully read the high-probability words and frequent & exclusive words (FREX) (Roberts, Stewart and Airoldi, 2016), as well as fifty representative documents per topic (Grimmer and Stewart, 2013). From the topics that all of us deemed as coherent, we picked ten domestic topics and ten international/military topics where the labels were most consistent. The final labels for each are shown in Table 2.[25] We refer to these labels as the "Careful Coder" labels.

To provide a contrast, we asked some graduate students in political science to create a different set of labels based on the high probability and FREX word output from the `stm` package (Roberts, Stewart and Tingley, 2019). These labels, which we refer to as "Cursory Coder" labels, are shown in the second column of Table 2. Note that conceptually these two label sets are not dramatically different (e.g. Agriculture vs. Farm Bill). We present them in random order to a different coder, asking whom to identify the better label. The coder picked 19 out of 20 labels from the Careful Coder.

5.1 Novel task structures

We imagine that a researcher wishes to validate *only* the ten domestic topics rather than the complete set of topic labels in the model. Thus, we focus only on the construct validity of

chunks of a book to this task. Our goal here is to provide a piece of evidence that can be presented concisely.

[25]High probability words and high probability documents for each of these ten topics are also provided in the Appendix.

Table 2: Labels to Validate

Careful Coder	**Cursory Coder**
Domestic Topics	
Equal Pay for Women	Working Class
Healthcare/Reproductive Rights	Planned Parenthood
Agriculture	Farm Bill
Student Loan/Debt	Economy
Drug Abuse	Prescription Medicine
Higher Education/Job Training	Grants for Colleges
Wall Street/Financial Sector	Banking
Government Shutdown/Congressional Budget	Government Spending
Obamacare/Tax Policy	Healthcare
Deficits/Debt/Budget	Debt Ceiling
International/Military Topics	
International Trade	Manufacturing
Praising Active Military/Military Units	"Welcome Home" Messages
Terrorism	Islamic Extremists
Military Sexual Assault	Military Affairs
Nuclear Deterrence/International Security	Foreign Affairs
Air Force	Military
Honoring Specific Veterans	Military Service
Honoring Veterans/Heroes	"Thank you" Messages
Military Operations/Armed Conflicts	Counter-terrorism
Veterans Affairs/Veterans Healthcare	Veterans

these ten. However, the task structures we present here could easily be extended to include all labels of interest depending on the substantive question.[26]

To validate topic labels, we designed two alternative task structures summarized in Table 3: *Label Intrusion* and *Optimal Label*. Note that for this second task, we explored two alternative approaches for generating intruding labels (discussed below). These tasks, intended to provide evidence of construct validity, were again designed to be generalizable, discrimative, easy-to-use, and reliable.

First, we considered a *Label Intrusion* (LI) task where the coder is shown a text and asked to identify a label that does not apply. Three of the labels come from topics highly associated with the document (the top three topics for that document) and one is selected from the remaining seven labels ("Within Category") or seven plus the ten international labels ("Across Category"). This structure intentionally mimics the word set intrusion design above.

The second task has the same basic layout, but the goal is not to find the intruder but rather the "best" label for the document. *Optimal Label* (OL) presents a document and four labels. One label is for the highest probability topic and the other three labels are chosen randomly from the remaining nine domestic labels ("Within Category") or nine plus the ten international labels ("Across Category").

We were attracted to the "optimal" label task structure because it is similar to the validation exercises already common in the literature where research assistants are asked to divide documents into predefined categories to assess topic quality (Grimmer, 2013). Further we expected this task to generally be easier for coders to complete. Finally, this task structure has the advantage of being the most directly interpretable since it essentially asks coders to confirm or refute the conceptual labels assigned to the documents.

In addition, we anticipated that discriminating between only domestic topics would be

[26]The Appendix includes additional information about two additional task structures that were focused on word sets rather than documents. We found these task structures to be unsuitable for differentiating between strong and weak label sets.

Table 3: Task Structures for Label Validation

	A. Text Presented	B. Question Asked	C. Options Given
LI	A randomly selected document[a]	Please read the four labels below and click on the label that is most UNRELATED to the passage.	*Within Category:* Three labels for the top three high-probability topics and one label for other domestic topics; *Across Categories:* Three labels for the top three high-probability topics and one label for other domestic or international/military topics
OL	A randomly selected document[b]	Please read the four labels below and click on the label that BEST summarizes the passage.	*Within Category:* One label for the highest-probability topic and three labels for other domestic topics; *Across Categories:* One label for the highest-probability topic and three labels for other domestic or international/military topics

[a]Top three predicted topics among the ten domestic topics.
[b]Top one predicted topic among the ten domestic topics.

harder than discriminating between domestic and international topics. That is, discriminating between conceptually similar topics (e.g. Drug Abuse vs. Healthcare/Reproductive Rights) is understandably a "harder test" than discriminating between clearly distinct topics (e.g. Drug Abuse vs. Terrorism). Which approach is better will depend on specific researcher needs, but we tested both variants of both task structures to confirm this intuition.

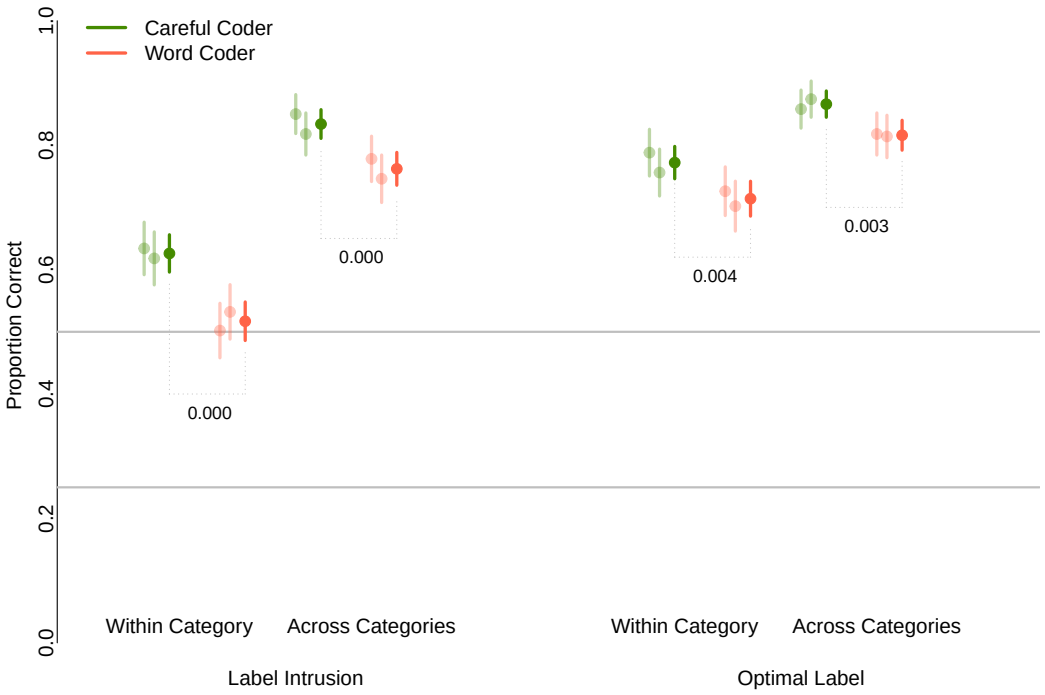
5.2 Results

To test these task structures we followed the same basic procedures discussed above. For each task/coder combination we created 500 tasks (plus 50 gold-standard HITs for evaluation purpose) that were coded by trained workers on AMT.[27] These were then repeated so that we could assess worker quality and replace work from low-quality workers. In total, workers completed 8,800 high quality HITS. The results are shown in Figure 3.

First, the results across runs seem fairly reliable with rank orderings of the label sets being indistinguishable across repetitions. Second, the results are consistent across task structures in identifying the Careful Coder labels as being superior. However, there is some

[27]Workers were paid 0.08/HIT.

Figure 3: Results for Label Validation



Note: The 95% confidence intervals are presented, where the two transparent bars represent two identical trails (500 HITs each) and the solid bar represents the pooled result (1000 HITs). P-values are based on the pooled set of tasks based on a difference-in-proportions test. No identical trails are significantly different from each other.

heterogeneity in the magnitude of the difference.[28] In Table 4, we note that workers achieve much higher correct rate when the discrimination is across the two broad categories.

In all, based on these results we would recommend that labels be evaluated using some form of the Label Intrusion task and the Optimal Label task. Whether it is better to draw intruder labels from a set of highly trusted and conceptually related topics or from across broad categories will depend on the purposes of the researcher. The closely related topics represent a harder test, but this in turn may artificially lower the number of correct responses and make fine-grained distinctions more difficult.

[28]See the Appendix for two additional task structures where this pattern did not hold. These tasks were based on word sets instead of documents, which significantly improved the perceived performance of the "Cursory Coder" labels.

Table 4: Accuracy within and across Two Broader Categories

Label Intrusion	Within	Across[a]		
Careful Coder	0.703	0.928		
Cursory Coder	0.490	0.939		
Optimal Label	Within	1 Across[b]	2 Across	3 Across
Careful Coder	0.788	0.816	0.896	0.964
Cursory Coder	0.717	0.788	0.835	0.918

[a]The intruder is one of the ten international labels
[b]One of the three intruders is one of the ten international labels

6 Limitations and Future Directions

The task designs we offer are not a replacement for other validation exercises such as careful reading of the texts or demonstrating agreement with alternative supervised or hand-coding measures. Their central advantage is that they are low cost, reliable, and easy to communicate to a reader. The difficulty of alternate strategies, particularly careful reading of the texts, is that the way they appear in the final article amounts to a statement that says "trust me, I checked this." For any given application, custom designed solutions will likely be superior, but our tasks provide something that researchers can reach for in most circumstances.

As we stated at the outset our goal here is not to present the final word on this methodological question, but rather to begin a dialogue about how and when it is appropriate to make inferences about latent concepts from topic models. Towards that end, we conclude with a discussion of the limitations of this approach and specific areas where future research may make improvements. We begin this discussion by reconsidering our four principles of generalizability, discrimination, ease-of-use and reliability.

Generalizability The validations we consider have several built-in assumptions that limit generalizability. First, the documents have to be *accessible* to the workers who are completing

the tasks. This means that documents have to be in a language the workers can read,[29] short enough to be readable[30], and require little background to understand. The analyst must also be allowed to post them in a semi-public way.[31] For some these concerns will inhibit the use of our designs, but they would not preclude most designs that have been published in the literature thus far.

A more subtle consideration is that basing the representation on a fixed number (e.g. 20) of the most probable words can present challenges in certain model fits. Topic models can have very sparse distributions over the vocabulary, particularly with large number of topics, large vocabularies or when fit with collapsed Gibbs sampling. If the topic is too sparse, the later words in the top twenty might have close to zero probability, making the words essentially random. If stop words are not removed, the vocabulary can include high frequency words which are probable under all topics and thus also not informative of the topic.[32] This is another instance of text pre-processing decisions playing a consequential role in unsupervised learning (Denny and Spirling, 2018). In our setting, it is straightforward to apply these steps after the model has been run just for the purposes of the validation tasks.

Discrimination The tasks were largely able to discriminate between different options (either models or label sets) but the challenge throughout this paper is that we don't have access to a ground truth. That is, we can see that the tasks discriminate among options but we have only circumstantial evidence that they discriminate *correctly*. The nature of measurement validity is that there is likely no way to actually do this kind of discrimination in the abstract, but we believe that these tasks are a useful part of a broader assessment.

[29]See Benoit et al. (2016) for a discussion of crowdsourced coding using international coders in multiple languages via the CrowdFlower system. AMT relies primarily on a US-based workforce since workers must have a US bank account to participate.

[30]We can use excerpts for long documents, but this similarly implies that a short summary can capture the gist of the document.

[31]See e.g. Romney, Stewart and Tingley (2015) on data access issues and intellectual property restrictions as limitations to transparency in statistical text analysis.

[32]There are also some concerns that may arise when not stemming or lemmatization as some word lists will be uninformative if they include many variants on the same word (e.g. `love`, `loves` and `loved`). This can also make the word set intrusion task trivially easy in some cases if multiple versions of the same word appear across different word sets (thus ruling them out as the intruder).

The coherence evaluations help to ensure that the topics convey a clear concept and are distinguishable from each other while the label validation exercises ensure that the researcher-assigned labels are at least somewhat accurate.[33]

Easy-to-use The tasks are easy and relatively cheap to deploy using our R package. While not as simple as statistics which can be easily calculated from the model, they are about as straightforward to implement as a human task is likely to be. These evaluations were all completed in less than three days and sometimes in only a few hours. Further, while certainly not free, the 500 task runs we used here are fairly affordable with costs ranging between $20 and $40.

An important easy-to-use limitation that we have not yet addressed is the difficulty of interpreting the results in isolation. Above, we focus on the relative accuracy of the tasks across models or label sets in large part because it is not clear exactly what the accuracy levels themselves mean. For example, Model 3 scores 61.6% on the top 8 word set intrusion task. Is this good or bad? Is it comparable to performance on a completely different data set? Documents which involve more complex material or technical vocabularies may lead to poorer scores not because the models are worse, but simply because the task is inherently harder. Readers may naturally want to assess some cut-off heuristic where models or labels that score below a particular threshold are not acceptable for publication. We note that this would be problematic and would fall into many of the traps that bedevil the debate over p-values. Finding the right way to compare evidence across datasets remains an open challenge. Authors will need to provide readers with context for evaluating and interpreting these numbers, preferably by evaluating multiple models using multiple validation methods.

[33]The tension arising from the lack of a ground truth is present in early parts of the literature as well. Chang et al. (2009) simply assert that their task designs select the most "semantically meaningful" topic models, but do not have any empirical evidence for that claim. More problematically, it isn't clear what empirical evidence for this claim could look like. Probably the closest analog would be using the judgment of subject matter experts as in Grimmer and King (2011) (two teams of political scientists) and Mimno et al. (2011) (NIH staff members). This kind of evidence is very costly to collect and the experience in specific applications does not necessarily generalize. The design as presented rests on the argument that being able to pass these tests is a reasonable consequence of a semantically coherent model.

With that said, random guessing would lead to a 25% correct rate for tasks with four choices. A minimal standard would be that coders should be able to substantially exceed this number. As we accumulate more evidence about such validation exercises, however, it may become possible to get a better sense of what an "adequate" score will be.

Reliability The task designs replicate across runs using the same population of Mechanical Turkers.[34] Our software helps to ensure that future iterations of the task would appear in the same way. With that said, researchers must take care to adequately train and screen workers, monitor data quality, and watch out for low quality workers who might flood researchers with low quality tasks.[35]

One final limitation is worth emphasizing. These tasks will not evaluate all properties we would need to see in a measurement. For example, many researchers use topics as outcomes in a regression. When estimating a conditional expectation, we want to know not only that the label is associated with the topic loadings but that they are proper interval scales (so that the mean is meaningful). These validation designs do nothing to assess these properties, and further work is needed to establish under what circumstances topic probabilities can be used as interval estimates of latent traits.

7 Conclusion

One reason the text as data movement is exciting is because it comes with a rapidly expanding evidence base in the social sciences (King, 2009). The conventional sources of evidence such as large surveys, summaries of voting records or economic data are giving way to individual study-specific datasets collected using text analysis techniques. This means that increasingly individual scholars are taking on the role of designing unique measurements for their study.

[34]Information on agreement rates for the same prompt across two different batches of workers are included in the appendix and are generally quite high.

[35]Anecdotally, we have found that worker quality is higher during normal work hours in US time zones.

Because we can't assume that these measurements will be used, examined and tested across many studies (e.g. as has happened with NOMINATE scores), it becomes imperative to develop best practices for regular diagnostic tests that researchers can run quickly on their own measurements and convey quickly.

We have offered a first step in this direction by improving upon the existing crowd-sourced designs of Chang et al. (2009) and extending them to create new designs that assess how well a label represents a corresponding topic. We tested these task structures using a novel topic model fit to Facebook posts by US Senators, and provided evidence that the method is reliable and allows for discrimination between models, based on semantic coherence, and labels, based on their conceptual appropriateness for specific documents. These kinds of crowd-sourced judgments allow us to capture human judgment and the linguistic knowledge that comes with it to examine our models, without experiencing the scale issues of relying on experts. The tasks are quick to complete and relatively inexpensive. They have relatively little cost in terms of the analyst's time.These designs will offer a fruitful space for innovation, however, and our collective work on validating topics as measures is just getting started.

References

Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

Airoldi, Edoardo M. and Jonathan M. Bischof. 2016. "Improving and Evaluating Topic Models and Other Models of Text." *Journal of the American Statistical Association* 111(516):1381–1403.

Al-Saggaf, Yeslam. 2016. "Understanding Online Radicalisation Using Data Science." *International Journal of Cyber Warfare and Terrorism* 6(4):13–27.

Armstrong, J. Scott. 1967. "Derivation of Theory by Means of Factor Analysis or Tom Swift and His Electric Factor Analysis Machine." *The American Statistician* 21(5):17–21.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu and Michael Zhu. 2013. A Practical Algorithm for Topic Modeling with Provable Guarantees. In *International Conference on Machine Learning*. pp. 280–288.

Bagozzi, Benjamin E. 2015. "The Multifaceted Nature of Global Climate Change Negotiations." *Review of International Organizations* 10(4):439–464.

Bagozzi, Benjamin E. and Daniel Berliner. 2018. "The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of US State Department Human Rights Reports." *Political Science Research and Methods* 6(4):661–677.

Barnes, Lucy and Timothy Hicks. 2018. "Making Austerity Popular: The Media and Mass Attitudes toward Fiscal Policy." *American Journal of Political Science* 62(2):340–354.

Bauer, Paul C., Pablo Barberá, Kathrin Ackermann and Aaron Venetz. 2017. "Is the Left-Right Scale a Valid Measure of Ideology?: Individual-Level Variation in Associations with "Left" and "Right" and Left-Right Self-Placement." *Political Behavior* 39(3):553–583.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–295.

Benoit, Kenneth, Michael Laver and Slava Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2):495–513.

Blaydes, Lisa, Justin Grimmer and Alison McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *Journal of Politics* 80(4):1150–1167.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022.

Bollen, Kenneth A. 2014. Measurement Models: The Relation between Latent and Observed Variables. In *Structural Equations with Latent Variables*. John Wiley & Sons, Ltd pp. 179–225.

Catalinac, Amy. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *Journal of Politics* 78(1):1–18.

Chaney, Allison June-Barlow and David M. Blei. 2012. Visualizing Topic Models. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*. pp. 288–296.

Chuang, Jason, Christopher D. Manning and Jeffrey Heer. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM pp. 74–77.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

Denny, Matthew J. and Arthur 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.

Dietrich, Bryce J., Matthew Hayes and Diana Z. O'brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review* 113(4):941–962.

DiMaggio, Paul, Manish Nag and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding." *Poetics* 41(6):570–606.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts." *arXiv preprint arXiv:1802.02163* .

Freeman, M. K., J. Chuang, M. E. Roberts, B. M. Stewart and D. Tingley. 2015. "stmBrowser: Structural Topic Model Browser.".

Gibson, James L. and Richard D. Bingham. 1982. "On the Conceptualization and Measurement of Political Tolerance." *The American Political Science Review* 76(3):603–620.

Gilardi, Fabrizio, Charles R Shipan and Bruno Wueest. forthcoming. "Policy Diffusion: The Issue-definition Stage." *American Journal of Political Science* .

