

# 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

July 11, 2019

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 is 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

<b>1</b>	<b>Alternative Task Designs and Results</b>	<b>SI: 1</b>
1.1	A Variant for R4WSI . . . . .	SI: 1
1.2	Two Alternative Label Validation Tasks . . . . .	SI: 1
<b>2</b>	<b>Working Example: Posting Tasks</b>	<b>SI: 5</b>
<b>3</b>	<b>Workers’ Training and Performance</b>	<b>SI: 8</b>
3.1	An Example of the Training Module . . . . .	SI: 8
3.2	Workers’ Performance . . . . .	SI: 17
<b>4</b>	<b>More on the Corpus and Model Fit</b>	<b>SI: 18</b>

# 1 Alternative Task Designs and Results

In this section, we present a variant 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 A Variant for R4WSI

The different version of R4WSI is identical to the task presented in the main text with one exception: the documents shown are always one of the ten most representative documents for a topic. Thus the only difference between these task structures is how the document is selected. For convenience, we call it R4WSI-Representative.

Same as other tasks, we post 500 HITs twice, summing to 1000 HITs. The result can be seen in Figure SI1. Models 1 and 2 are not distinguishable from each other, nor does Model 2 and 3. 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 only draw from the representative documents for R4WSI.

## 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 WSI described in the topic validation section in the main text, simply replacing the document with 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 the for the relevant topic and the others are for other economy-related topics. An example is shown in Panel (b) of Figure 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.

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.

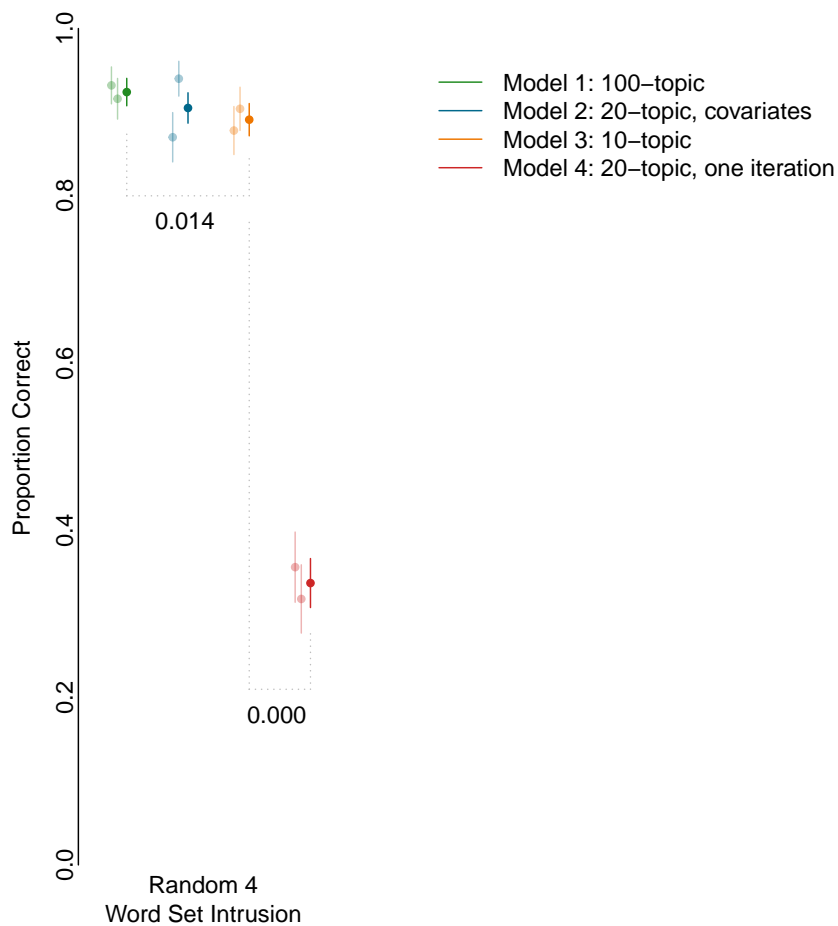
To test these task structures we followed the same basic procedures discussed above. For each task/coder combination we created 500 tasks that were coded by trained workers on AMT.<sup>1</sup> The results from 4,000 high quality HITs are shown in Figure SI3.

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

---

<sup>1</sup>Workers were paid 0.02/HIT for the OLW task, 0.03/HIT for the WSI task

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. The two identical trials for Model 2 are significantly different.)