Goldberg, Mitchell D., Heather Kilcoyne, Harry Cikanek and Ajay Mehta. 2013. "Joint Polar Satellite System: The United States next Generation Civilian Polar-Orbiting Environmental Satellite System." *Journal of Geophysical Research: Atmospheres* 118(24):13–463.

Greene, Derek and James P. Cross. 2017. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *Political Analysis* 25(1):77–94.

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* 21(3):267–297.

Grimmer, Justin and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Hayden, Jessica M., Matthew J. Geras, Nathan M. Gerth and Michael H. Crespin. 2017. "Land, Wood, Water, and Space: Senator Robert S. Kerr, Congress, and Selling the Space Race to the American Public." *Social Science Quarterly* 98(4):1189–1203.

Horowitz, Michael, Brandon M Stewart, Dustin Tingley, Michael Bishop, Laura Resnick Samotin, Margaret Roberts, Welton Chang, Barbara Mellers and Philip Tetlock. 2019. "What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting." *The Journal of Politics* 81(4):1388–1404.

Kalish, Michael L., Thomas L. Griffiths and Stephan Lewandowsky. 2007. "Iterated Learning: Intergenerational Knowledge Transmission Reveals Inductive Biases." *Psychonomic Bulletin & Review* 14(2):288–294.

Karell, Daniel and Michael Raphael Freedman. 2019. "Rhetorics of Radicalism." *American Sociological Review* .

King, Gary. 2009. The Changing Evidence Base of Social Science Research. Routedge pp. 91–93.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.

Lau, Jey Han, David Newman and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 530–539.

Lau, Jey Han, Karl Grieser, David Newman and Timothy Baldwin. 2011. Automatic Labelling of Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics pp. 1536–1545.

Lowe, Will and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2):254–277.

Lund, Jeffrey, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber and Kevin Seppi. 2019. "Automatic Evaluation of Local Topic Quality.".

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics pp. 262–272.

Minhas, Shahryar, Peter D. Hoff and Michael D. Ward. 2019. "Inferential Approaches for Network Analysis: AMEN for Latent Factor Models." *Political Analysis* 27(2):208–222.

Newman, David, Jey Han Lau, Karl Grieser and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics pp. 100–108.

Nielsen, Richard A. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad*. Cambridge University Press.

Nowlin, Matthew C. 2016. "Modeling Issue Definitions Using Quantitative Text Analysis." *Policy Studies Journal* 44(3):309–331.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Cambridge University Press.

Poole, Keith T. and Howard Rosenthal. 2000. *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press on Demand.

Pratt, Tyler. 2018. "Deference and Hierarchy in International Regime Complexes." *International Organization* 72(3):561–590.

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.

Renshon, Jonathan and Arthur Spirling. 2015. "Modeling "Effectiveness" in International Relations." *Journal of Conflict Resolution* 59(2):207–238.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2016. "Navigating the Local Modes of Big Data." *Computational Social Science* 51.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2019. "stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91(2):1–40.

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.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley and Edoardo M. Airoldi. 2013. The Structural Topic Model and Applied Social Science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. Harrahs and Harveys, Lake Tahoe pp. 1–20.

Roberts, Margaret E., Brandon M. Stewart and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Romney, David, B. Stewart and Dustin Tingley. 2015. "Plain Text? Transparency in Computer-Assisted Text Analysis." *Qualitative & Multi-Method Research* .

Ryoo, Joseph and Neil Bendle. 2017. "Understanding the Social Media Strategies of U.S. Primary Candidates." *Journal of Political Marketing* 16(3-4):244–266.

Sievert, Carson and Kenneth Shirley. 2014. LDAvis: A Method for Visualizing and Interpreting Topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. pp. 63–70.

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.

Spirling, Arthur. 2012. "US Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911." *American Journal of Political Science* 56(1):84–97.

Stewart, Brandon M. and Yuri M. Zhukov. 2009. "Use of Force and Civil–Military Relations in Russia: An Automated Content Analysis." *Small Wars & Insurgencies* 20(2):319–343.

Terman, Rochelle. 2017. "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage." *International Studies Quarterly* 61(3):489–502.

Velden, Mariken Van Der, Gijs Schumacher and Barbara Vis. 2018. "Living in the Past or Living in the Future? Analyzing Parties' Platform Change In Between Elections,The Netherlands 1997–2014." *Political Communication* 35(3):393–412.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov and David Mimno. 2009. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM pp. 1105–1112.

Supplementary Information

Inferring Concepts from Topics:
Towards Procedures for Validating Topics as Measures

Luwei Ying
Washington University in St. Louis

Jacob M. Montgomery
Washington University in St. Louis

Brandon Stewart
Princeton University

November 6, 2020

The supplementary information to "Inferring Concepts from Topics: Towards Procedures for Validating Topics as Measures" includes four sections. The first section presents and discusses several "less ideal" validation task designs that we have tried but decided not to recommend for future researchers. The second section introduces our R package, `validateIt`, and provides a manual walking people through the procedures of posting their own tasks. The third section provides an example of the training module we used and discusses workers' performance. The last section contains additional information on the corpus and model fit.

Contents

1 Alternative Task Designs and Results

In this section, we present two variants of the Random 4 Word Set Intrusion (R4WSI) topic selection task and two other task structures for label validation. We experimented with them but decided not to recommend them as the primary choices for the reasons discussed below.

1.1 Two Variants for R4WSI

A different version of R4WSI, R4WSI-Random, chooses a random document from the corpus to display. Workers are then asked to identify the word set that does not belong. They must then choose from four different word sets, which are drawn following the same procedure as we did for R4WSI in the main text: the three non-intruder word sets are all chosen from the same topic (the topic most associated with the document) and the intruder set from a different topic. Structurally, this variant is similar to the T8WSI task.

The second variant is identical to the above R4WSI-Random task with one exception: the documents shown are always one of the ten most representative documents for a topic. Thus the only difference between these two variants is how the document is selected. For convenience, we call it R4WSI-Representative. Both variants are summarized in Table SI1.

Table SI1: Additional Task Structures for Coherence Evaluations

	A. Text Presented	B. Question Asked	C. Options Given
R4WSI-Random	A randomly selected document	After reading the above passage, please click on the word set below that is most UNRELATED to the passage.	Three word sets (each containing four mass-based selected words) from the top twenty high-probability words of one topic and one word set (the intruder) mass-based selected from the top twenty high-probability words of another topic
R4WSI-Representative	A document randomly selected from the top ten most representative documents for a given topic	After reading the above passage, please click on the word set below that is most UNRELATED to the passage.	Three word sets (each containing four randomly selected words) from the top twenty high-probability words of one topic and one word set (the intruder) randomly selected from the top twenty high-probability words of another topic

Same as other tasks, we post 500 HITs twice, summing to 1000 HITs, for each task-model combination. Note that the R4WSI-Representative task was implemented with four different structural topic models from Models 1-5 in the main text:[1]

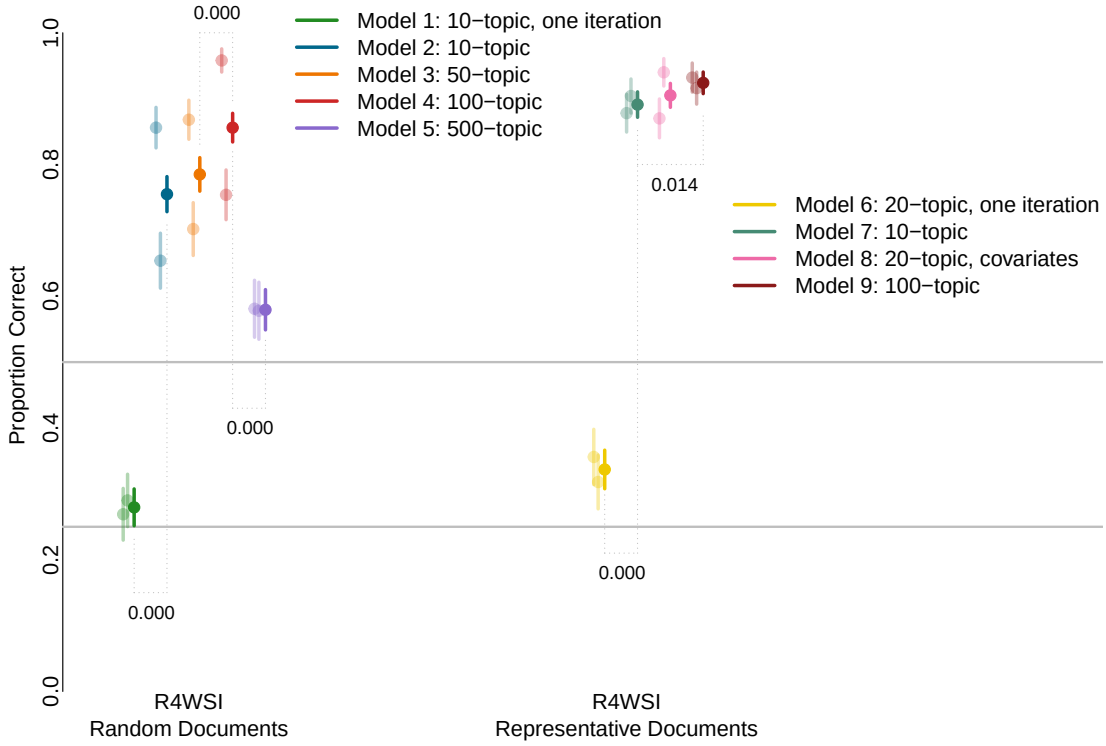
1. a 20 topic model (same as Model 2) limited to one EM iteration (which keeps the model from properly converging) with random initialization;
2. a model with 10 topics and no covariates;
3. a model with 20 topics and adding the senators' names as a covariate, and;

[1]Models 6-9 were from an earlier version of the paper where we fit topic models without holding out any documents.

4. a model with 100 topics.

The result can be seen in Figure SI1. Results from R4WSI-Random are not consistent and the two identical trials for Model 2-4 are significantly different. In R4WSI-Representative, the two identical trials for Model 8 are significantly different. Additionally, Models 2 and 3 are not distinguishable from each other, nor does Model 3 and 4. We think this is because the R4WSI-Representative task only uses the clearest documents and thus results in a ceiling effect. For that reason, we do not recommend future researchers to adopt either of the variants presented here.

Figure SI1: Results for Coherence Evaluations



Note: The 95% confidence intervals are presented. The two transparent bars represent two identical trials (500 HITs each). The solid bar represents the pooled result (1000 HITs). When two models yield significantly different results, the p-value is noted. (Significance tests are difference in proportions as calculated by the `prop.test` function in R.)

1.2 Two Alternative Label Validation Tasks

We have also designed two alternative label validation tasks that rely only on the labels and high probability words. The *Word Set Intrusion* task copies the R4WSI described in the topic validation section in the main text, simply adding the label at the top of the task. An example is shown in Panel (a) of Figure SI2. The *Optimal Label for Word Sets* (OLW) task presents a topic, represented by the eight highest-probability words, and four labels. One label is for the relevant topic and the others are for other topics in the model but within the specified broad category. An example is shown in Panel (b) of Figure SI2. Both task structures are summarized in words in Table SI2. These various task structures (two presented in the main text and two here) vary with respect to whether topic content is primarily inferred from word sets or documents, whether the goal is to find intruders or optimal relationships between topics and labels.

Table SI2: Additional Task Structures for Label Validation

	A. Text Presented	B. Question Asked	C. Options Given
WSI	A label[a]	Please click on the word set below that is most UNRELATED to the above label.	Three labels for the top three high-probability topics and one label for other economy-related topics
OLW	The eight highest-probability words for a topic	Please read the four labels below and click on the label that BEST summarizes the word set.	One label for the highest-probability topic and three labels for other economy-related topics

[a]The label for one of ten economy-related topics in this empirical illustration.

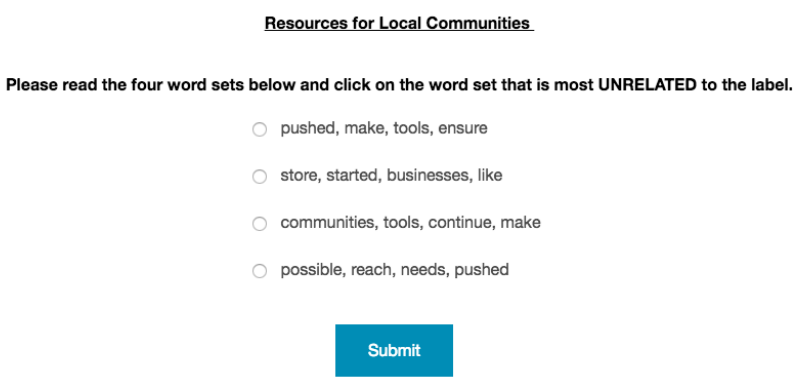
Past experience teaches that word sets are easier to associate with a topic than an actual document since language out of context is easier to re-interpret. This can make associating broad concepts with word sets easier than with documents, but often at the expense of verisimilitude. Thus, it should not be too surprising to find that validations based primarily on word sets will give less clear guidance.

To test these task structures we followed the same basic procedures discussed above. For each task/coder combination we created 500 tasks that were coded by trained workers on AMT.[2] The results from 4,000 high quality HITs are shown in Figure SI3.

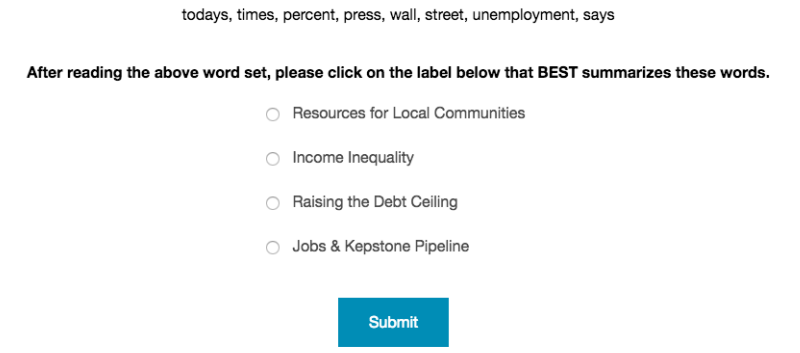
The expected result that the "Careful Coder" labels were superior, however, was not confirmed by the tasks based around word sets (WSI and OWS). Ex ante this makes sense since the "Word Coder" labels were based only on the word sets to begin with. Given that the "Careful Coder" labels were intentionally designed to more accurately represent the underlying meaning of the topics as they appear in the actual documents, we infer that label validation should be based on documents rather than word sets alone. In all, based on these results we would recommend that labels be evaluated using the LI and OL.

[2]Workers were paid 0.02/HIT for the OLW task, 0.03/HIT for the WSI task

Figure SI2: Example HITs for Alternative Label Validation Tasks

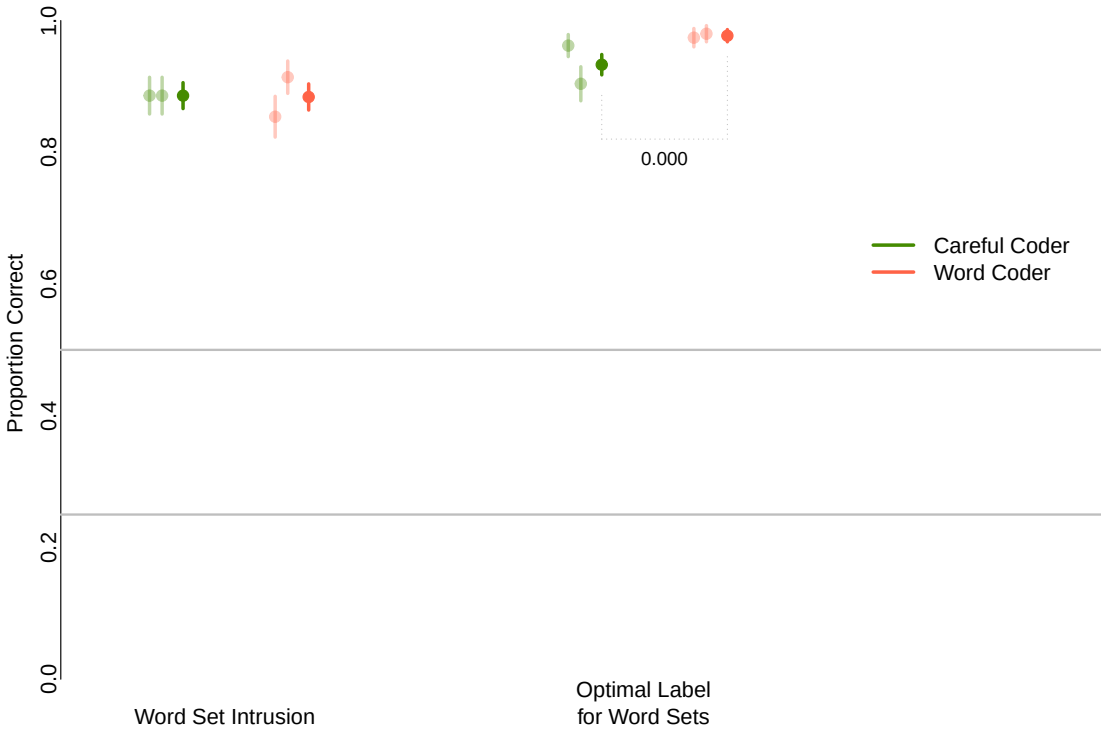


(a) Word Set Intrusion



(b) Optimal Label for Word Sets

Figure SI3: Results for Label Validation



Note: The 95% confidence intervals are presented. The two transparent bars represent two identical trials (500 HITs each). The solid bar represents the pooled result (1000 HITs). When two models yield significantly different results, the p-value is noted. (Significance tests are difference in proportions as calculated by the `prop.test` function in R.)

2 Software and Working Example

Along with the paper, we provide free-to-use software to facilitate the topic and label validation procedures for future researchers. This section serves as a user manual.

2.1 R Package: `validateIt`

The R package, `validateIt`, to implement the methods we proposed in the paper is currently being prepared for submission to the *Comprehensive R Archive Network* (CRAN). Users can install the most recent development version from `Github` using the `devtools` package. First, users would have to install `devtools` using the following code. Note that this step only has to be done once:

```
if(!require(devtools)) install.packages("devtools")
```

Then you can load the package and use the function `install_github`

```
library(devtools)
install_github("Luwei-Ying/validateIt", dependencies = TRUE)
```

Note that this will install all the packages suggested and required to run our package. It may take a few minutes the first time, but this only needs to be done on the first use. In the future users can update to the most recent development version using the same code.

2.2 Fitting Candidate Topic Models and Labeling

From here, we provide example code to post validation tasks. It shows the entire procedure of validating topic model outcomes using our proposed methods. This procedure consists five major steps: 1) Fitting Candidate Topic Models and Labeling; 2) Training and Certification; 3) Preparing Tasks Locally; 4) Posting Tasks to Mturk and Maintaining Them Online; and 5) Retrieving and Evaluating Results.

Researchers need to fit several candidate topic models to be validated later. To illustrate the methods, we fit 5 topic models using the `R` packages `stm` with 500 topics, 100 topics, 50 topics, 10 topics, 10 topics but restricted to only 1 EM iteration, as described in the main manuscript. Topic models generated by other software will also work as long as the users keep track of a) the documents in the topic model; b) the vocabulary in the topic model; c) the matrix of topic conditional word probabilities, usually known as the beta matrix; and d) the matrix of the topic distribution for each document, usually known as the theta matrix.