Figure SI2: Example HITs for Alternative Label Validation Tasks

**Resources for Local Communities**

Please read the four word sets below and click on the word set that is most UNRELATED to the label.

- pushed, make, tools, ensure
- store, started, businesses, like
- communities, tools, continue, make
- possible, reach, needs, pushed

Submit

(a) Word Set Intrusion

today's, times, percent, press, wall, street, unemployment, says

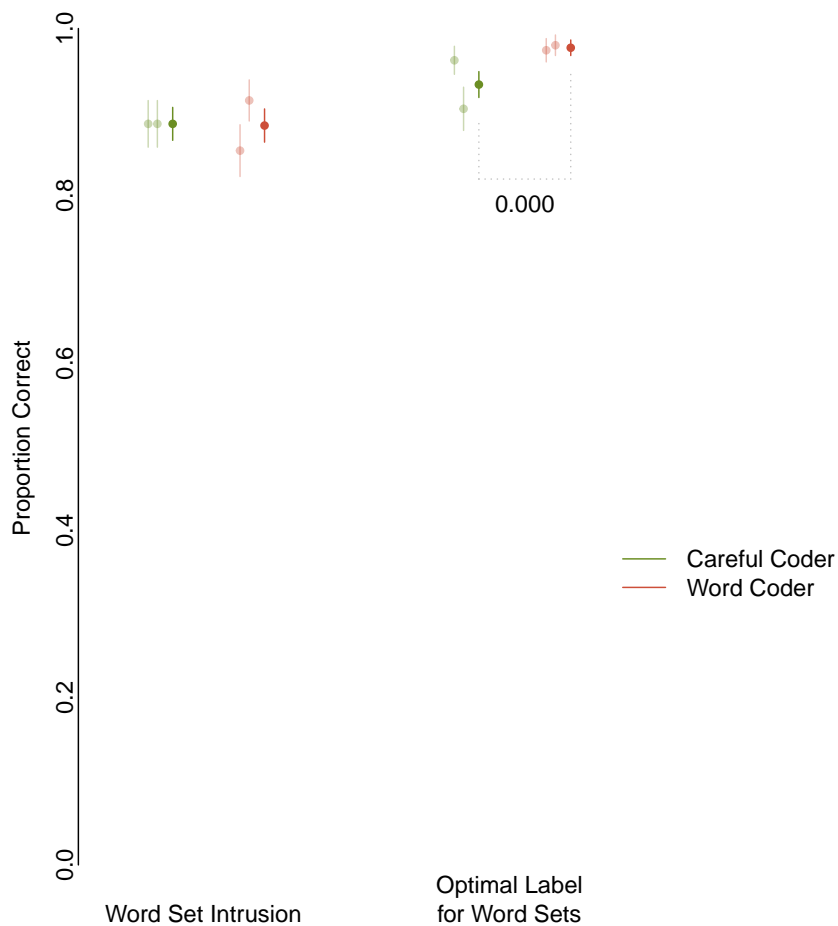
After reading the above word set, please click on the label below that BEST summarizes these words.

- Resources for Local Communities
- Income Inequality
- Raising the Debt Ceiling
- Jobs & Keystone Pipeline

Submit

(b) Optimal Label for Word Sets

Figure SI3: 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.)

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 some form of OLD.

## 2 Working Example: Posting Tasks

In this section, we provide example code to post validation tasks. We are building an R package with the following mentioned functions, with which any future researchers will be able to post their own validation tasks via the `SentimentIt` platform.<sup>2</sup>

First and foremost, the researcher needs to register with `SentimentIt`. Go to [www.sentimentit.com/](http://www.sentimentit.com/) and follow the links or go to [https://www.sentimentit.com/researchers/sign\\_up](https://www.sentimentit.com/researchers/sign_up). The following page includes instructions: [https://www.sentimentit.com/documentation/getting\\_started](https://www.sentimentit.com/documentation/getting_started). The email address and password (“researcher@school.edu” and “uniquePassword”) for this account will be required in the following steps.

Next, we show the entire procedure from making the desired tasks with topic model outputs to interacting with the workers on the Mechanical Turk platform. The following example code is for the *Label Intrusion* (LI) task in the paper, but it can be easily altered to post other tasks by changing an argument in the `MakeTaskLabel/MakeTaskTopic` and `SendTaskLabel/SendTaskTopic` functions, which will be discussed in more detail below.

Before considering our crowd-sourced topic validation methods, researchers need to fit their own topic models using any software of their choice, such as the R packages `lda` and `stm`. The outputs of different software would be of different structures, and our package `SentimentIt` mainly needs a) the high probability (or frequent and exclusive) words for each topic and b) the documents with indicators of the top three topics. Both a) and b) could be in either matrix or data frame format. For label validations, labels are also required from human coders — preferably external coders who are not the authors — after they carefully reading both the topic words and the representative documents.

First of all, we need to prepare validation tasks locally. For the *Label Intrusion* (LI) task, the code is as follows. For other label validation task structures, simply specify the `structure` argument as “OLD”, “WSI”, etc. For topic validation task, use the function `MakeTaskTopic()`.

```
TasksLI <- MakeTaskLabel(email = 'researcher@school.edu',
                        password = 'uniquePassword',
                        n = 500,
                        structure = 'LI',
                        text = document[, "text"],
                        toptopics = document[, "Top3", "Top2", "Top1"],
                        labelsource = label_from_RA1)
```

The `n` argument specifies the number of tasks and `structure` the type of these tasks. The `text` argument looks for a column (of a matrix or a data frame) that contains all the

---

<sup>2</sup>The R package for interacting with `SentimentIt` is hosted on GitHub at: <https://github.com/carlson9/SentimentIt/>. Additional information on setting up certifications for tasks is available here: <https://www.sentimentit.com/documentation/certifications.markdown>.

original documents. The `toptopics` argument specifies several indicators (in the example code, three) of the machine predicted top topics, in correspondence with each document. Its values should match the column names of the label file, as is required by the `labelsource` argument.

The output of `MakeTasksLabel()` is a list of three elements. In the above example, it is assigned to the object “TasksLI.” The first element, “Tasks[[1]],” is the randomly selected texts that are going to be presented. The second element, “Tasks[[2]],” is the options for the Mturk workers to choose. The column name (the last column) will indicate which words are the intruders for future reference. When the task is later posted, the order of the words will be randomized. The third element, “Tasks[[3]],” is of identical structure to the second one, except the values are IDs returned from the `SentimentIt` server. When the function `MakeTaskLabel` is called, it posts the potential options on the server and gets a unique ID for each of them, using the authentication specified by `email` and `password`. The platform will later “recognize” these ID, recall the corresponding options, and present them to the Mturk workers.

Next, researchers will post the tasks to the `SentimentIt` server. To achieve this, create a batch first. Keep record of the returned batch id. `task_setting_id` is the HIT setting identifier. To specific a HIT setting, go to [https://www.sentimentit.com/hit\\_settings](https://www.sentimentit.com/hit_settings).<sup>3</sup>

```
batch_id <- createBatches(email = 'researcher@school.edu',
                          password = 'uniquePassword',
                          task_setting_id = 44,
                          num_batches = 1)
```

Then, send the prepared tasks to the above empty batch. The call for the label validation tasks is as follows. For topic validations, use `SendTaskTopic()`.

```
SendTaskLabel(email='researcher@school.edu',
              password = 'uniquePassword',
              structure = 'LI',
              batch_id = batch_id,
              text = TasksLI[[1]],
              option_ids = TasksLI[[3]],
              question = "After reading the above passage, please click on
the label below that is most UNRELATED to the passage.")
```

Note that the type of these tasks needs to be specified again with the `structure` argument. The `batch_id` argument asks for the desired batch id to which the tasks will be sent. The `text` argument refers to the text presented to the Mturk workers, which, in our example, is the original documents. The `option_ids` argument takes in the IDs — as opposed to the actual words — for each tasks. The IDs are in the three elements returned by the `MakeTaskLabel()` or `MakeTaskTopic()` function. For the LI task, the question is default to “After reading the above passage, please click on the label below that is most

---

<sup>3</sup>Additional information on this is available at: [https://www.sentimentit.com/documentation/running\\_analysis#step3](https://www.sentimentit.com/documentation/running_analysis#step3)

UNRELATED to the passage.” But it can be specified differently when needed through the `question` argument.

Now, all the tasks should be visible on the `SentimentIt` platform. Researchers can log into <https://www.sentimentit.com/> and preview the tasks under the “Data – Comparisons” page. These tasks would look very similar to the ones Mturk workers are going to see.