While this section does not intend to provide detailed instructions on fitting topic models, we want to point out that we did not stem the words. We also randomly select 10% of the documents (16364 out of 163642 documents) and **held out** 50% of the tokens in these documents so we will be able to compare the results from our methods with held-out likelihood. Code for this procedure is included below.

```
# step 1: pre-process, obtain the stm object
# define the customized stop words outside of the function
customstopwords <- c('senator', 'senate', 'sen', 'facebook', 'please', 'today',
                     'will', tolower(state.name), tolower(state.abb), 'rhode',
                     'hampshire', 'jersey', 'york', 'carolina', 'dakota')
docs = textProcessor(documents = as.character(Senate$post_text),
                     metadata = Senate,
                     lowercase = TRUE,
                     removestopwords = TRUE,
                     removenumbers = TRUE,
                     removepunctuation = TRUE,
                     stem = FALSE,
                     wordLengths = c(3, 20),
                     language = "en",
                     striphtml = TRUE,
                     customstopwords = customstopwords)
stmPrep <- prepDocuments(docs[[1]], docs[[2]], meta = docs$meta)
save(stmPrep, file = 'Corpus/stmPrep.Rdata')

# step 2: heldout some words from some documents
heldout <- make.heldout(stmPrep$documents, stmPrep$vocab, seed = 123)
save(heldout, file = 'Corpus/heldout.Rdata')
documents <- heldout$documents
vocab <- heldout$vocab

# step 3: fit stm with different topic numbers, i.e., different k
stm <- stm(documents, vocab, k = 10)
save(stm, file = 'Models/unstemmed/stm.Rdata')
```

To label the topics, we strongly recommend users to follow the practice of the "Careful Coder": carefully read the high-probability words and frequent & exclusive words (FREX), as well as fifty representative documents per topic. Ideally, future researchers will have more than one coder to independently label the topics so they can assess inter-coder reliability.

2.3 Training and Certification

First of all, researchers need to register a Mturk requester account here: `https://www.mturk.com/`.

For each of the task structures, Word Intrusion (WI), Top 8 Word Set Intrusion (T8WSI), Random 4 Word Set Intrusion (R4WSI), Label Intrudion (LI), and Optimal Label (OL), researchers need to make a training module to introduce the background and get online workers familiar with the tasks. While doing so, we provide five practice questions, where people's answers are NOT scored, then eight test questions. The workers need to get 7 or more out of the 8 questions correct to receive the qualification.

The training modules are written in .xml files, each with a separate .xml file coding the answers. Section 3.1 - 3.5 present the training modules used for this paper. For .xml formatting, see the example files "Question.xml" and "Answer.xml" in our replication package.

We then create the qualifications on Mturk:

1) Specify the Mturk options.

```
library(pyMTurkR)
Sys.setenv(AWS_ACCESS_KEY_ID = AWS_ACCESS_KEY_ID)
Sys.setenv(AWS_SECRET_ACCESS_KEY = AWS_SECRET_ACCESS_KEY)

# change sandbox = F when ready to run on MTurk
options(pyMTurkR.sandbox = T)

# use this to test that the pyMturkR settings are correct
AccountBalance()
```

2) Read in the .xml file for questions and answers.

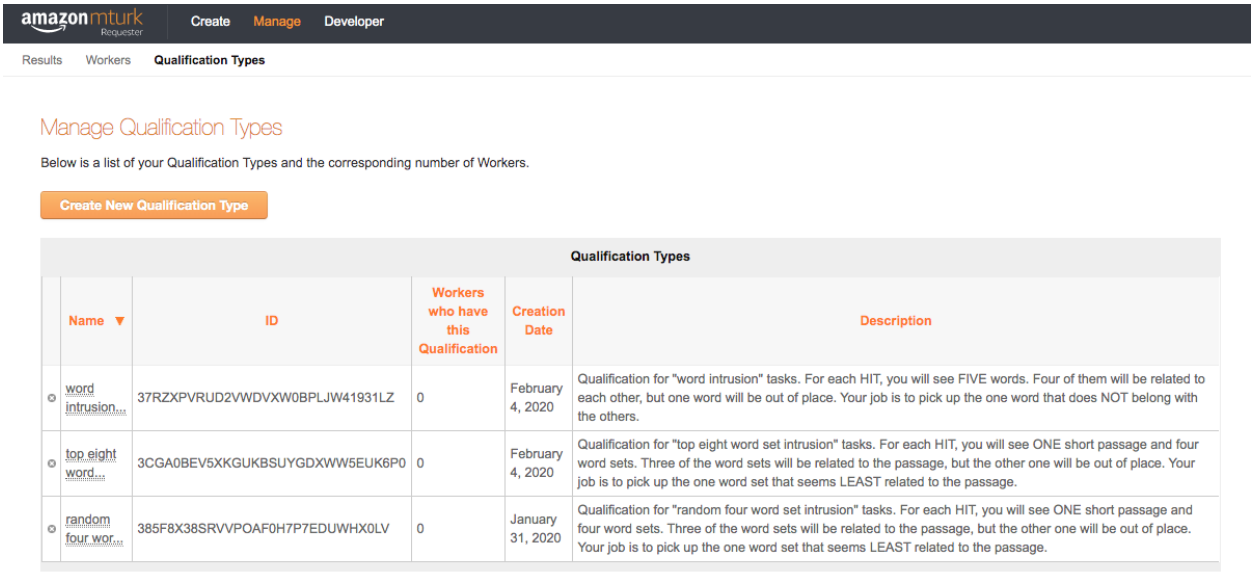
```
TestQuestions <- paste0(readLines("T8WSIQuestion.xml", warn = FALSE),
                        collapse = "")
TestKey <- paste0(readLines("T8WSIAnswer.xml", warn = FALSE),
                  collapse = "")
```

3) Create the qualification.

```
T8WSIQual <- CreateQualificationType(
             name = "top eight word set intrusion qualification",
             description = 'Qualification for "top eight word
                            set intrusion" tasks.',
             status = "Active", # allows qual to remain active for users
             test = TestQuestions, # pass questions for test
             test.duration = 60 * 60, # test duration, in seconds
             retry.delay = NULL, # how long until worker can retry test;
                                   NULL means never
             answerkey = TestKey)
```

The created qualification can be seen on the requester's dashboard under the "Manage" tab after a few minutes.

Figure SI4: Check the qualification, which should appear on the dashboard



4) Save the qualification in case changes need to be made in the future.

```
save.image('Qualifications.RData') # you may need to set file path
```

5) Update a qualification if needed, e.g., fix a typo.

```
load('Qualifications.RData')
T8WSIQual <- UpdateQualificationType(
             qual = T8WSIQual$QualificationTypeId, # keep same qualification id
             description = T8WSIQual$Description,
             status = 'Active', # manually set status
             test = TestQuestions, # update test
             test.duration = 60 * 60, # manually set test duration
             retry.delay = NULL, # manually set retry delay
             answerkey = TestKey) # update answer key
```

2.4 Preparing Tasks

Researcher will prepare tasks locally before sending them to Mturk.

2.4.1 Preparing Tasks for Topic Validation

For topic validation, our package, `validateIt`, requires a) the documents in the topic model (`docs`); b) the vocabulary in the topic model (`vocab`); c) the beta matrix in the topic model (`beta`); and d) the theta matrix in the topic model (`theta`).

In the main manucript, we fit *unstemmed* models. The `combMass()` function takes in a `stm` output and combine mass for words with the same root. The output, `newMass`, is a list

of two: the combined vocabulary matrix where words are completed to the most frequent form in that specific topic (`newvocab`) and the combined beta matrix in correspondance with the combined vocabulary (`newbeta`). This step is not required in using our crowd-sourcing methods. We see this procedure as analogous to fitting a stemmed topic model but retain the complete form of words.

```
newMass <- combMass(stm)
```

Now, researchers start formatting the tasks:

1) The `validateTopic()` function creates tasks of the desired type and number. Below is the example code for T8WSI, the structure of which task is the most complicated. We have left out the documents where a portion of words have been held out. The default threshold is 20, meaning that our algorithm draws from the pool of top 20 highest probability words.

```
T8WSItasks <- validateTopic(type = "T8WSI",
                            n = 500,
                            text = stmPrep$meta$post_text[-heldout$missing$index],
                            vocab = newMass[[1]],
                            beta = newMass[[2]],
                            theta = stm50k$theta[-heldout$missing$index,],
                            thres = 20)
```

For WI and R4WSI tasks, leave out the `docs` and `theta` arguments. The default threshold is still 20.

```
# WI
WItasks <- validateTopic(type = "WI",
                         n = 500,
                         vocab = newMass[[1]],
                         beta = newMass[[2]],
                         thres = 20)

# R4WSI
R4WSItasks <- validateTopic(type = "R4WSI",
                            n = 500,
                            vocab = newMass[[1]],
                            beta = newMass[[2]],
                            thres = 20)
```

All word-drawing processes are probability-based (mass-based). Specifically, in preparing WI tasks, the function a) orders the word probability (beta[k,]) for each topic k, then b) randomly draws 4 words from top 1 to `thres` words **based on their corresponding probabilities in** k, and then c) randomly draws 1 intruder word from top 1 to `thres` words from another topic, $\neg k$ **based on it corresponding probability in** $\neg k$.

In preparing R4WSI, the function a) for each topic k, randomly draws 12 words from top 1 to `thred` words **based on their corresponding probabilities in** k and randomly assign them to 3 different word sets, then b) randomly draws 4 intruder words from top 1

to `thred` words from another topic, $\neg k$ **based on their corresponding probabilities in** $\neg k$.

In preparing T8WSI, the function a) uses `vocab` and `beta` to find the top 8 words for each topic k, then b) randomly sample a document from the document pool (`docs`) and calculate the top 3 high probability topics associated with that document using `theta`, and then c) sample an intruder topic from other lower probability topics. Notice that the word drawing procedure here is not probability-based as the function is always looking for the top 8 words regardless of their mass.

The output while specifying "T8WSI" looks as below, where each row represents a task. Each task contains an indicator of the topic (`topic`), a randomly drawn document associated with that topic (`doc`), three non-intruder word sets (`opt2` - `opt3`), and an intruder word set (`optcrt`).

Figure SI5: Output of `validateTopic()`

	topic	doc	opt1	opt2	opt3	optcrt
1	1, 5, 8	Courage and a love of liberty hav...	day, family, wo...	president, obam...	people, vote, rig...	health, care, veterans,
2	10, 3, 1	Watch Senator Leahy announce t...	work, bill, act, n...	discuss, watch, ...	day, family, wo...	people, vote, right, pre
3	5, 8, 1	I congratulate the Argentine peo...	president, obam...	people, vote, rig...	day, family, wo...	discuss, watch, news,
4	3, 5, 10	Our latest newsletter went out to...	discuss, watch, ...	president, obam...	work, bill, act, n...	people, vote, right, pre
5	1, 3, 2	When I studied for a year at the U...	day, family, wo...	discuss, watch, ...	great, visit, mee...	work, bill, act, need, h
6	1, 2, 10	Events across North Carolina hav...	day, family, wo...	great, visit, mee...	work, bill, act, n...	jobs, businesses, ener
7	1, 8, 10	Like many others, I still remembe...	day, family, wo...	people, vote, rig...	work, bill, act, n...	discuss, watch, news,
8	10, 4, 8	The St. Louis Post-Dispatch & ST...	work, bill, act, n...	national, service...	people, vote, rig...	jobs, businesses, ener
9	2, 1, 4	I would like to announce the serv...	great, visit, mee...	day, family, wo...	national, service...	people, vote, right, pre
10	9, 6, 1	Social Security and Veterans bene...	health, care, vet...	tax, budget, gov...	day, family, wo...	president, obama, adn

2) In addition to generating tasks for validation, we suggest preparing a series of gold-standard HITs to monitor the quality of the works in the future. These gold-standard HITs are the "easy" tasks whose answers are unambiguous.

While preparing gold-standard HITs for the current paper, we select 50 HITs for each task structure from pilot runs on Mturk where two different workers have agreed on the answers in two identical rounds. These selection processes were **not random** and we also modified the documents and words to make the answers even clearer. Future researchers could simply generate more tasks than they need and hand pick the gold-standard HITs they deem as clear from the excessive tasks. An example of T8WSI gold-standard HITs reads like this:

> *The Keystone XL pipeline represents not only thousands of jobs and growth for the nation's economy, but also a big step toward American energy independence. We can become energy independent in America within five to seven years, but we must commit to moving forward with important projects like the Keystone XL pipeline.*

After reading the above passage, please click on the set of words below that is most unrelated to the passage.

- o jobs, business, energy, new, economy, create, state, economic
- o work, project, forward, need, american, legislation, support, make
- o oil, energy, security, pipeline, administration, states, strategy, must
- o day, family, holiday, summer, beach, play, sunshine, vacation

3) Randomly mix in the gold-standard HITs. The `mixGold()` function ensures that one gold-standard HIT would show up every $\frac{\#\text{ total tasks}}{\#\text{ gold standard tasks}}$ number of tasks. It also assigns a unique id for each of the task.

```
goldT8WSI <- read.csv("goldT8WSI.csv", stringsAsFactors = FALSE)
allT8WSItasks <- mixGold(tasks = T8WSItasks, golds = goldT8WSI)
```

The output looks as below, where the topic column for gold-standard HITs indicates "gold."

Figure SI6: After mixing in gold-standard HITs

	topic	doc	opt1	opt2	opt3	optcrt	id
1	1, 5, 8	Courage and a love o...	day, family, wo...	president, obam...	people, vote, rig...	health, care, vet...	1
2	10, 3, 1	Watch Senator Leahy ...	work, bill, act, n...	discuss, watch, ...	day, family, wo...	people, vote, rig...	2
3	5, 8, 1	I congratulate the Ar...	president, obam...	people, vote, rig...	day, family, wo...	discuss, watch, ...	3
4	gold	I have substantial co...	health-care, car...	budget, govern...	people, right, pr...	overseas, iraq, s...	4
5	3, 5, 10	Our latest newsletter ...	discuss, watch, ...	president, obam...	work, bill, act, n...	people, vote, rig...	5

4) Next, record the prepared tasks to a specified path. While doing so, the function will create a list of two, where the first element is the above data frame of prepared tasks and the second element only keeps the documents and word sets (in the case of WI only keeps the words and in the case of R4WSI only keeps the word sets) with randomized order, leaving out other meta data, e.g., the ids. Users can assign the record to an object for immediate use. They can always load it later as well.

```
record <- recordTasks(type = "T8WSI", tasks = allT8WSItasks,
                      path = "T8WSI/record.Rdata")
```

2.4.2 Preparing Tasks for Label Validation

The procedure for preparing label validation tasks locally is almost identical to that for preparing topic validation tasks. One extra step is that users need to predict the high probability topics for their documents and define a document pool, from which they would like to draw documents. Typically, the document pool for Label Intrusion (LI) tasks contains all documents whose top three high probability topics are among the pre-defined category (in our example, domestic topics). The document pool for Optimal Label (OL) tasks, on the other hand, contains all documents whose top one high probability topic is among the pre-defined category (in our example, domestic topics). The document pool file – like the one depicted in Figure SI7 – must contain a document column and one or three top column(s) named as "top1" (and "top2"/"top3").

Figure SI7: Structure of the Document Pool Document

	post_text	top1	top2	top3
1	Tomorrow once again we will vote to protect women's...	9	86	45
2	Yesterday I grilled U.S. financial watchdogs on the col...	65	67	83
3	The payroll tax cut expires in two weeks. Letting it ex...	35	88	89
4	Today we passed the payroll tax cut extension so that...	88	9	22
5	Here's an excellent editorial from the Pioneer Press ab...	62	83	56
6	Ive been fighting to help MN farmers and ranchers rec...	65	94	88
7	For years Ive been pushing the big phone companies ...	65	16	26
8	Today our bipartisan working group in the Senate ann...	44	89	33
9	Today I was in St. Paul to highlight public-private part...	88	46	98
10	Today I held a workforce development forum at Dunw...	62	83	56

Now the users will use the function `validateLabel()`, specifying either `type = "LI"` or `type = "OL"`. `text.predict` specifies the document pool. `text.name` specifies the name of the document column of the document pool data frame. `labels` are the labels given by human coders, in the same order with `labels.index`. Users could choose to add additional intruder labels through the `labels.add` argument. In our example for illustration, we validate 10 domestic labels while adding 10 international labels as cross-category intruders.

```
documentpool <- read.csv("documentpool.csv", stringsAsFactors = FALSE)
OLtasks <- validateLabel(type = "OL",
                         n = 500,
                         text.predict = documentpool,
                         text.name = "post_text",
                         labels = c("Equal Pay for Women",
                                    "Healthcare/Reproductive Rights",
                                    "Agriculture",
                                    "Student Loan/Debt",
                                    "Drug Abuse",
                                    "Higher Education/Job Training",
                                    "Wall Street/Financial Sector",
                                    "Government Sutdown/Congressional Budget",
                                    "Obamacare/Tax Policy",
                                    "Deficits/Debt/Budget"),
                         labels.index = c(1, 9, 21, 35, 44, 62, 65, 70, 88, 89),
                         labels.add = c("International Trade",
                                        "Praising Active Military/Military Units",
                                        "Terrorism",
```

```
"Military Sexual Assault",
"Nuclear Deterrence/International Security",
"Air Force",
"Honoring Specific Veterans",
"Honoring Veterans/Heroes",
"Military Operations/Armed Conflicts",
"Veterans Affairs/Veterans Healthcare"))
```

Like the topic intrusion tasks, we have randomly mixed 50 gold-standard HITs into 500 tasks for both task structure. We recommend future researchers to adopt this "gold-standard HITs" approach as well. The `recordTasks()` function records the tasks at a specified local directory. The two elements in the output list of `recordTasks()` is shown in Figure SI8 and SI9.