After checking everything, we can send the tasks to the Mechanical Turk platform by running the following command. The batch is no longer empty but contains  $n$  tasks.

```
createTasks(email = 'researcher@school.edu',
            password = 'uniquePassword',
            batch_id = batch_id)
```

To check the batch status, run:

```
batchStatus(email = 'researcher@school.edu',
            password = 'uniquePassword',
            batch_id = batch_id)
```

If tasks expire before all are completed, run the following to repost:

```
repostExpired(email = 'researcher@school.edu',
              password = 'uniquePassword',
              batch_id = batch_id)
```

When all HITs in the batch have been completed, read in the data. The `write_document_to` argument writes the results to a local file for future reference, as the `SentimentIt` platform will delete the data after six months.

```
readInData(email = 'researcher@school.edu',
            password = 'uniquePassword',
            batch_id = batch_id,
            write_document_to = "batch_id_result.csv")
```

To check the proportion correct for the batch. Simply find the original tasks and compare it with the workers’ answers using the following command.

```
CheckCorrect(org.task = TasksLI,
             answer = "batch_id_result.csv")
```

To check the workers’ performance, researchers need to post the **exact** same tasks and compare the results using the following command. If any tasks are identified as being of low quality, re-pack them in a different batch using the same procedure as above.

```
CheckAgreement(Round1 = "batch_id_result.csv",
               Round2 = "batch_id2_result.csv")
```



## 3 Workers' Training and Performance

The first part of this section is an example of a full training module for the random 4 word set intrusion task used in the paper. The second part provides some statistics and discussion about the workers' performance.

### 3.1 An Example of the Training Module

If you finish this training module with a passing score, you will be qualified to complete HITs posted by the requester `SentimentIt` with the title **Word Set Intrusion Task**.

#### 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 passage and 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).

This training module has two parts. In Part 1, we will provide five practice HITs to help you better understand the task. In Part 2, we will give you eight example HITs to complete. To receive the qualification for the Word Set Intrusion task, you must complete seven out of eight of these example HITs correctly.

Before proceeding, please enter your MTurk Worker ID (not the email login ID) below in order to record your qualification information.

MTurk Worker ID:

Confirm MTurk Worker ID:

Click the next button. It will be disabled if you have already taken the survey or have been banned.

#### Part 1

This part includes five practice questions to help you better understand the nature of the task. Your answers will NOT be scored.

#### Practice HIT #1

Please read the following passage:

Alexander Cosponsors 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 word sets below and click on the set of words that is most unrelated to the passage.

- A. bill, govern, vote, cut
- B. senate, congress, pass, plan
- C. tax, budget, house, reform
- D. report, today, press, emergency

**Practice HIT #1** - Answer [Correct/Incorrect].

The passage talks about a legislation on the federal budget cut. 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 "report, today, press, emergency" is not. Rather it seems to relate to time sensitive news media reports and emergencies.

[Original Document Pasted Here]

**Practice HIT #2**

Please read the following passage:

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

Please read the four word sets below and click on the set of words that is most unrelated to the passage.

- A. great, thank, visit, city
- B. state, enjoy, tour, center
- C. reform, obamacare, federal, year
- D. happy, team, celebrate, park

**Practice HIT #2** - Answer [Correct/Incorrect].

The passage celebrates a town called "Warroad" for having many hockey medalists. 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.

[Original Document Pasted Here]

### **Practice HIT #3**

Please read the following passage:

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 word sets below and click on the set of words that is most unrelated to the passage.

- A. nation, student, service, school
- B. honor, education, veteran, state
- C. social, call, secure, post
- D. learn, high, college, proud

### **Practice HIT #3 - Answer [Correct/Incorrect].**

The passage congratulates a student for receiving the U.S. Presidential Scholar Award. Two of the word sets (”nation, student, service, school” and ”learn, high, college, proud”) clearly relate to recognizing local figures and educational issues. The other word set (”honor, education, veteran, state”) is slightly more ambiguous but is still more closely related to recognizing local educational achievements than the remaining word set (”social, call, secure, post”).

[Original Document Pasted Here]

### **Practice HIT #4**

Please read the following passage:

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 word sets below and click on the set of words that is most unrelated to the passage.