Figure SI8: Local Record with Identifiers (`recordTasks()` Output 1)

	topic	doc	opt1	opt2	opt3	optcrt	id
1	1	Americans deserve a #fairsho...	Deficits/De...	Government...	Agriculture	Equal Pay fo...	1
2	9	Did you know the Republican...	Student Loa...	Higher Educ...	Government...	Healthcare/...	2
3	21	ICYMI read my op-ed in the l...	Deficits/De...	Healthcare/...	Equal Pay fo...	Agriculture	3
4	35	Congress has a responsibility...	Equal Pay fo...	Drug Abuse	Agriculture	Student Loa...	4
5	44	The time is NOW to address t...	Government...	Equal Pay fo...	Student Loa...	Drug Abuse	5
6	62	The Crofton girl wrote to Pre...	Equal Pay fo...	Obamacare/...	Healthcare/...	Higher Educ...	6
7	65	Read my op-ed in the Wall St...	Student Loa...	Drug Abuse	Deficits/De...	Wall Street/...	7
8	70	New House Ukraine bill show...	Agriculture	Student Loa...	Drug Abuse	Government...	8
9	88	I remain committed to repeali...	Student Loa...	Equal Pay fo...	Deficits/De...	Obamacare/...	9
10	gold	Yesterday, the Senate voted t...	Happy Moth...	Linking to S...	Condolence...	Trade	10
11	89	Admiral Mullen, Chairman of ...	Equal Pay fo...	Drug Abuse	Wall Street/...	Deficits/De...	11

Figure SI9: Local Record, Randomized Order, i.e., (`recordTasks()` Output 2)

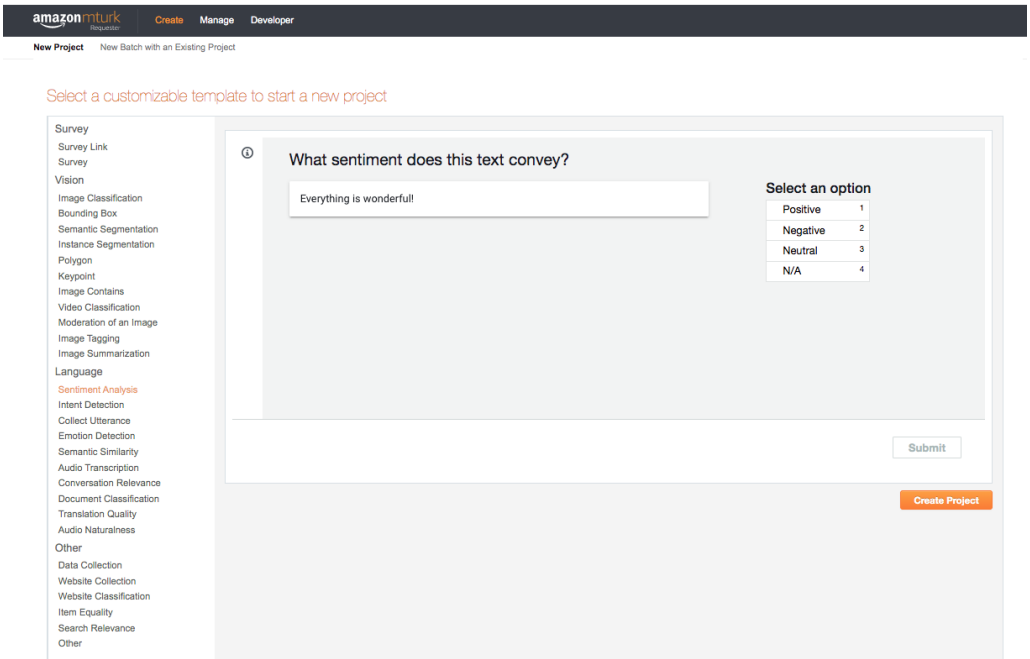
	passage	word1	word2	word3	word4
1	Americans deserve a #fairshot ...	Deficits/Debt/B...	Equal Pay for Wo...	Agriculture	Government Sut...
2	Did you know the Republican bi...	Healthcare/Repr...	Government Sut...	Higher Educatio...	Student Loan/De...
3	ICYMI read my op-ed in the In...	Agriculture	Deficits/Debt/Bu...	Healthcare/Repr...	Equal Pay for Wo...
4	Congress has a responsibility t...	Agriculture	Student Loan/Debt	Drug Abuse	Equal Pay for Wo...
5	The time is NOW to address thi...	Student Loan/De...	Government Sut...	Equal Pay for Wo...	Drug Abuse
6	The Crofton girl wrote to Presid...	Higher Educatio...	Healthcare/Repr...	Obamacare/Tax...	Equal Pay for Wo...
7	Read my op-ed in the Wall Stre...	Deficits/Debt/B...	Wall Street/Finan...	Student Loan/D...	Drug Abuse
8	New House Ukraine bill shows ...	Student Loan/De...	Agriculture	Government Sut...	Drug Abuse
9	I remain committed to repealin...	Equal Pay for Wo...	Student Loan/Debt	Deficits/Debt/B...	Obamacare/Tax ...
10	Yesterday, the Senate voted to ...	Linking to Speec...	Trade	Condolences me...	Happy Mother's ...
11	Admiral Mullen, Chairman of th...	Equal Pay for Wo...	Drug Abuse	Wall Street/Fina...	Deficits/Debt/B...

2.5 Posting Tasks and Maintaining

Now, researchers should be able to interact with Mturk and post tasks. They will do so by first manually specifying the basic tasks settings and then sending tasks through the API.

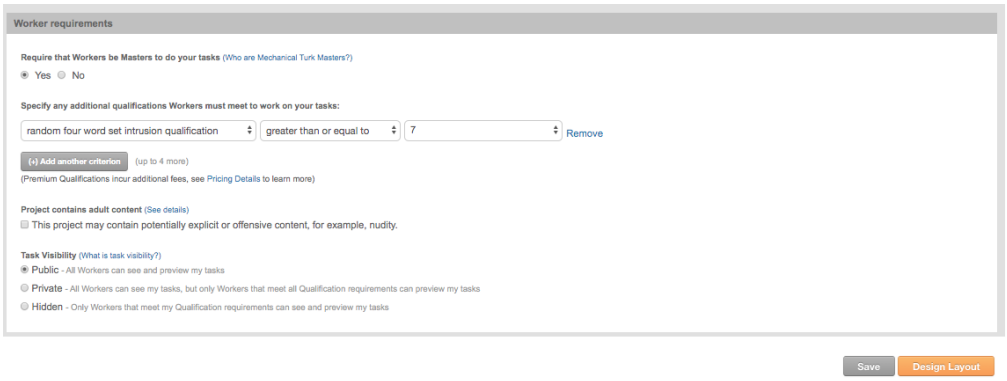
1) Login to the Mturk requester page, click "New Project" under the "Create" tab. Click the "Sentiment Analysis" tab (which allows us to specify our customized layout later) on the left. Then, click "Create Project."

Figure SI10: Create a Project: Step One



2) Following the prompts, specify the properties. In particular, be sure to add an additional qualification created in section "Training and Certification" under the "Worker requirements" tab. Our tasks in the paper require a "master" qualification as well. Then, click "Design Layout."

Figure SI11: Create a Project: Step Two



3) Replace the default layout file with our customized layout. The example .html file can be found as a separate file in our replication package (Layout.html). Notice that researchers can change the instructions for their own tasks, which will later appear on the left of the workers' screen. For the tasks in our paper, we specifically tell them:

> Some of these choices will be very clear, but others will require you to use your best judgment. We understand that in many cases it will be hard to tell what the "right" choice is. Use your best judgement, but please be attentive. **We have randomly mixed in a number of gold-standard HITs whose answers are less ambiguous to evaluate the quality of your work.** If you miss a large number of gold-standard HITs, you may be **blocked** from continued participation in this study (and future studies).
>
> After you've done a relatively large number of HITs of a specific task structure, your qualification might be temporarily suppressed. This does not mean you've done anything wrong, but we need to ensure that a variety of workers complete our tasks. This is **NOT a block** and will have **NO negative impact** on you Mturk record.

4) Preview the Layout and finish creating tasks.

From here, we can do everything through the API, sending instructions from `R` to interact with the Mturk platform.

5) Load the prepared tasks from section 2.4.

```
load("T8WSI/record.Rdata")
```

6) Send tasks to Mturk using the `sendTasks()` function. Users must specify a path to **record the HITids immediately** as they are vital for future procedures. The function

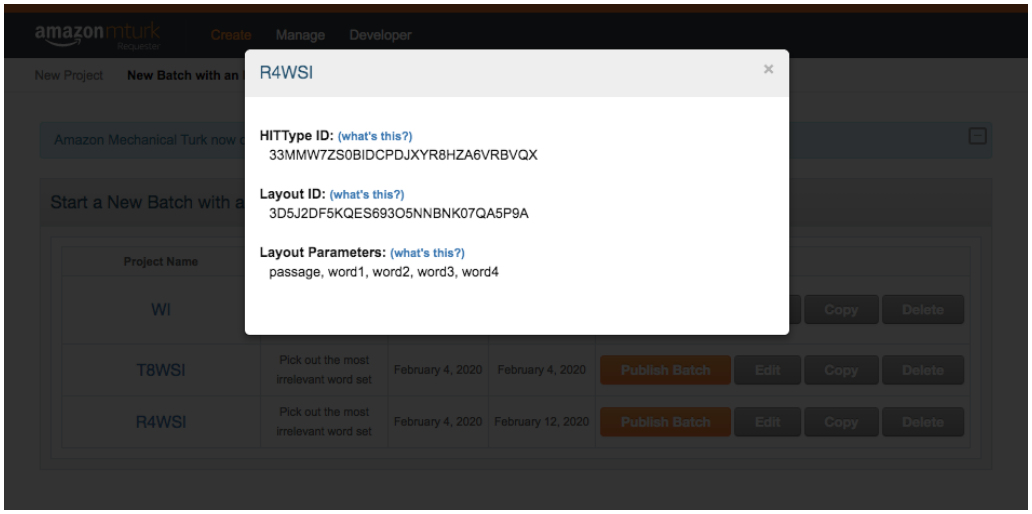
will record the HITids in a list format where the first element is a vector of HIT ids returned by Mturk and the second element is **the mapping of the local HIT ids to the Mturk HIT ids**. The `tasksids` argument allows users to post specific tasks by specifying the task ids (numeric form) in a vector. If `tasksids` left unspecified, all tasks in the record will be posted.

```
HITids <- sendTasks(hit_type = 'FIND_IT_FROM_DASHBOARD',
                    hit_layout = 'FIND_IT_FROM_DASHBOARD',
                    type = "T8WSI",
                    tasksrecord = record,
                    tasksids = c(1:10, 15),
                    HITidspath = "T8WSI/HITids/testIDs.Rdata")

# Console responds
> Sending task to MTurk
> HITids saved to T8WSI/HITids/testIDs.Rdata
```

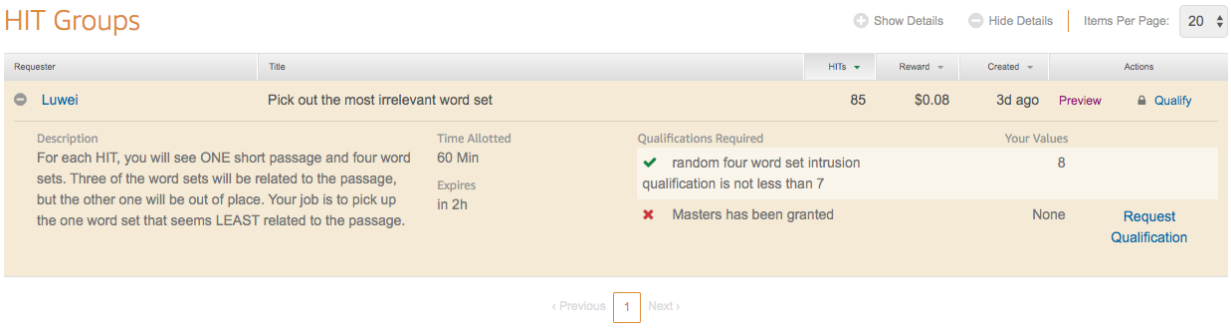
Notice that `hit_type` and `hit_layout` can be found from the MTurk requester's dashboard by clicking the project name, e.g., "T8WSI":

Figure SI12: Finding HITType ID and Layout ID



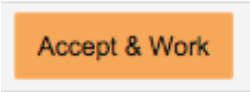
Once posted, worker should be able to see the tasks on the Mturk platform. They can get the certification by going through the training module and passing the test. Figure SI13 depicts an example task from the workers' perspective, we see that a "master" qualification is also required.

Figure SI13: How tasks look on Mturk



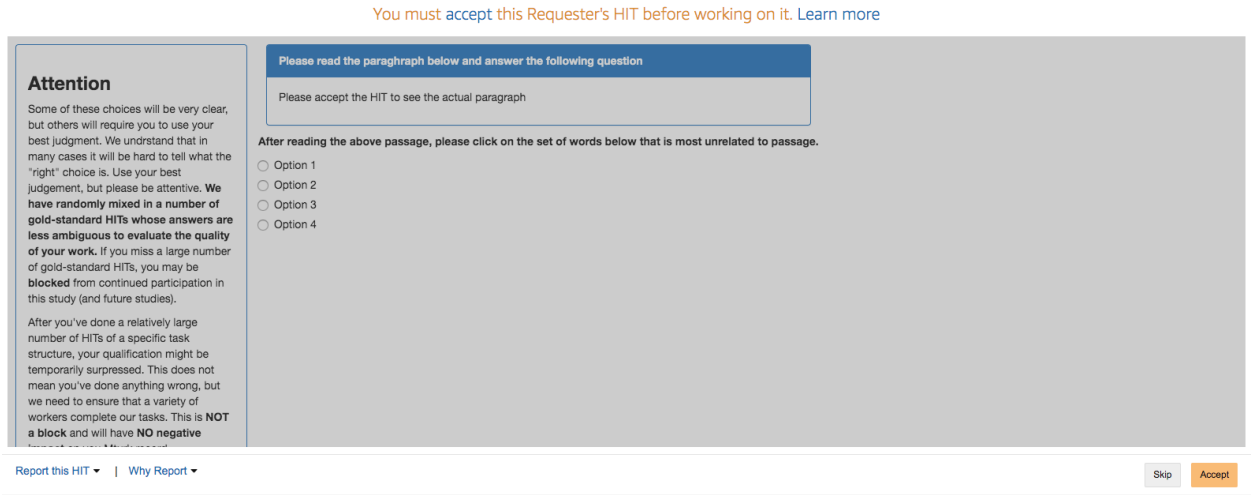
Once the workers get the qualification, the "lock" icon will become "Accept & work".

Figure SI14: Workers can start working after seeing this "Accept & work" icon



Before they accept the HIT, they would **NOT** be able to see the actual options or document, which feature prevents them from selecting the tasks.

Figure SI15: Before Accepting the HIT

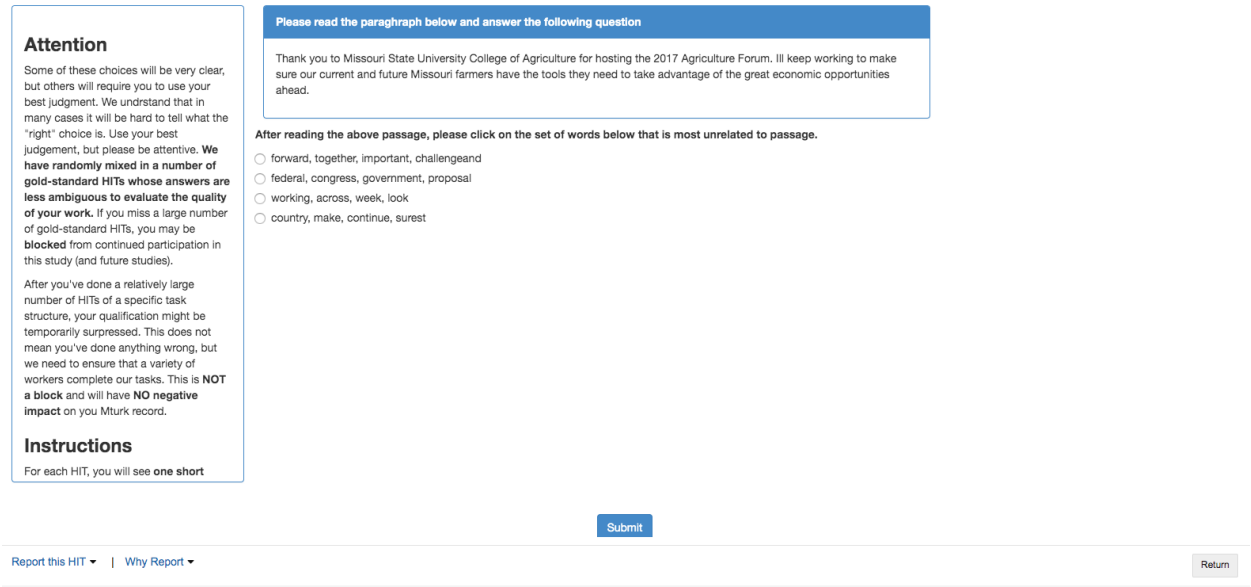


They'll only be able to see the actual tasks after accepting. Notice that they can ALWAYS see the instructions on the left.

Researchers are able to extend the time of their HITs before or after the HITs expire.

```
for(i in HITids[[1]]){
  ExtendHIT(hit = i, add.seconds = 3600)
}
```

Figure SI16: After Accepting the HIT



2.6 Retrieving and Evaluating Results

The `getResults()` function allows the researchers to retrieve results from a batch. It would work **no matter the batch has been completed or not**. In that sense, this function can also be used to check batch status. Notice that `batch_id` here can be any string that helps the researchers to refer to the batch in the future. It is first specified here. The `retry`, if `TRUE`, retries retrieving results from Mturk API at most five times, given that it fails retrieving all results the first time. The default is `retry = TRUE`.

```
testresults <- getResults(batch_id = "testbatch",
                          hit_ids = HITids,
                          retry = FALSE)

# Console returns
> Start getting HITs...
> 208 / 250 results retrieved
```

We highly recommend users to record the results immediately, especially when the batch is finished as the results might be deleted by the Mturk platform at some point.

To evaluate results, use `evalResults()`. This function identifies workers who consistently give poor-quality work and also returns the rate that human workers agree with the machine prediction.

```
evalResults(results = testresults,
            key = record,
            type = "T8WSI")
```

Figure SI17: Evaluating Results: Output from `evalResults()`

```
208 / 250 results will be evaluated
35 / 35 gold-standard HITs are answered correct
123 / 173 non-gold-standard HITs are answered correct
$`Gold-standard HIT Correct Rate`
[1] "1"       "35 / 35"

$`Gold-standard HIT Correct Rate by Workers`

                 TRUE
0                   0
A1[redacted]F      19
A2[redacted]J4      8
AY[redacted].W      8

$`Gold-standard HIT Correct Rate`
[1] "0.710982658959538" "123 / 173"
```

In rare cases, researchers will encounter workers who keep giving work of poor quality. Therefore, researcher can reject HITs from these workers and ban them from participating in future tasks in two steps.

1) Identify these workers and revoke their qualifications.

```
workers_to_ban <- c("WORKER1", "WORKER2")
for(worker in workers_to_ban){
  AssignQualifications(qual = "QUALIFICATION_ID",
                       workers = worker,
                       notify = TRUE,
                       value = 0)
}
```

2) Reject their qualifications.

```
AssignmentsToReject <-
    testresult$assignment_id[testresults$worker_id %in% workers_to_ban]
RejectAssignments(assignments = unique(AssignmentsToReject),
                  feedback = 'The quality of your work in the most recent
                  batch of HITs did not pass an audit. Your HITs from this
                  batch have been rejected and your qualification has
                  been deactivated.')
```

3 Workers' Training and Performance

This section first presents the training modules for the five tasks used in the paper. The second part provides some statistics and discussion about the workers' performance.

3.1 Training Module for Word Intrusion

Completing this training module qualifies you to complete Word Intrusion HITs.

Basic instructions
1. For each HIT, you will see FIVE words.
2. Four of them will be related to each other, but one word will be out of place.
3. Your job is to pick up the one word that does NOT belong with the others.

Background

The words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o job
- o business
- o day
- o work
- o economy

Answer: The correct answer is "day". "job", "business", "work" and "economy" are all related to economic activities. "day" is not.

Practice HIT #2

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o energy

- o water
- o climate
- o small
- o clean

Answer: The correct answer is "small". "energy", "water", "climate" and "clean" are all connected to one another under the theme of environmental policy concerns. However, "small" is irrelevant to this theme.

Practice HIT #3

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o country
- o america
- o oil
- o nation
- o freedom

Answer: The correct answer is "oil". "country", "america", "nation" and "freedom" are related to America and American values. However, "oil" is irrelevant to this theme.

Practice HIT #4

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o farm
- o school
- o students
- o university
- o high

Answer: The correct answer is "farm". "school", "students", "university" and "high" are all related to education. However, "farm" is not.