- A. day, family, country, people
- B. proud, learn, university, service
- C. women, american, live, make
- D. work, life, nation, love

**Practice HIT #4** - Answer [Correct/Incorrect].

The passage honors victims of the attacks on Pearl Harbor. While the three other word sets are related to patriotism, celebrations of American ideals and values, and important civic holidays ("day, family, country, people", "women, american, live, make", and "work, life, nation, love"), the set of words "proud, learn, university, service" is instead more closely related to celebrating the achievements of local citizens in areas such as education and service.

[Original Document Pasted Here]

**Practice HIT #5**

Please read the following passage:

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 word sets below and click on the set of words that is most unrelated to the passage.

- A. job, business, help, work
- B. confirm, judge, nominate, rule
- C. energy, economy, state, new
- D. develop, continue, import, protect

**Practice HIT #5** - Answer [Correct/Incorrect].

The passage advertises that North Dakota received federal funds supporting small businesses and local economies. While the other three word sets are related to the job, business, help, work confirm, judge, nominate, rule energy, economy, state, new develop, continue, import, protect economy, local economic development, and business concern ("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.

[Original Document Pasted Here]

Now we are moving on to the second part. It includes eight questions to test your qualification to take the task. Thus, your answers will be scored. To receive the qualification for the Word Intrusion Task, you must answer correctly to at least seven out of eight following questions.

## Part 2

In this part, your answers will be scored. To receive the qualification for the Word Intrusion Task, you must answer correctly to at least **seven** out of **eight** following questions.

### Test HIT #1

Please read the following passage:

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.

Please read the four word sets below and click on the set of words that is most unrelated to the passage.

- A. discuss, join, week, senate
- B. hear, talk, share, live
- C. many, world, nation, america
- D. office, morning, watch, meet

### Test HIT #1 - Answer [Correct/Incorrect].

The passage announces the schedule of a town hall meeting. While the other three word sets are related to activities that legislators do in their home districts/states ("discuss, join, week, senate", "hear, talk, share, live" and "office, morning, watch, meet"), the set of words "many, world, nation, america" is rather related to America as a nation on the world stage.

[Original Document Pasted Here]

### Test HIT #2

Please read the following passage:

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: <http://1.usa.gov/1KoWSIL>

Please read the four word sets below and click on the set of words that is most unrelated to the passage.

- A. thank, visit, good, time
- B. health, care, veteran, help
- C. need, provide, support, work
- D. drug, program, prevent, ensure

**Test HIT #2** - Answer [Correct/Incorrect].

The passage talks about a legislation that prevents veterans' suicide and underscores the importance of taking care of veterans' mental health. While the other three word sets are related to providing health care support and programs ("health, care, veteran, help", "need, provide, support, work" and "drug, program, prevent, ensure"), the set of words "thank, visit, good, time" is rather related to legislators' visits to local events and appreciation of local constituents' achievements.

[Original Document Pasted Here]

**Test HIT #3**

Please read the following passage:

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 word sets below and click on the set of words that is most unrelated to the passage.

- A. right, law, court, senate
- B. vote, justice, supreme, constitution
- C. judge, nominate, rule, investigate
- D. news, report, read, said

**Test HIT #3** - Answer [Correct/Incorrect].

The passage endorses the nomination of the new U.S. Supreme Court judge. While the other three word sets are related to legal institutions and offices ("right, law, court, senate", "vote, justice, supreme, constitution" and "judge, nominate, rule, investigate"), the set of words "news, report, read, said" is rather related to news reports of ongoing events.

[Original Document Pasted Here]

**Test HIT #4**

Please read the following passage:

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 word sets below and click on the set of words that is most unrelated to the passage.

- A. president, action, security, administration
- B. elect, decision, political, confirm
- C. nation, threat, trump, deal
- D. secretary, congress, immigrant, defense

**Test HIT #4** - Answer [Correct/Incorrect].

The passage supports more sanctions to be placed on Iran for its on-going development of the nuclear missile program. While the other three word sets are related to national security and those who have a say on security decisions ("president, action, security, administration", "nation, threat, trump, deal" and "secretary, congress, immigrant, defense"), the set of words "elect, decision, political, confirm" is rather related to electing political figures and confirming decisions nominating such figures.

[Original Document Pasted Here]

**Test HIT #5**

Please read the following passage:

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: <https://www.disasterassistance.gov/>

Please read the four word sets below and click on the set of words that is most unrelated to the passage.

- A. family, legislation, children, bill
- B. news, today, said, report
- C. governor, state, press, call
- D. federal, emergency, post, continue

**Test HIT #5** - Answer [Correct/Incorrect].

The passage talks about the application for federal disaster assistance. While the other three word sets are related to news reports and emergency announcements ("news, today, said, report", "governor, state, press, call" and "federal, emergency, post, continue"), the set of words "family, legislation, children, bill" is rather related to legislation on family issues.

[Original Document Pasted Here]

**Test HIT #6**

Please read the following passage:

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 word sets below and click on the set of words that is most unrelated to the passage.

- A. west, proud, community, learn
- B. university, force, military, nation
- C. student, service, member, thank
- D. veteran, help, need, support

**Test HIT #6** - Answer [Correct/Incorrect].

The passage calls for applications for military academies. While the other three word sets are related to education ("west, proud, community, learn", "university, force, military, nation" and "student, service, member, thank"), the set of words "veteran, help, need, support" is not but rather relates to supporting veterans.

[Original Document Pasted Here]

**Test HIT #7**



Please read the following passage:

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. <http://1.usa.gov/1nr1get>

Please read the four word sets below and click on the set of words that is most unrelated to the passage.

- A. administration, deal, threat, state
- B. small, support, community, economy
- C. create, continue, import, protect
- D. job, business, help, work

**Test HIT #7** - Answer [Correct/Incorrect].

The passage talks about the development of local businesses in North Dakota. While the other three word sets are related to the economic activities and small local businesses ("small, support, community, economy", "create, continue, import, protect" and "job, business, help, work"), the set of words "administration, deal, threat, state" is not but rather relates to national security.

[Original Document Pasted Here]

**Test HIT #8**

Please read the following passage:

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 word sets below and click on the set of words that is most unrelated to the passage.

- A. justice, elect, supreme, political
- B. decision, protect, general, constitution
- C. right, law, court, senate

D. bill, american, tax, people

**Test HIT #8** - Answer [Correct/Incorrect].

The passage denounces Attorney General for his misleading testimonies at the confirmation hearing. While the other three word sets are related to legal institutions and offices (justice, elect, supreme, political”, ”decision, protect, general, constitution” and ”right, law, court, senate”), the set of words ”bill, american, tax, people” is rather related to US tax bills.

[Original Document Pasted Here]

[Pass block] You have successfully completed the training module with a score of [automatically caculated score]. You will now be eligible to participate in Word Intrusion HITs. Please be sure to continue carefully reading each future Word Intrusion HIT. We will be monitoring your responses to ensure response quality. If we determine that your answers do not accurately assess the statement and the sets of words in question, you will no longer be eligible to work on Word Intrusion HITs. Thank you for completing this training. Your turk ID is now included for this task. We are currently Beta testing our qualification system. If you have any issues, please contact [researcher’s email] by email or through AMT.

[Fail Block] Thank you for completing the training module. Unfortunately your score of [automatically caculated score] prevents you from participating in the Word Set Intrusion HITs. By clicking here, I acknowledge that my performance is being monitored as indicated above.

### 3.2 Workers’ Performance

To assess the consistency of workers’ performance across rounds, we present the “agreement rate” in the following two tables. 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 SI1: Workers’ Agreement Rate in the Topic Validation Tasks

	Model 1	Model 2	Model 3	Model 4
<b>Tasks in Paper</b>				
WI	0.722	0.746	0.742	0.790
T4WSI	0.710	0.624	0.562	0.622
R4WSI-Rdm	0.854	0.836	0.810	0.630
<b>Tasks in SI</b>				
R4WSI-Rep	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 SI2: Workers' Agreement Rate in the Label Validation Tasks

	Careful Coder	Word Coder
<b>Tasks in Paper</b>		
LI	0.684	0.688
OL-Econ	0.826	0.758
OL-All	0.890	0.634
<b>Tasks in SI</b>		
WSI	0.860	0.856
OLW	0.908	0.962

## 4 More on the Corpus and Model Fit

In this section, we provide more details on the corpus and model fit. 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 three models (Model1, Model2, Model3) are of different quality.