Practice HIT #5

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o veterans
- o service
- o proud
- o meeting
- o force

Answer: The correct answer is "meeting". "veterans", "service", "proud" and "force" are all related to veterans affairs and the armed services. However, "meeting" is irrelevant.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o tax
- o budget
- o debt
- o bipartisan
- o spending

Answer: [**Not Available in the Real Test**] "bipartisan"

Test HIT #2

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o homework
- o security
- o nuclear
- o iran
- o hearing

Answer: [**Not Available in the Real Test**] "homework"

Test HIT #3

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o watch
- o news
- o committee
- o live
- o show

Answer: [**Not Available in the Real Test**] "committee"

Test HIT #4

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o vegetable
- o congress
- o president
- o republicans
- o house

Answer: [**Not Available in the Real Test**] "vegetable"

Test HIT #5

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o health
- o bill
- o act
- o airplane
- o legislation

Answer: [**Not Available in the Real Test**] "airplane"

Test HIT #6

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o agriculture
- o happy
- o farmers
- o industry
- o trade

Answer: [**Not Available in the Real Test**] "happy"

Test HIT #7

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o sanders
- o social
- o night
- o billion
- o spending

Answer: [**Not Available in the Real Test**] "night"

Test HIT #8

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

- o nuclear-power
- o discuss
- o watch
- o spoke
- o read

Answer: [**Not Available in the Real Test**] "nuclear-power"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

3.2 Training Module for Top 8 Word Set Intrusion

Completing this training module qualifies you to complete Top Eight Word Set Intrusion HITs.

Basic instructions

1. For each HIT, you will see ONE SHORT PASSAGE and FOUR word sets.
2. Three of the word sets will be related to the passage, but the other one will be out of place.
3. Your job is to pick up the one word set that seems LEAST related to the passage.

Background

The passages and words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will highlight the important text and provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

> *Heartening story in the Lawrence Journal-World about 89-year-old WWII Veteran and The University of Kansas Football Team Alumnus Bryan Sperry. He stole the show in Saturday's alumni flag football game.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o veterans, service, national, honor, military, thank, proud, women
- o great, students, school, office, visit, thanks, state, county
- o watch, senate, morning, last, discuss, live, read, news
- o sanders, tax, budget, new, bernie, said, debt, pay

Answer: The correct answer is "sanders, tax, budget, new, bernie, said, debt, pay". The passage is about Bryan Sperry. It is a local interest story that focuses on the achievements of a veteran a football alumni involved in a flag football game.

The word set "veterans, service, national, honor, military, thank, proud, women" clearly relates to his identity as "an 89-year-old WWII Veteran.

The word set "great, students, school, office, visit, thanks, state, county" relates to his identity as an alumnus of the University of Kansas Football Team and the fact that this post is celebrating a local event/achievement.

The passage also references local news coverage (the post is actually about a news story), which is somewhat related to the set "watch, senate, morning, last, discuss, live, read, news."

The word set "sanders, tax, budget, new, bernie, said, debt, pay", however, seems to relate public policy and are not related to a local non-politcial event.

Practice HIT #2

> *Earlier this week I spoke at a town hall in support of Referendum 74 at First Baptist Church in Seattle. Stand with me – and thousands of Washington families – and click below to volunteer in support of marriage equality!*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o watch, senate, morning, last, discuss, live, read, news
- o day, family, people, one, life, every, years, world
- o energy, water, climate, change, oil, federal, clean, regulations
- o great, students, school, office, visit, thanks, state, county

Answer: The correct answer is "energy, water, climate, change, oil, federal, clean, regulations". The passage is in support of a state referendum on gay marriage. It also references a town hall (a local political meeting) held at a church.

Two word sets "great, students, school, office, visit, thanks, state, county" and "watch, senate, morning, last, discuss, live, read, news" relates to the fact that the post references the Senator appearing at a local political gathering.

Another word set "day, family, people, one, life, every, years, world" seems to be related to the fact that the issue concerns "marriage" and "families."

The final word set "energy, water, climate, change, oil, federal, clean, regulations", however, seems to primarily be about environmental policy.

Practice HIT #3

> *The London Metropolitan Police Department has great taste in cars! They're patrolling the streets of London in BMW X5s made in – you guessed it – South Carolina.*
>
> *I'm in London this weekend attending the annual Farnborough International Airshow in support of the Boeing 787 Dreamliners also made in South Carolina. #PalmettoPride #PalmettoPrideintheUK*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o new, jobs, state, help, businesses, work, business, economy
- o veterans, service, national, honor, military, thank, proud, women
- o president, congress, people, senate, american, americans, government, obama
- o great, students, school, office, visit, thanks, state, county

Answer: The correct answer is "president, congress, people, senate, american, americans, government, obama". The passage expresses pride in the cars and airplanes made in South

Carolina. The passage mentions cars, the police, and Boeing airplanes manufactured in South Carolina, and an airshow in London.

One word set "new, jobs, state, help, businesses, work, business, economy" clearly relates to the economy, and is related to the cars and planes manufactured in South Carolina.

A second word set "great, students, school, office, visit, thanks, state, county" again relates to the local nature and focus of the post.

A third word set "veterans, service, national, honor, military, thank, proud, women" is a bit more difficult. Probably it is related to both the reference to the police and the international airshow.

The fourth set of words "president, congress, people, senate, american, americans, government, obama", has no real connection with the passage.

Practice HIT #4

> *Today's executive order from President Trump is more about extreme xenophobia than extreme vetting. This executive order is the equivalent of a "Keep Out" sign posted at America's borders.*
>
> *Turning away immigrants based on their nationality and religion is un-American and in direct opposition to everything for which our Founding Fathers fought. President Trump may not call it a Muslim ban, but it is, and runs afoul of our morals and values.*
>
> *We must ensure that we have the strongest safeguards in place to keep terrorists from ever reaching our shores. And we must fully and thoroughly vet all refugees to screen out any potential terrorist threats.*
>
> *But as conflict and war force millions around the world from their homeland, the United States should welcome more refugees, not less. Suspending the U.S. refugee resettlement program will endanger refugees' lives and tear families apart.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o day, family, people, one, life, every, years, world
- o great, students, school, office, visit, thanks, state, county
- o president, congress, people, senate, american, americans, government, obama
- o committee, security, hearing, secretary, united, states, department, senate

Answer: The correct answer is "great, students, school, office, visit, thanks, state, county". The passage is about President Trump's executive order on immigrants, which is a national-level policy discussion.

The word set "president, congress, people, senate, american, americans, government, obama" is about national politics and, therefore, relevant.

Another word set "committee, security, hearing, secretary, united, states, department, senate" seems to relate to politics and perhaps national security.

The word set "day, family, people, one, life, every, years, world" is a bit tricky. However, it seems to relate to social issues and is triggered by the mention of families.

The final word set "great, students, school, office, visit, thanks, state, county" seems to refer to local politics and local events and is not clearly related to the discussion of national

immigration policy.

Practice HIT #5

> *Women in New Mexico and across the country deserve to make their own decision about family planning. Laws to protect a women's right to choose what's best for her body and well-being should not be restricted to what state she lives in. On the 42nd anniversary of the landmark Roe v. Wade decision, I remain committed to strengthening women's reproductive rights and freedom to make their own personal health care choices. #Roe42*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o president, congress, people, senate, american, americans, government, obama
- o energy, water, climate, change, oil, federal, clean, regulations
- o health, bill, act, care, legislation, help, need, bipartisan
- o day, family, people, one, life, every, years, world

Answer: The correct answer is "energy, water, climate, change, oil, federal, clean, regulations". The passage advocates for a pro-choice position and frames abortion as personal health care choices (family planning).

The word set "health, bill, act, care, legislation, help, need, bipartisan" is about health care and, therefore, relevant.

Another word set "day, family, people, one, life, every, years, world" seems to relate to social issues and is triggered by the mention of families.

The word set "President, congress, people, senate, american, americans, government, obama" is a bit more ambiguous. It relates to legislative politics and the passage is talking about an important issue in legislative politics. Thus, it is also relevant.

The final word set "energy, water, climate, change, oil, federal, clean, regulations" is about environmental issues and not relevant to the passage.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

> *Being conservative means controlling spending and costs. Those without insurance are forced to seek care in expensive ER's after their conditions have worsened. Hospitals are legally required to treat patients in ERs regardless of whether they can pay and they pass the cost of treatment onto the privately insured, increasing total costs. Increasing coverage and allowing more people to manage their health care is cheaper for society. The Cassidy-Collins would accomplish this.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o conservative, congress, people, senate, american, society, government, obama
- o spending, tax, budget, new, cost, price, debt, pay
- o military, veterans, service, national, defence, threat, attack, iran
- o health, bill, act, care, legislation, help, need, insurance

Answer: [**Not Available in the Real Test**] "military, veterans, service, national, defence, threat, attack, iran"

Test HIT #2

> *Today Sen. Judd Gregg and I introduced The Bipartisan Tax Fairness and Simplification Act of 2010, which will help middle-class taxpayers by streamlining and modernizing the outdated tax code. The proposal includes fiscally responsible tax cuts to help working families struggling to make ends meet and also eliminates the corporate tax break that encourages companies to invest overseas rather than creating jobs in the U.S.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o water, river, environment, change, new, future, proud, great
- o senate, bill, act, introduce, legislation, help, need, bipartisan
- o invest, jobs, state, help, campany, work, business, economy
- o sanders, tax, budget, cut, bernie, said, debt, pay

Answer: [**Not Available in the Real Test**] "water, river, environment, change, new, future, proud, great"

Test HIT #3

> *Obama simultaneously could ruin Putin's day and brighten the lives of millions of Americans. All Obama needs is the courage to tell the environmental Left to let him do the right thing.*
>
> *Today is the last day for the public to comment on whether the State Department should approve a presidential permit for the #KeystoneXL pipeline. This article should give you some ideas of things to include in a comment. Here is where you can leave your comment: [link omitted]*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o day, family, people, public, life, every, millon, world
- o energy, water, climate, change, oil, federal, clean, regulations
- o president, congress, people, senate, american, state, government, obama
- o veterans, service, national, honor, military, thank, proud, women

Answer: [**Not Available in the Real Test**] "veterans, service, national, honor, military, thank, proud, women"

Test HIT #4

Look, here is the bottom line. We remain the only nation in the industrialized world that doesn't guarantee health care to all people. We have 29 million people who are uninsured, yet we are spending far, far more per capita on health care than do the people of any other country.

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o health, bill, act, care, legislation, help, need, bipartisan
- o students, school, office, class, young, educare, college, county
- o president, congress, people, senate, american, americans, government, obama
- o spending, tax, budget, new, bernie, said, debt, pay

Answer: [**Not Available in the Real Test**] "students, school, office, class, young, educare, college, county"

Test HIT #5

Like so many Rhode Islanders, I am deeply disturbed by President-elect Trump's appointment of Steve Bannon to be his chief strategist in the White House. As Executive Chairman of Breitbart, Bannon served as a conduit for some of the worst sentiments in our society – hatred and violence on the basis of race, religion, gender, and way of life. As CEO of Trump's campaign, he employed the hateful code of the white supremacist movement to leverage prejudice and fear, for political gain. Now, the President-elect wants Bannon to help guide his administration. Many Rhode Islanders have called and written to tell me how unsettling Bannon's presence in the White House is, and I share their concern. If President-elect Trump will not denounce the dangerous ideas Steve Bannon has represented, it is up to the rest of us to stand firm against them.

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o students, school, office, class, young, educare, college, county
- o president, congress, people, senate, american, americans, government, obama
- o day, family, people, one, life, every, equal, world
- o committee, appointment, hearing, represent, united, states, department, senate

Answer: [**Not Available in the Real Test**] "students, school, office, class, young, educare, college, county"

Test HIT #6

I wish Democrats would show some interest in how and why President Obama and Susan Rice got it so wrong about the true nature of the Benghazi attack. I wish Democrats would show a little interest about why Secretary Clinton was clueless about the multiple security requests coming from Benghazi and how she allowed our mission to become a death trap. I wish Democrats would show a little interest in finding out whether Mike Morell, the former #2 at the CIA, lied to Congress and the American people about a protest that never happened. Democrats seem

to be more interested in protecting the Obama Administration than they are in getting the truth.

If Republicans win a Senate majority in 2014, one of the first things I will insist on is hearings that actually get to the bottom of what happened before, during, and after the Benghazi attack.

The families of those who lost loved ones, and the American people, deserve nothing less than a full accounting.

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o day, family, people, one, life, every, years, world
- o committee, security, hearing, secretary, united, states, department, senate
- o president, congress, people, senate, american, americans, government, obama
- o company, jobs, invest, help, business, work, unemployment, economy

Answer: [**Not Available in the Real Test**] "company, jobs, invest, help, business, work, unemployment, economy"

Test HIT #7

Tomorrow, I'll be attending the White House bipartisan summit on health care reform hosted by President Obama. Share your ideas on health care reform with me now – so I can share your feedback with President Obama and Congressional leaders on Thursday!

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o military, veterans, service, national, defence, threat, attack, iran
- o health, bill, act, care, legislation, help, need, bipartisan
- o president, congress, people, senate, american, americans, government, obama
- o watch, senate, morning, share, discuss, live, feedback, news

Answer: [**Not Available in the Real Test**] "military, veterans, service, national, defence, threat, attack, iran"

Test HIT #8

Today Chairman Johnson held a hearing on duplication, waste, and fraud in federal programs.

Seven years later, with less than half of GAO's recommendations even implemented, GAO estimates that this report has resulted in actual savings of $75 billion. One simple idea has saved American taxpayers tens of billions of dollars. However, there are still hundreds of recommendations that have gone unimplemented, and very little actual duplication in the federal government has been addressed. I am pleased that the Trump administration is taking this problem seriously. The Executive Order signed by President Trump and the memorandum by OMB Director Mulvaney that followed will result in a plan to reorganize and

streamline the federal government and help it better serve the American people. This is long overdue.

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

- o sanders, tax, budget, new, bernie, said, debt, pay
- o committee, order, hearing, federal, united, states, department, senate
- o president, administration, congress, people, senate, americans, government, obama
- o students, school, office, class, young, educare, college, county

Answer: [**Not Available in the Real Test**] "students, school, office, class, young, educare, college, county"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

3.3 Training Module for Random 4 Word Set Intrusion

Completing this training module qualifies you to complete Random Four Word Set Intrusion HITs

Basic instructions

1. For each HIT, you will see FOUR word sets.
2. Three of the word sets will be related to one another, but the other one will be out of place.
3. Your job is to pick up the one word set that seems LEAST related to others.

Background

The words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

Please click on the word set below that is most UNRELATED to the other three.

- o bill, govern, vote, cut
- o senate, congress, pass, plan
- o tax, budget, house, reform
- o report, today, press, emergency

Answer: The correct answer is "respond, today, first-aid, emergency". While three word sets are related to policy-making institutions and procedures ("bill, govern, vote, cut", "senate, congress, pass, plan", "tax, budget, house, reform"), the set of words "respond, today, first-aid, emergency" is not. Rather it seems to relate to time sensitive response and emergencies.

Practice HIT #2

Please click on the word set below that is most UNRELATED to the other three.

- o great, thank, visit, city
- o state, enjoy, tour, center
- o reform, obamacare, federal, year
- o happy, team, celebrate, park

Answer: The correct answer is "reform, obamacare, federal, year". While the other three word sets relate to celebration of local events ("great, thank, visit, city", "state, enjoy, tour, center", and "happy, team, celebrate, park"), the set of words "reform, obamacare, federal, year" is not but rather relates to health care reform policy.

Practice HIT #3

Please click on the word set below that is most UNRELATED to the other three.

- o nation, student, service, school
- o honor, education, veteran, state
- o social, call, secure, post
- o learn, high, college, proud

Answer: The correct answer is "social, call, secure, post". Three of the word sets ("nation, student, service, school", "honor, education, professor, state", and "learn, high, college, proud") relate to education. The other word set ("social, call, secure, post") is not.

Practice HIT #4

Please click on the word set below that is most UNRELATED to the other three.

- o day, family, country, people
- o proud, learn, university, service
- o women, american, live, make
- o work, life, nation, love

Answer: The correct answer is "proud, learn, university, college". While the three other word sets seems to be generally about American ideals and values ("day, family, country, people", "women, american, live, make", and "work, life, nation, love"), the set of words

"proud, learn, university, college" is instead more closely related to education.

Practice HIT #5

Please click on the word set below that is most UNRELATED to the other three.

- o job, business, help, work
- o confirm, judge, nominate, rule
- o energy, economy, state, new
- o develop, continue, import, protect

Answer: The correct answer is "confirm, judge, nominate, rule". While the other three word sets are related to the economy and businesses ("job, business, help, work", "energy, economy, state, new" and "develop, continue, import, protect"), the set of words "confirm, judge, nominate, rule" is more clearly related to judicial appointment and legal issues.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

Please click on the word set below that is most UNRELATED to the other three.

- o discuss, join, week, senate
- o hear, talk, share, live
- o many, world, nation, america
- o office, morning, watch, meet

Answer: [**Not Available in the Real Test**] "many, world, nation, america"

Test HIT #2

Please click on the word set below that is most UNRELATED to the other three.

- o thank, visit, good, time
- o health, care, veteran, help
- o need, provide, support, work
- o drug, program, prevent, ensure

Answer: [**Not Available in the Real Test**] "thank, visit, good, time"

Test HIT #3

Please click on the word set below that is most UNRELATED to the other three.

- o right, law, court, senate
- o vote, justice, supreme, constitution
- o judge, nominate, rule, investigate
- o news, report, read, said

Answer: [**Not Available in the Real Test**] "news, report, read, said"

Test HIT #4

Please click on the word set below that is most UNRELATED to the other three.

- o president, action, security, administration
- o elect, decision, political, confirm
- o nation, threat, trump, deal
- o secretary, congress, immigrant, defense

Answer: [**Not Available in the Real Test**] "elect, decision, political, confirm"

Test HIT #5

Please click on the word set below that is most UNRELATED to the other three.

- o family, legislation, children, bill
- o news, today, said, report
- o governor, state, press, call
- o federal, emergency, post, continue

Answer: [**Not Available in the Real Test**] "family, legislation, children, bill"

Test HIT #6

Please click on the word set below that is most UNRELATED to the other three.

- o west, proud, community, learn
- o university, force, military, nation
- o student, service, member, thank
- o veteran, help, need, support

Answer: [**Not Available in the Real Test**] "veteran, help, need, support"

Test HIT #7

Please click on the word set below that is most UNRELATED to the other three.

- o administration, deal, threat, state
- o small, support, community, economy
- o create, continue, import, protect
- o job, business, help, work

Answer: [**Not Available in the Real Test**] "administration, deal, threat, state"

Test HIT #8

Please click on the word set below that is most UNRELATED to the other three.

- o justice, elect, supreme, political
- o decision, protect, general, constitution
- o right, law, court, senate
- o bill, american, tax, people

Answer: [**Not Available in the Real Test**] "bill, american, tax, people"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

3.4 Training Module for Label Intrusion

Completing this training module qualifies you to complete Label Intrusion HITs

Basic instructions
1. For each HIT, you will see ONE SHORT PASSAGE and FOUR labels.
2. Three of the labels will be related to the passage, but the other one will be out of place.
3. Your job is to pick up the one label that seems LEAST related to the passage.

Background

The passages you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Using a computer algorithm, we divided these posts into different categories based on the words they contain. The labels you see are our tentative summaries of those issues or topics.

We want to see if you can identify the label that does not belong with the post.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will highlight the important text and provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

> *Alexander Cosponsor's Cut, Cap, and Balance Act: At a time when we're borrowing 40 cents of every dollar we spend, I will support serious proposals like the "Cut, Cap and Balance Act" to reduce out-of-control Washington spending. The final version of any such legislation should have an appropriate balance between reductions in both entitlement and discretionary spending, but this bill is a good start.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Government Spending
- o Legislation
- o American Politics
- o Emergency

Answer: The correct answer is "Emergency". The passage talks about a legislation on the federal budget cut. Thus, the labels "Legislation" and "Government Spending" are clearly

relevant. This is an important issue in "American Politics," which is also relevant as a label. The remaining label "Emergency," however, is out of place because the budget cut is usually not an emergent event.

Practice HIT #2

> *Earlier this week I spoke at a town hall in support of Referendum 74 at First Baptist Church in Seattle. Stand with me – and thousands of Washington families – and click below to volunteer in support of marriage equality!*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Local Events
- o Public Meeting
- o Environment
- o Human Wellbeing

Answer: The correct answer is "Environment". The passage is in support of a state referendum on gay marriage. It also references a town hall (a local political meeting) held at a church. Two word sets "Local Events" and "Public meeting" relate to the fact that the post references the Senator appearing at a local political gathering. Another word set "Human Wellbeing" seems to be related to the fact that the issue concerns "marriage" and "families." The final label "Environment," however, is not mentioned.

Practice HIT #3

> *The London Metropolitan Police Department has great taste in cars! They're patrolling the streets of London in BMW X5s made in – you guessed it – South Carolina.*
>
> *I'm in London this weekend attending the annual Farnborough International Airshow in support of the Boeing 787 Dreamliners also made in South Carolina. #PalmettoPride #PalmettoPrideintheUK*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Economy
- o Local Events
- o Presidential Politics
- o Police

Answer: The correct answer is "Presidential Politics". The passage expresses pride in the cars and airplanes made in South Carolina. Those products are about "Economy." South Carolina is local and the airshow belongs to "Local Events." Since the police department is mentioned, the label "Police" should also be relevant. "Presidential Politics" has no real connection with the passage.

Practice HIT #4

Today's executive order from President Trump is more about extreme xenophobia than extreme vetting. This executive order is the equivalent of a "Keep Out" sign posted at America's borders.

Turning away immigrants based on their nationality and religion is un-American and in direct opposition to everything for which our Founding Fathers fought. President Trump may not call it a Muslim ban, but it is, and runs afoul of our morals and values.

We must ensure that we have the strongest safeguards in place to keep terrorists from ever reaching our shores. And we must fully and thoroughly vet all refugees to screen out any potential terrorist threats.

But as conflict and war force millions around the world from their homeland, the United States should welcome more refugees, not less. Suspending the U.S. refugee resettlement program will endanger refugees' lives and tear families apart.

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Immigration
- o Local Events
- o Presidential Politics
- o Security and Defence

Answer: The correct answer is "Local Events". The passage is about President Trump's executive order on immigrants. The labels "Presidential Politics" and "Immigration" are thus clearly relevant. The sentence "We must ensure that we have the strongest safeguards in place to keep terrorists from ever reaching our shores." implies that "Security and Defence" are also concerned. "Local Events" is not clearly related to the discussion of national immigration policy.

Practice HIT #5

Women in New Mexico and across the country deserve to make their own decision about family planning. Laws to protect a women's right to choose what's best for her body and well-being should not be restricted to what state she lives in. On the 42nd anniversary of the landmark Roe v. Wade decision, I remain committed to strengthening women's reproductive rights and freedom to make their own personal health care choices. #Roe42

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Legislation
- o Environment
- o Healthcare
- o Daily Life

Answer: The correct answer is "Environment". The passage advocates for a pro-choice position in legislative politics and thus "Legislation" is clearly relevant. It frames abortion as personal "Healthcare" choices (family planning). The label "Daily Life" is triggered by

the mention of families. "Environment" is not relevant to the passage.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

> *The unexpected and tragic attacks on Pearl Harbor that occurred 75 years ago today–a date that even now still lives in infamy–triggered America to go to war. As a result of the unprovoked attacks in 1941, many of our brave servicemen and women lost their lives. Today and every year on December 7th, I hope we each take a moment to reflect on and pay tribute to the sacrifices made by those who were injured or lost at Pearl Harbor. And let us each re-dedicate ourselves to ensuring their lives were not lost in vain by honoring and upholding the quintessential American values and freedoms they were fighting for.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o American History
- o Security and Defence
- o Education
- o Honoring Veterans

Answer: [**Not Available in the Real Test**] "Education"

Test HIT #2

> *Today Sen. Judd Gregg and I introduced The Bipartisan Tax Fairness and Simplification Act of 2010, which will help middle-class taxpayers by streamlining and modernizing the outdated tax code. The proposal includes fiscally responsible tax cuts to help working families struggling to make ends meet and also eliminates the corporate tax break that encourages companies to invest overseas rather than creating jobs in the U.S.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Immigration
- o Legislation
- o Economy
- o Tax and Budget

Answer: [**Not Available in the Real Test**] "Immigration"

Test HIT #3

Obama simultaneously could ruin Putin's day and brighten the lives of millions of Americans. All Obama needs is the courage to tell the environmental Left to let him do the right thing.

Today is the last day for the public to comment on whether the State Department should approve a presidential permit for the #KeystoneXL pipeline. This article should give you some ideas of things to include in a comment. Here is where you can leave your comment: [link omitted]

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Human Wellbeing
- o Environment
- o Presidential Politics
- o Honoring Veterans

Answer: [**Not Available in the Real Test**] "Honoring Veterans"

Test HIT #4

Look, here is the bottom line. We remain the only nation in the industrialized world that doesn't guarantee health care to all people. We have 29 million people who are uninsured, yet we are spending far, far more per capita on health care than do the people of any other country.

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Healthcare
- o Education
- o Government Spending
- o Human Wellbeing

Answer: [**Not Available in the Real Test**] "Education"

Test HIT #5

In Warroad, MN, Hockeytown USA, with 3 proud U.S. Hockey silver medalists: Rubin Bjorkman (1952), Gordon Christian (1956) and Henry Boucha (1972). Warroad boasts 7 Olympic hockey medalists with Gigi Marvin and TJ Oshie set to win more. Oshie's grandpa played on 1948 Warroad team! Go USA hockey!

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Healthcare
- o Announcement
- o Game
- o Celebration

Answer: [**Not Available in the Real Test**] "Healthcare"

Test HIT #6

I wish Democrats would show some interest in how and why President Obama and Susan Rice got it so wrong about the true nature of the Benghazi attack. I wish Democrats would show a little interest about why Secretary Clinton was clueless about the multiple security requests coming from Benghazi and how she allowed our mission to become a death trap. I wish Democrats would show a little interest in finding out whether Mike Morell, the former #2 at the CIA, lied to Congress and the American people about a protest that never happened. Democrats seem to be more interested in protecting the Obama Administration than they are in getting the truth.

If Republicans win a Senate majority in 2014, one of the first things I will insist on is hearings that actually get to the bottom of what happened before, during, and after the Benghazi attack.

The families of those who lost loved ones, and the American people, deserve nothing less than a full accounting.

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o The Public
- o Political Parties
- o Presidential Politics
- o Economy

Answer: [**Not Available in the Real Test**] "Economy"

Test HIT #7

Tomorrow, I'll be attending the White House bipartisan summit on health care reform hosted by President Obama. Share your ideas on health care reform with me now – so I can share your feedback with President Obama and Congressional leaders on Thursday!

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Honoring Veterans
- o Healthcare
- o American Politics
- o Public Participation

Answer: [**Not Available in the Real Test**] "Honoring Veterans"

Test HIT #8

Yesterday I met with Anjali Lall and Maneesh Apte, U.S. Presidential Scholar recipients who recently graduated from Davies High School in Fargo, ND. The United States Presidential Scholar Program annually awards up to 141 high school graduates nationwide, honoring students who demonstrate exceptional academic achievement. Congratulations Anjali and Maneesh!

Please read the four labels below and click on the label that is most UNRELATED to the passage.

- o Honorary Statement
- o Education
- o President
- o Economy

Answer: [**Not Available in the Real Test**] "Economy"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

3.5 Training Module for Optimal Label

Completing this training module qualifies you to complete Optimal Label HITs.

Basic instructions

1. For each HIT, you will see ONE SHORT PASSAGE and FOUR labels.
2. Your job is to pick up the one label that seems BEST summarizing the passage.

Background

The passages you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Using a computer algorithm, we divided these posts into different categories based on the words they contain. The labels you see are our tentative summaries of those issues or topics.

We want to see if you can identify the label that best summarizes the post.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will highlight the important text and provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

Alexander Cosponsor's Cut, Cap, and Balance Act: At a time when we're borrowing 40 cents of every dollar we spend, I will support serious proposals like the "Cut, Cap and Balance Act" to reduce out-of-control Washington spending. The final version of any such legislation should have an appropriate balance between reductions in both entitlement and discretionary spending, but this bill is a good start.

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Insurance
- o Legal/Law
- o Immigration
- o Budget Cut

Answer: The correct answer is "Budget Cut". The passage talks about a legislation on the federal budget cut. The label "Budget Cut" best summarizes it. "Legal/Law", "Immigration", and "Immigration" are not relevant.

Practice HIT #2

In Warroad, MN, Hockeytown USA, with 3 proud U.S. Hockey silver medalists: Rubin Bjorkman (1952), Gordon Christian (1956) and Henry Boucha (1972). Warroad boasts 7 Olympic hockey medalists with Gigi Marvin and TJ Oshie set to win more. Oshie's grandpa played on 1948 Warroad team! Go USA hockey!

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Economy
- o Holiday Greetings
- o Celebration Messages
- o Information Sources

Answer: The correct answer is "Celebration Messages". The passage celebrates a town called "Warroad" for having many hockey medalists. "Celebration Messages" is thus the optimal label. The purpose of this message is not greeting people, so "Holiday Greetings" is not the best answer. While providing people with some information, the passage is not mainly about "Information Sources." "Economy" is not relevant.

Practice HIT #3

Yesterday I met with Anjali Lall and Maneesh Apte, U.S. Presidential Scholar recipients who recently graduated from Davies High School in Fargo, ND. The United States Presidential Scholar Program annually awards up to 141 high school graduates nationwide, honoring students who demonstrate exceptional academic achievement. Congratulations Anjali and Maneesh!

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Law Enforcement
- o Honoring Veterans
- o Education
- o Presidential Politics

Answer: The correct answer is "Education". The passage congratulates a student for receiving the U.S. Presidential Scholar Award. It is clearly an educational issue and, thus, the label "Education" is optimal. Although the award's name has the word, presidential, in it, the passage is not about presidential politics. "Law Enforcement" and "Honoring Veterans" are not implied.

Practice HIT #4

> *The unexpected and tragic attacks on Pearl Harbor that occurred 75 years ago today–a date that even now still lives in infamy–triggered America to go to war. As a result of the unprovoked attacks in 1941, many of our brave servicemen and women lost their lives. Today and every year on December 7th, I hope we each take a moment to reflect on and pay tribute to the sacrifices made by those who were injured or lost at Pearl Harbor. And let us each re-dedicate ourselves to ensuring their lives were not lost in vain by honoring and upholding the quintessential American values and freedoms they were fighting for.*

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Terrorist Attack
- o Honoring Veterans
- o Water Pollution
- o American History

Answer: The correct answer is "Honoring Veterans". The passage honors victims of the attacks on Pearl Harbor and specifically mentions "brave servicemen and women." "Honoring Veterans" is the best summary. Although the attack took place 75 years ago, this message is not mainly about "American History." The Pearl Harbor attack is not a "Terrorist Attack." "Water Pollution" is not relevant.

Practice HIT #5

> *The Red River Corridor Fund received more than $1.1 million in federal funding to incentivize private companies to help small businesses and local economies grow. When our small businesses have an environment they can thrive in, it creates more jobs and helps grow our communities and state. These funds will help encourage more private-public partnerships across the state so small businesses, private lenders, and the federal government can work collaboratively to build North Dakota's already growing economy.*

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Unemployment Rate
- o Economy
- o Local Issues
- o Healthcare

Answer: The correct answer is "Economy". The passage advertises that North Dakota received federal funds supporting small businesses and local economies. "Economy" is relevant. "Unemployment Rate" is attempting given that the passage mentioned "creates more

jobs". However, it is not optimal because it cannot cover other aspects of the economy such as "funding" and "private-public partnerships." "Local Issues" is too vague. "Healthcare" is not relevant.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

> *I'm hosting a live telephone town hall call this Monday, April 24 at 5:45pm MT. If you're available to talk on the phone, sign up to get a call. You can also participate and ask questions by streaming the event online. Click on the link below to do both: [link omitted]*

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Drug Abuse
- o Legislative Politics
- o Education
- o Town Hall Meeting

Answer: [**Not Available in the Real Test**] "Town Hall Meeting"

Test HIT #2

> *Today, Congress has unanimously passed the Clay Hunt SAV Act to improve veteran suicide prevention programs. We must do everything we can to ensure our veterans and military families have access to high-quality mental health services. By improving treatment, intervention, and outreach, the Clay Hunt SAV Act will help more veterans get the help that they need and have earned. No veteran should ever have to wait for critical mental health care. Learn more: [link omitted]*

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Veteran Healthcare
- o Economy
- o Drug Abuse
- o Nation Building

Answer: [**Not Available in the Real Test**] "Veteran Healthcare"

Test HIT #3

> *Today I met with Judge Neil Gorsuch, nominee to serve on the U.S. Supreme Court. Judge Gorsuch is a smart, diligent and thoughtful jurist. I am impressed by his remarkable commitment to the Constitution and the separation of power our founders envisioned. Oklahoma is within the jurisdiction of the 10th Circuit*

Court of Appeals, on which Judge Gorsuch sits, and I have seen his judicial philosophy first hand. In Burwell v. Hobby Lobby, he wrote a concurring opinion in favor of the Oklahoma company's position, and upholding our First Amendment right to religious freedom. Judge Gorsuch is a qualified mainstream jurist who is well respected by conservatives and liberals alike and I urge Senate Democrats to allow and up or down vote on Gorsuch. I look forward to confirming him to the bench.

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Supreme Court
- o Congratulation Messages
- o Military
- o Religious Freedom

Answer: [**Not Available in the Real Test**] "Supreme Court"

Test HIT #4

Recent actions by Iran have once again demonstrated their reluctance to curb their confrontational nuclear missile program. These dangerous, provocative actions by a country considered a state sponsor of terror should carry consequences. The decision to increase sanctions should come at no surprise. A line was drawn, and Iran deliberately defied the agreement. I stand by the President on his decision to implement a quick and deliberate strategy. It should send a clear message that Iran's reckless behavior will not be tolerated.

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Justice
- o Security
- o Honoring Veterans
- o Government Spending

Answer: [**Not Available in the Real Test**] "Security"

Test HIT #5

Deadline extended: January 23 is the new deadline to apply for federal disaster assistance from Hurricane Matthew. Register with FEMA Federal Emergency Management Agency online: [link omitted]

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Emergency Management
- o Holiday Greetings
- o Foreign Policy
- o Public Hearing

Answer: [**Not Available in the Real Test**] "Emergency Management"

Test HIT #6

Interested in attending one of our nation's military academies? My office is currently accepting applications from North Dakota students interested in attending the U.S. Military Academy at West Point, the U.S. Naval Academy, U.S. Air Force Academy or the U.S Merchant Marine Academy. The deadline is Oct. 5.

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Healthcare
- o Doctor
- o American History
- o Application

Answer: [**Not Available in the Real Test**] "Application"

Test HIT #7

Today I brought together local business development and community leaders to discuss the essential role of rural communities in North Dakota's economic success and ways to support new investment and growth in small towns across the state. Energy development in the Bakken has created unprecedented growth but we have to make sure this development's success flows across the entire state. Rural communities like Ellendale are an important part of that effort, and by supporting hardworking North Dakotans and small businesses in these areas, we can continue to sustainably grow our state for future generations.

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Local Businesses
- o Nation Building
- o Infrastructure Construction
- o Public Meeting

Answer: [**Not Available in the Real Test**] "Local Businesses"

Test HIT #8

While under oath during his confirmation hearing, Attorney General Jeff Sessions misled his colleagues in U.S. Senate about his communications with Russia Ambassador Sergey Kislyak. He should resign immediately. This is exactly why we need an independent special counsel to investigate the Russian hacking of our presidential election and any ties the Trump campaign may have had to Russia.

Please read the four labels below and click on the label that BEST summarizes the passage.

- o Family Memory
- o Information Sources
- o Games
- o Legal Institutions and Offices

Answer: [**Not Available in the Real Test**] "Legal Institutions and Offices"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

3.6 Workers' Performance

To assess workers' performance, we randomly mixed in a gold-standard HIT into every ten HITs. Once posted a batch, we checked its progress sporadically. We suppressed the qualification of workers who have missed more than 2 gold-standard HITs or who have done a relatively large number of HITs of a specific task structure. This operation has no negative impact on their Mturk records. Among all workers who have completed HITs in the batches demonstrated in the main manuscript, we have rejected and replaced work from only three of them who missed more than 4 gold-standard HITs each. These thresholds (2 and 4) are not iron rules. Future researchers can decide on their own standards based on the difficulty of their designed gold-standard HITs.

The gold-standard HIT approach yields consistent results. In the main manuscript, none of the identical pairs we presented are significantly different from each other.

Another way to assess the consistency of workers' performance across identical trails is to assess the "agreement rate," which we present in Tables SI3 and SI4. The numbers reflect the frequency that workers from the two trials get the task either right or wrong simultaneously, as opposed to one get it right but the other get it wrong.

Table SI3: Workers' Agreement Rate in the Topic Validation Tasks

	Model 1	Model 2	Model 3	Model 4	Model 5
Tasks in Paper					
WI	0.816	0.696	0.724	0.696	0.684
T8WSI	0.62	0.606	0.68	0.72	0.704
R4WSI	0.714	0.754	0.85	0.884	0.728
Tasks in SI					
R4WSI-Random	0.780	0.912	0.914	0.932	0.851
	Model 6	Model 7	Model 8	Model 9	
R4WSI-Representative	0.920	0.858	0.854	0.566	

We realize that these numbers are hard to interpret. In general, a high agreement rate is desirable. However, the agreement rate would, theoretically, be a function of the correct rate. This introduces a floor effect. Consider the case, for instance, where workers got 95% of the HITs correct, agreement rates can only go as low as 90%. For this reason, we did not include these tables in the main text, but only as supplementary information.

Table SI4: Workers' Agreement Rate in the Label Validation Tasks

	Careful Coder	Word Coder
Tasks in Paper		
LI-Within	0.764	0.746
LI-Across	0.864	0.896
OL-Within	0.836	0.824
OL-Across	0.904	0.856
Tasks in SI		
WSI	0.860	0.856
OLW	0.908	0.962

4 More on the Corpus, Topic Model Fit, and Labeling

This section presents more background information and discussions on the structural topic models in the paper and their labeling process. Specifically, we first discuss the distributions of word probability in the topic models, which is the reason we chose to draw words based on their probability in the topic validation tasks. We then present some word clouds to give people a different perspective of the topics in different topic models. Finally, we present the representative documents based on which we came up with the topic labels.

4.1 Word Mass Distribution

One major difference between our topic validation tasks and those from Chang et al. (2009) is that we randomly draw words based on their probabilities in a given topic from a given model. We made this choice because the word probabilities could vary a lot among the top 20 high probability words. Figures SI18 and SI19 demonstrate this point with Model 2 (the 10-topic model) from our paper, where `beta` refers to the word probability.

Furthermore, Figure SI20 depicts the total word probability covered by the top n words for the four converged topic models (Model2, Model3, Model4, and Model5) from our paper. The shapes of the curves are similar, although the specific numbers vary across different models. On average, except Model2, the top 20 words cover about 50% of the total word probability.

Figure SI18: Word Mass: Beta

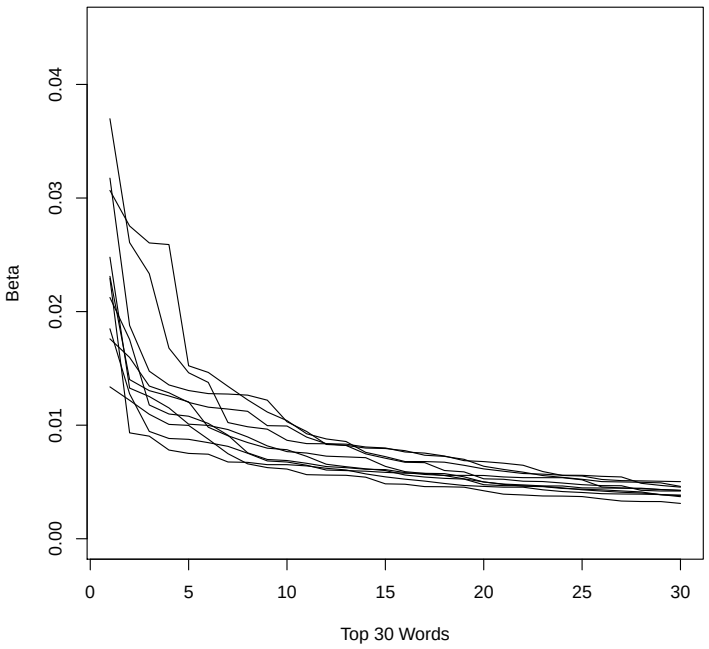


Figure SI19: Word Mass: Log-Beta

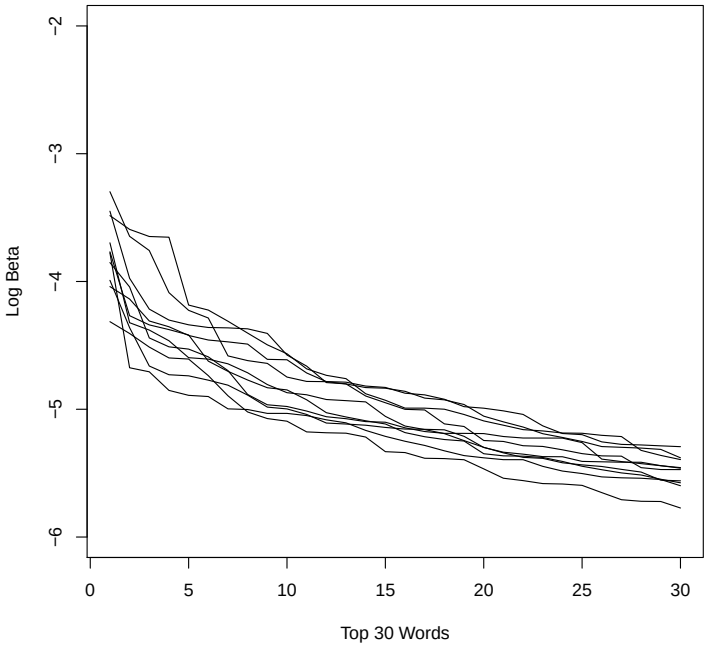
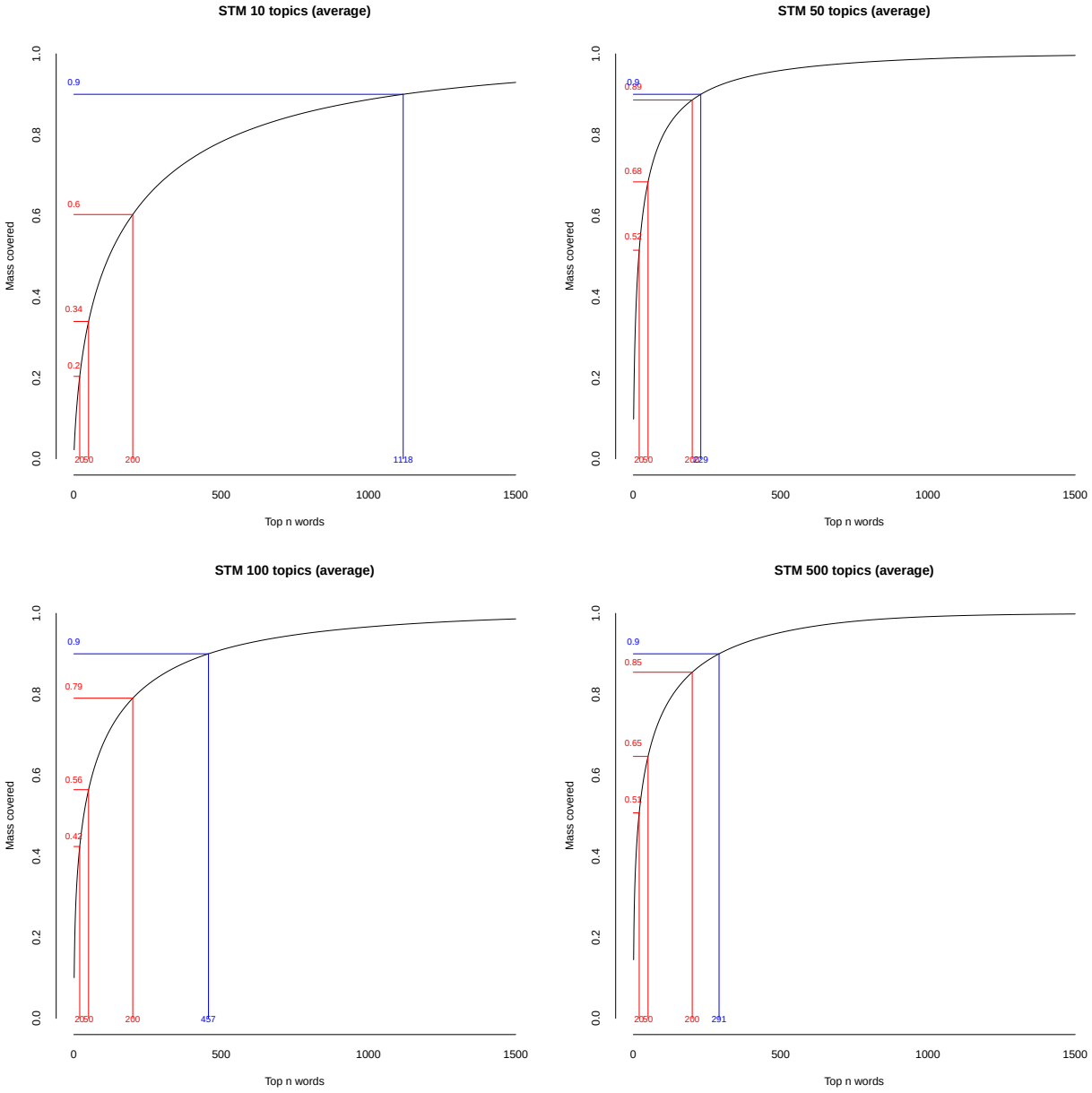


Figure SI20: Word versus Mass



4.2 Word Clouds

For the four structural topic models validated in the paper, we randomly select 6 topics from each of them and present the word clouds below. Just from the word cloud, it does not seem obvious that the four converged models (Model2, Model3, Model4, and Model5) are of different quality.

Figure SI21: Word Cloud: Six topics from **model 1** (the 10-topic model, 1 iteration)

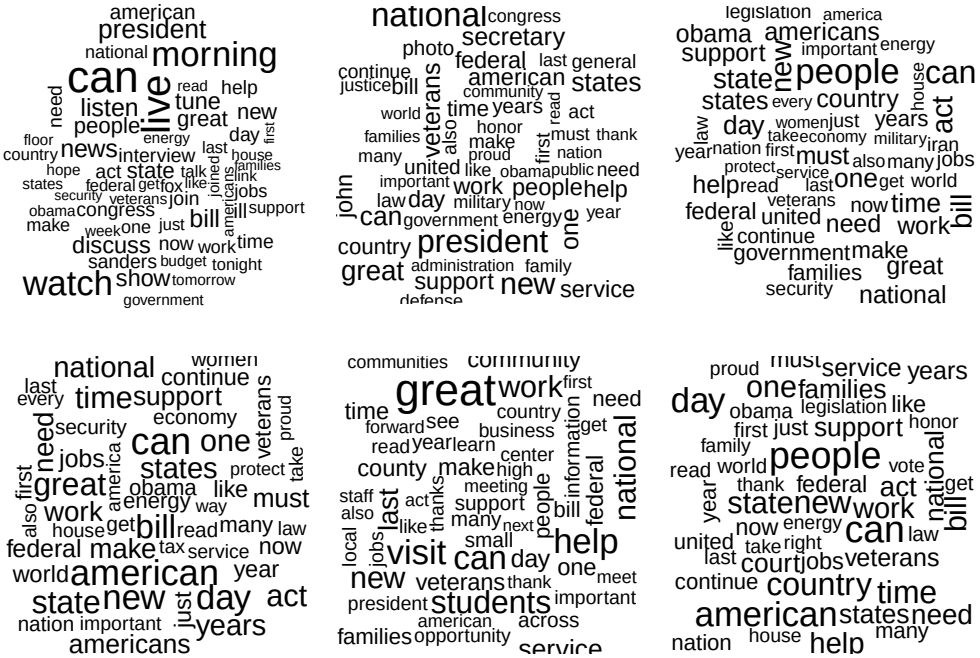


Figure SI22: Word Cloud: Six topics from **model 2** (the 10-topic model)

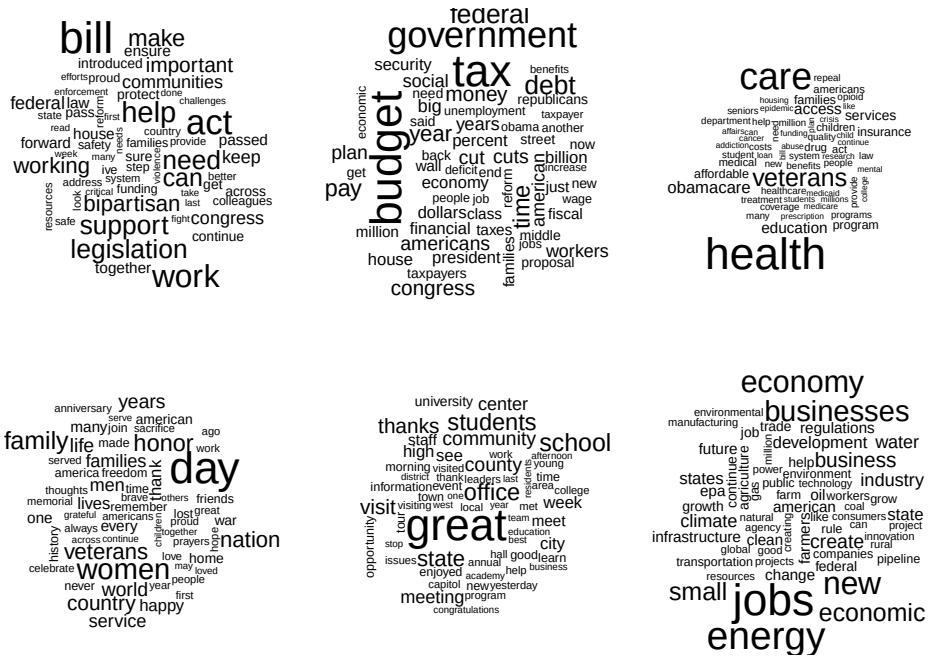


Figure SI23: Word Cloud: Six topics from **model 3** (the 50-topic model)

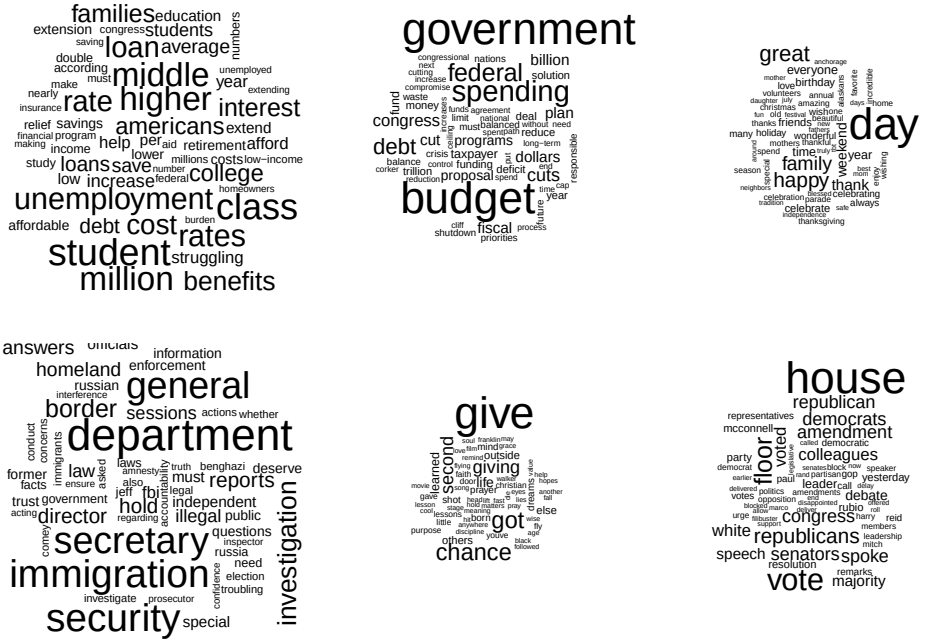
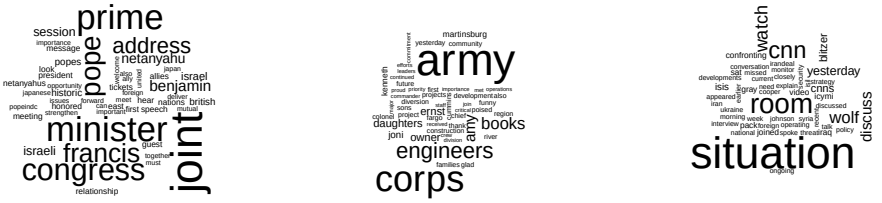


Figure SI24: Word Cloud: Six topics from **model 4** (the 100-topic model)

Figure SI25: Word Cloud: Six topics from **model 5** (the 500-topic model)



4.3 Representative Documents for Labeling

For the label validation tasks, we asked the two careful coders to read carefully the top 50 representative documents. Table SI5 shows the top 5 documents for each domestic topic with which they come up with the labels.

Table SI5: Representative Documents for Domestic Topics

Label (Careful Coder)	Document
Equal Pay for Women	(1) In Indiana, women working full-time make an average of 75 cents for every dollar earned by their male colleagues. Women in Indiana and around the country deserve to earn equal pay for equal work. When women are paid less, it not only impacts their ability to save for retirement but also means families have less to spend on groceries, rent, and other basic expenses. #EqualPayDay (2) Sweat, sweat, sweat! Work and sweat, cry and sweat, pray and sweat! Zora Neale Hurston #Hardwork (3) Women earn less than men regardless of any other factor. I support #EqualPay for equal work & that's why I voted for the Paycheck Fairness Act that builds on the Lilly Ledbetter Fair Pay Act by closing loopholes that enable pay discrimination against women. Share this graphic if you agree that women should earn equal pay for equal work! (4) Today is Equal Pay Day. North Dakota women earn 73 cents for every $1 a man earns. I support the Paycheck Fairness Act because there should be equal pay for equal work. Like/share if you agree. (5) In Michigan, women earn 74 cents for every dollar men earn. The Gender Pay Gap doesnt just hurt women - it hurts entire families, affecting retirement, health care coverage, pensions and much more. Today is Equal Pay Day, a date that symbolizes how far into the year women must work to earn what men earned in the previous year. It is time to end the Gender Pay Gap and pay women fair wages for their work.
Healthcare/ Reproductive Rights	(1) Planned Parenthood provides critical access to health care for thousands of Montanans. #StandWithPP (2) Courtney from Wisconsin stands with Planned Parenthood because without them, she wouldn't have access to quality and affordable womens health care. TrumpCare rolls back the clock on womens health by cutting funding for maternity care and defunding Planned Parenthood. (3) I #StandwithPP for the millions who depend on their critical health care services. #PinkOut

	(4) California health centers received funds to expand primary care services by hiring qualified health care providers. (5) Planned Parenthood is providing women across Michigan with critical preventative care, including cancer screenings and wellness exams. Today, I met with Planned Parenthood Advocates of Michigan to let them know I #StandWithPP and the millions of women who rely on them for quality health care. #PinkOut
Agriculture	(1) Antibiotic overuse unchecked: "About four-fifths of all antibiotics sold in the U.S. go to livestock and poultry." (2) Montana ranchers produce the best beef in the world and we cant have beef imports from Brazil and Argentina jeopardizing the livelihood of Montana producers. (3) Today we recognize and thank all who contribute to Iowas agriculture industry, among them Iowas farmers, producers, ranchers: [Link Omitted] #AgDay (4) SenatorLeahy: #VT farmers, sugar makers and rural farms benefited from REAP in 2008 Farm Bill. #FarmBill 2012 continues this good work. (5) Senator Shaheen discussed improvements made to the dairy program in this years farm bill during a tour of McNamara Dairy farm in Plainfield. Employees of McNamara Dairy milk 175 Holstein cows, bottle their own milk and sell it to customers up and down the Connecticut River Valley. (Plainfield, NH. July 20, 2012)
Student Loan/ Debt	(1) #ObamaEconomy: Inflation-adjusted median household income has fallen from $54K in '08, to only $52K in 2013 while inflation-adjusted per capita income has fallen from $29,173 in 2008 to $28,829 (2) Recent college graduates are struggling in this slow economic recovery. We must keep student loan interest rates at the current rate and prevent them from doubling, but we must pay for it. (3) In our latest installment of Correspondence from the Commute, Chris replies to Bailey from Hockessin about Joseph Kony and the LRA. (4) Americans have over $1.3 trillion of student debt. My In #TheRedAct will allow struggling borrowers to refinance their student loans and take advantage of lower interest rates the same way people refinance a mortgage, a car loan or business debt. (5) FACT: Total student loan debt now surpasses both credit card & auto loan debt. #HigherEdNotDebt
Drug Abuse	(1) DEA Implements Cornyn-Klobuchar Law to Help Curb Prescription Drug Abuse:

	(2) Today, I teamed up with Senator Sheldon Whitehouse in hosting a roundtable with the Alliance to Prevent the Abuse of Medicines on how we can work together to combat prescription drug abuse and help struggling families and communities in Ohio. To prevent drug abuse and better help the tens of thousands of Ohioans struggling with addiction, we need a comprehensive strategy that starts from the bottom up. Our bipartisan Comprehensive Addiction and Recovery Act builds on proven methods to enable law enforcement to respond to the heroin epidemic and supports long-term recovery by connecting prevention and education efforts with treatment programs. Like my video to join me in the fight against drug abuse. [Link Omitted] (3) Today I re-introduced the Budgeting for Opioid Addiction Treatment (LifeBOAT) Act to establish reliable funding to expand access to substance abuse treatment in West Virginia. (4) Help prevent the misuse of addictive prescription drugs on Prescription Drug Take-Back Day at a location near you [Link Omitted] (5) KSDK NewsChannel 5 previews Claire's roundtable discussion today at NCADA to tackle the opioid and heroin epidemic with advocates, providers, and law enforcement.
Higher Education/ Job Training	(1) Very productive college affordability roundtable with students from University of Wisconsin-Parkside, Gateway Technical College, Carthage College & University of Wisconsin - Whitewater. Higher education should be a path to shared prosperity, not a path into suffocating debt. (2) Honored to be at Northwestern Michigan College yesterday for the expansion of their Community College Skilled Trades Equipment Program. With this new program, NMC will be able to offer wider array of educational opportunities - from marine and aviation tech to welding and nursing. (3) Happy to join Community College of Rhode Island (CCRI) today to announce a $2.5 million federal grant to create new pathways and opportunities in advanced manufacturing. CCRI's Accelerated Pathways in Advanced Manufacturing program emphasizes a learn and earn model that provides opportunities for adult learners to acquire new knowledge and skills that are linked with jobs in high-growth industries. (4) As part of her efforts to help New Hampshire workers develop the skills and innovative thinking needed for jobs in the 21st century economy, Governor Maggie Hassan announced today that five New Hampshire companies have been awarded job training grants to help them train 171 workers in new skills. The job training grants total $144,973.50, and the companies contributed matching funds to bring the total amount for training workers to $289,947.

	(5) Continuing her efforts to help New Hampshire workers develop the skills, knowledge and innovative thinking necessary for jobs in the 21st century economy, Governor Hassan announced today that five companies will receive job training grants to help them train 125 workers in new skills. The job training grants total $72,536.48. The five companies contributed matching funds, bringing the total funds for training workers to $145,072.96.
Wall Street/ Financial Sector	(1) SenJohnMcCain: Wall Street bonuses: Goldman Sachs $16.7 billion; JP Morgan $8.7 billion; main street $0 your tax doll... (2) This recession was caused by the greed, the recklessness and the illegal behavior on Wall Street. [Link Omitted] (3) Will question CEO Jamie Dimon about J.P. Morgan's $2 billion plus losses at a Senate Banking hearing around 11am. Tune in here: [Link Omitted] (4) I'm working with Consumer Financial Protection Bureau head Rich Cordray to protect consumers from predatory for-profit schools & big bank tricks. (5) No, AIG. Criticism of Wall Street bonuses is not equivalent to lynchings. Offensive.
Government Shutdown/ Congressional Budget	(1) Republicans in Congress are passing bills to reopen and fund essential functions of our government. It's the Democrats in Congress who refuse to negotiate. #HarryReidsShutdown (2) How does federal government claim preemption in Arizona when federal government fails to act? (3) This week, Puerto Rico argued before the Supreme Court that it should be able to restructure its debt. If Puerto Rico were a country, it could go to the IMF. If it were a city, it could declare bankruptcy. But because Puerto Rico is a U.S. territory, it can do neither. Without help, Puerto Rico is at the mercy of Wall Street vulture funds, creditors that specialize in preying on borrowers who are in trouble. The Republican leadership needs to step in immediately and help millions of American families who are caught in an economic catastrophe. (4) Skipping the budgeting, authorizing and appropriating committee process in lieu of last-minute, omnibus, must-pass, stop-gap bills diminishes the role of Congress. It diminishes the role of individual members of Congress and reduces the input of the people from that members state or district. By skipping steps, Congress allows the Administration to regulate and spend with little fear of being checked by the legislative branch. Our countrys founders gave Congress the power of the purse. The Senate and House should right the ship no matter what party controls the White House. Using the committee system and considering legislation as it should be considered would go a long way. It would put a stop to a lot of partisanship, some of which is simply frustration from leaders preventing members from legislating.

	(5) An essential premise of good government is that Congress should authorize programs and activities before it funds them. If we relinquish our responsibility to regularly review and reform these programs, all of our government funding will essentially operate on auto-pay.
Obamacare/ Tax Policy	(1) #99countymeetings 50ppl at Mapleton Q&A EPAregs VoterID MedicareFraud Obamacare/SupremeCt Farmbill F&F ElectionFinanceRecorm SocSec (2) Courtesy of ObamaCare: Broken promises, plagued by failure & cost skyrocket for millions. (3) #ObamaCare creates fewer options at "much higher premiums, higher out-of-pocket costs, higher taxes on the costs that they[everyday Americans] do incur, and fewer jobs and fewer hours for those who are employed." (4) Millions who lost their insurance and reenrolled through OCare are facing higher premiums: [Link Omitted] (5) Premiums are soaring, patients choices are dwindling. Obamacare must be repealed and replaced. #RepealObamacare
Deficits/ Debt/ Budget	(1) The President's budget proposes deficits of $5.3 trillion and $8.1 trillion in new debt over the next decade. Where's the Balanced Budget? (2) Why we must cut spending and why simply raising taxes won't solve our fiscal problems (3) Today the Congressional Budget Office released a report that shows positive impact of the budget caps. If our nation is serious about balancing our budget and reducing Americas debt, real, substantive budget reforms and savings will have to be on the table during any spending negotiations. (4) Voted NO on raising debt ceiling without addressing spending problem. Need to both prevent a debt limit crisis today & debt crisis tomorrow. (5) The combination of new spending with the President's proposed budget freeze will only maintain an unsustainable status quo. We need serious cuts, not a continuation of the trillion dollar deficits that are jeopardizing our nations fiscal stability.

Table SI6 shows the top 5 documents for each domestic topic with which they come up with the labels.

Table SI6: Representative Documents for International Topics

Label (Careful Coder)	Document
International Trade	(1) Sherrod reintroduces China trade plan to discourage currency manipulation that harms American exports. As our trade deficit continues to widen, our need to level the playing field for American manufacturers and workers becomes more urgent.

	(2) RT @GrassleyOffice: Heres #TheScoop by @ChuckGrassley [Link Omitted] #4jobs @RedCross @Mail4Heroes #fastandfurious #NationalGuard ... (3) Chinas economy is simply too large for it to be artificially propped up by a blatantly manipulated Yuan. Chinas actions are deliberate and are designed to give China a competitive advantage in the marketplace. I believe free and fair trade is beneficial, but I also know that Chinese currency manipulation and intellectual property theft is doing harm to our economy. (4) I am pleased that the U.S. Trade Representative took action against Chinas unfair trade practices. Its policies give Chinese companies an unfair advantage and hurt American workers and products. (5) Today, I urged the International Trade Commission to protect Ohio Steelworkers. Ohio has the best steelworkers in the country and they deserve to compete on a level playing field with our competitors. That is why I am working to promote exports of Ohio steel and protect our steelworkers from unfair competition: [Link Omitted]
Praising Active Military/ Military Units	(1) 150 soldiers from the 1st Stryker Brigade Combat Team, 25th Infantry Division arrived home to Fort Wainwright yesterday from a deployment in Afghanistan. Welcome home! Check out these great homecoming pictures: (2) The South Dakota Army National Guard 196th Maneuver Enhancement Brigade returned after a 10-month deployment to Kuwait. I was honored to welcome them home today. (3) Welcome support from Hollywood to #SaveSaeed Abedini and Kenneth Bae. (4) Welcome home from Afghanistan SPC Jessica McKim, The U.S. Army! (5) Yesterday, the 82nd Airborne Division welcomed its 49th commander, Maj. Gen. Erik Kurilla. Join me in thanking Maj. Gen. Kurilla and the 48th commander, Maj. Gen. Richard Clarke for their service to the 82nd Airborne Division.
Terrorism	(1) Whether its in Jerusalem, or Paris, or New York, the civilized people of the world are under siege by violent Islamic extremists. We must redouble our efforts to win this war on terrorism. In 2016, we must remain vigilant against Islamic extremism. We need to do everything in our power to protect Americans from terrorism. (2) Washington must wake up to this fact: the crowd-sourcing of domestic terror is a reality. While slow-moving, federal bureaucracies look for card-carrying terrorists, the Islamic State and al Qaeda are crowd-sourcing their jihad. We must adapt and develop a long-term strategy to name and defeat our enemy. That enemy is not the empty label “extremism” but the ideology of militant Islam.

	(3) This weekend as Americans celebrate the birth of our nation and our freedoms, the horrific events yesterday are a sobering reminder of the radical Islamic terrorists who seek to destroy us. Twenty more innocent civilians are dead in the most recent attack by radical Islamic terrorism this time the target was an upscale neighborhood in Dhaka, #Bangladesh. Terrorists throwing grenades and shouting Allahu Akbar attacked a popular restaurant frequented by diplomats and students, reportedly slitting the throats of some of their victims and singling out those who could not recite verses from the Koran. Over the past few years, the authorities in Bangladesh have insisted that terrorist incidents in their country were somehow isolated from the global scourge of radical Islamic terrorism, the product of homegrown militants responding to domestic grievances. This attack, however, appears designed by ISIS to prove otherwise, and that the Islamic State is functioning not only in Syria and Iraq but also in Bangladesh, just as it is in Turkey, France, Belgium and the United States. This isnt just a wake-up call anymore it is a screaming siren of an alarm that demands our urgent attention. Just as the Bangladesh attack was a demonstration of the Islamic States determination to take their war on civilization to another Muslim-majority nation, the recent attacks in the United States from San Bernardino to Orlando demonstrate the United States is, likewise, not immune. We are fortunate to have friends and allies in Europe, the Middle East, and Asia who see this enemy for what it is, and we should do everything we can to partner with them not to struggle against generic violent extremism but to actually take the fight to the radical Islamic terrorists. (4) The attack at the Istanbul Ataturk Airport in Turkey is yet another tragic reminder of the threats we face from terrorists who indiscriminately employ senseless violence with no regard for innocent lives or their own. It's a reminder that we must remain vigilantnot only here at home, but also around the globein our efforts to prevent terrorist threats at our transportation hubs and other vulnerable locations. (5) While this is a milestone that we have all awaited, we must remember that al Qaeda and its affiliates are not dependent on one man and we must remain vigilant in our efforts to disrupt and destroy terrorist networks that threaten our Nation and allies.
Military Sexual Assault	(1) Momentum continues to build in support of the Military Justice Improvement Act, which would create an unbiased military justice system by transferring the decision-making authority over which cases go to trial from the chain of command to independent trained military prosecutors where it belongs. I hope you'll join me and my colleagues in supporting this commonsense measure to reform the way the military handles sexual assault cases so that victims can get the fair shot at justice they deserve. (2) Military sexual assault victims were silenced in Congress. Lend your voice in support of the Military Justice Improvement Act to give them a fair shot at justice. #MJIA

	(3) Yesterday, I stood alongside sexual assault survivors, advocates, and a bipartisan coalition of U.S. Senators to renew our push for real reform of the military justice system. Our bipartisan Military Justice Improvement Act would remove prosecutorial authority in sexual assault cases from the chain of command and place it in the hands of independent trained military prosecutors where it belongs. Only then will we truly create an unbiased system of military justice where assailants are held accountable and survivors, like Samantha Jackson, get the justice they deserve. (4) Of the estimated 26,000 incidents of unwanted sexual contact and sexual assaults that occured in 2012, only 3,374 were reported. Time and again, victims tell us the reason there is such a drop off in reporting is because the decision-making in cases of sexual assault lies within the chain of command. Until this authority is moved outside the chain of command, into the hands of an independent military prosecutor, victims will not have the confidence they will receive the justice they deserve. This is why I'm fighting to pass the Military Justice Improvement Act, to make sure the voices of victims–like Navy veteran and MST survivor Brian Lewis–are heard and that we create an objective and accountable military justice system. Please share this graphic to let Brian's voice, and that of so many others, be heard. #MJIA (5) Victims of sexual assault in the military tell us time and time again that the reason they don't report sexual assaults is because the decisionmaking lies within their own chain of command. If we want to truly reform the military justice system and give victims of sexual assault a real chance at justice, we must remove this decisionmaking power from the chain of command and place it in the hands of independent trained military prosecutors. Right now, victims' voices are not being heard in Congress, such as U.S. Marine Vet & #MST survivor Stacey Thompson. Please lend your voice in support of the Military Justice Improvement Act and make sure the voices of victims like Stacey are heard: [Link Omitted] #MJIA
Nuclear Deterrance/ International Security	(1) Today, Iran test-fired ballistic missiles in violation of international sanctions. This launch demonstrates the regimes continued disregard for international sanctions and the grave consequences of this Administration blindly trusting the Iranian regime to uphold its commitments. Instead of making concessions to the Iranian regime and freeing up more than $100 billion dollars in sanctions relief, the Obama Administration should change course and increase pressure on Iran by imposing additional unilateral sanctions to stop their belligerent behavior once and for all. (2) The more we learn about the Iran deal, the more concerned I am that it not only fails to prevent Iran from obtaining a nuclear weapon, it emboldens them through tens of billions of dollars in sanctions relief, a phased out lifting of a UN arms embargo and the ability to test more advanced centrifuges.

	(3) The long list of concessions in the Presidents Iran deal includes the end of a long-standing arms embargo. This misguided concession, coupled with the immediate and imprudent sanctions relief, makes this not only a weak deal, but a dangerous one, too. (4) Even President Obama is admitting that Iran would be able to develop a nuclear weapon in "a matter of months" after the Iran deal expires. I strongly oppose the deal, which paves the path for Iran to become a nuclear power and threatens the security of Israel and our partners in the Middle East. (5) Freeing American prisoners of conscience in Iranian prisons should be a pre-condition to any negotiations with Iran. #FreeAmir
Air Force	(1) Today, Brig. Udo K. "Karl" McGregor, commander of the 452nd Air Mobility Wing, March ARB, Calif., passed the 730th Air Mobility Training Squadron guidon to Lt. Col. Jonathan M. Philebaum, 730th AMS commander, Altus AFB, during the squadron reactivation and assumption of command at hangar 517. (Photo courtesy of U.S. Air Force) (2) It was an honor to spend Friday at Pease Air National Guard Base, onboard the KC-135A Stratotanker with the New Hampshire National Guard. During the refueling mission, I watched as the crew honed in on a EC-130J Commando Solo aircraft not far from Pittsburgh and, with laser like precision, guided a refueling hose into the aircraft. (3) Advocated for Warriors of the North-Grand Forks Air Force Base, Minot Air Force Base, and North Dakota National Guard during meeting with U.S.Airforce. (4) U.S. Sen. Cory Booker and the Adjutant General of the New Jersey National Guard, Brig. Gen. Michael L. Cunniff, discuss the flight parameters of an F-16 fighter jet during a visit to the 177th Fighter Wing in Egg Harbor Township, N.J. on Aug. 5, 2016. It was the first visit to the New Jersey Air National Guard base by Sen. Booker, who received a briefing by the Wing Commander, U.S. Air Force Col. John R. DiDonna, flew an F-16 simulator and got a chance to get an up close look at an F-16C Fighting Falcon fighter jet on the flight line at the Atlantic City Air National Guard base. (U.S. Air National Guard photo by Master Sgt. Andrew J. Moseley/Released) (5) Enjoyed meeting with Col. Gentry Boswell, Commander of the 28th Bomb Wing at Ellsworth Air Force Base and Command Chief, CMSgt Sonia Lee today to discuss the status of the B-1 fleet at Ellsworth.
Honoring Specific Veterans	(1) What an honor to present the Distinguished Flying Cross to WWII Veteran and Hudson-area resident Alfred LeFeber. A long overdue recognition for a true patriot. (2) Yesterday, I was honored to present Roger Piasecki of Kearney with a WWII Medal & Lapel Pin for his father Walter's military service. #Greatest-Generation

	(3) Today I participated in a ceremony honoring Mildred Pretzer of Charlottesville, a World War II veteran who served as part of the Womens Army Corps (WAC). I proudly presented Ms. Pretzer with the Army Good Conduct Medal, the WAC Service Medal, the American Campaign Medal, the World War II Victory Medal and the Honorable Service Lapel Pin. (4) Today, I had the privilege of escorting James Gray, a WWII Veteran from Malvern as he visited the WWII Memorial for the first time. Mr. Gray is a former POW and Purple Heart Recipient who fought in the Battle of the Bulge where he was captured and spent six months as a German POW. Thank you for your service to our great country, Mr. Gray! (5) It was an honor to present lost medals to World War II Veteran and hero Julious Elmore in Magnolia today.
Honoring Veterans/ Heroes	(1) Honor the heroes who ran into burning buildings. Honor the brave who sacrificed their lives for strangers. And never forget those lost on 9/11/2001 and 9/11/2012. (2) Today we remember and honor the brave men and women that sacrificed their lives for our freedom. We are forever grateful and you will never be forgotten. (3) Thank you to the men and women who have worn, and continue to wear, the uniform of our nation's military. As America pauses today to remember the service and sacrifice that our veterans and active duty members have endured, we are forever in debt to you and your families for your selfless service. (4) Today, let us pause to remember the heroes we have lost in the service of our great nation. Thank you to all who serve to protect our freedom. (5) This #MemorialDay, we honor the brave men and women who paid the ultimate sacrifice to preserve our nation's freedom
Military Operations/ Armed Conflicts	(1) Sen. Corker visited the Kilis refugee camp on the Turkish border with Syria today for the second time since the conflict began. Members of the camps refugee leadership expressed dismay and disappointment at the lack of U.S. support for the Syrian opposition. Specifically, refugee leaders noted the importance of helping arm and equip the vetted, moderate opposition under the leadership of Gen. Salim Idriss. (2) Congressional authorization for the use of military force against ISIS, Al-Qaeda, and the Taliban will make clear to our warfighters, our allies, and our adversaries that we are united #AUMF (3) U.S. Senator Tim Kaine leads a Congressional Delegation (CODEL) to the Middle East region focused on the U.S. mission against ISIL and the ongoing humanitarian crisis in Iraq and Syria. (4) The drawdown of US troops in Iraq and use of force in Libya without a plan have enabled ISIS to amass power and territory with little pushback.

	(5) ISIS has possession of and is now using illegal chemical weapons against the Kurds in Iraq. Earlier this year, I visited Iraq and met with the leaders of Iraqi Kurdistan, and I firmly believe that the U.S. must increase its support for the Iraqi Kurdish Peshmerga forces who are valiantly fighting against ISIS.
Veterans Affairs/ Veterans Healthcare	(1) I have been and will continue to be committed to reforming the VA to ensure our veterans receive the care they deserve. For our veterans in need of help with the VA, my office is here to help! (2) Senator Grassley has made it clear that the comments from the Secretary of Veterans Affairs comparing wait times for VA treatment with wait times for rides at Disneyland were unacceptable and that the Secretary should make amends. Senator Grassley has worked to improve veterans experience at the VA, including pushing for and receiving an acknowledgement from the VA that veterans have experienced problems with the Veterans Choice program and a pledge to improve services to Iowa veterans and veterans across the country. (3) "Our veterans deserve the best medical care available, but far too often, veterans suffering from PTSD, Traumatic Brain Injury, and other psychological impacts slip through the cracks." (4) The Department of Veterans Affairs needs to fix problems in helping veterans get appointments outside the VA when needed. Senator Grassley is working with the VA on veterans' frustrations. (5) Today, I toured the Portland Vet Center with Steven Reeves from the U.S. Department of Veterans Affairs. Our veterans deserve the highest quality care possible, and our 5 Vet Centers in Maine need to be adequately staffed in order to provide vital services to those who served our country.

References

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*. pp. 288–296.