Figure SI4: Word Cloud: Six topics from **model 1** (the 100-topic model)



Figure SI5: Word Cloud: Six topics from **model 2** (the 20-topic model, covariates)



Figure SI6: Word Cloud: Six topics from **model 3** (the 10-topic model)

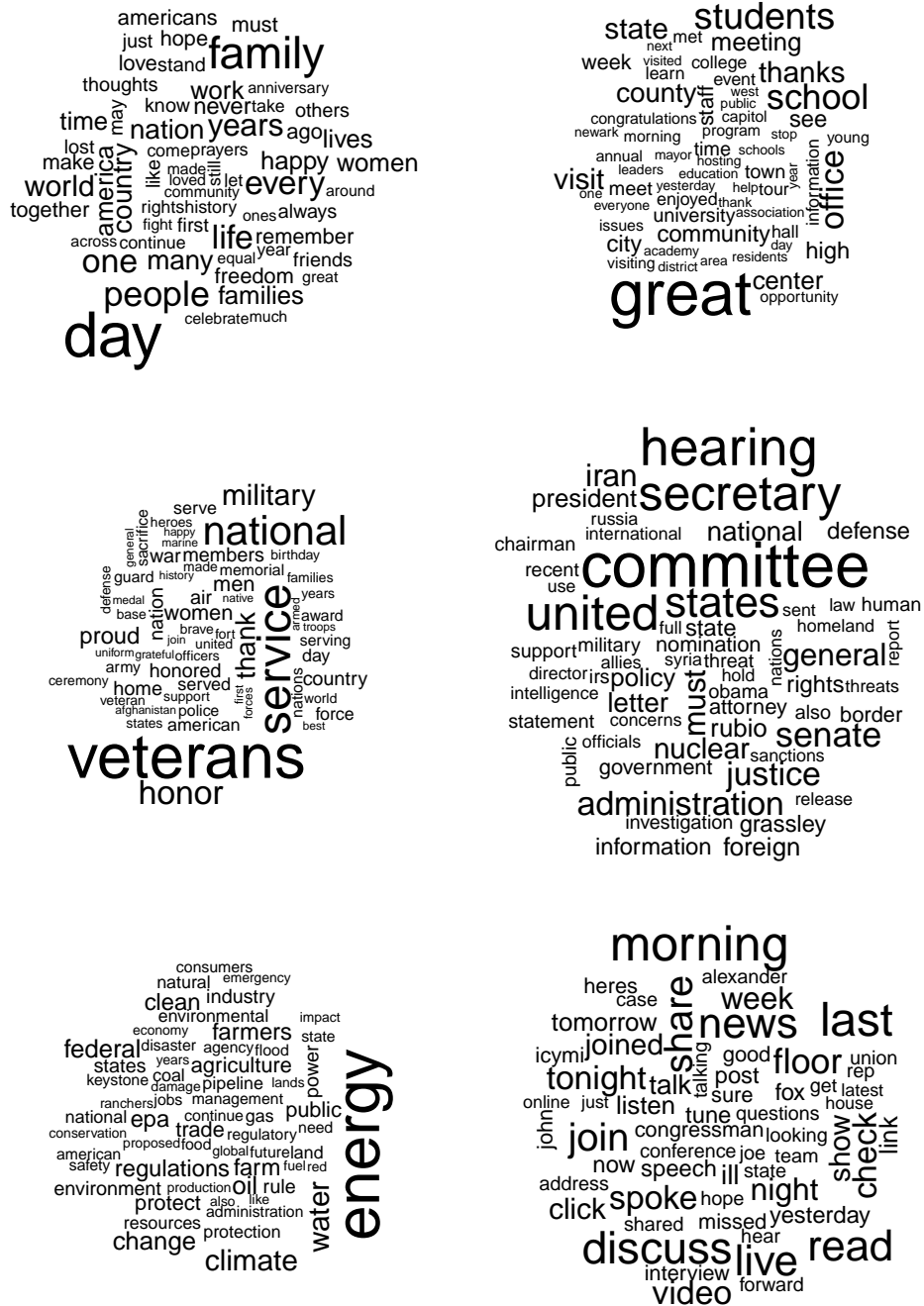
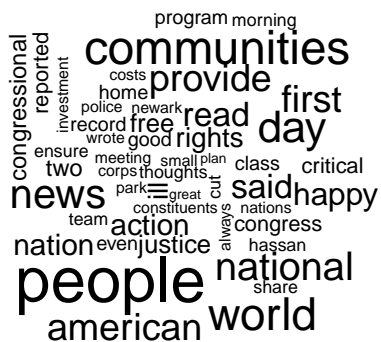
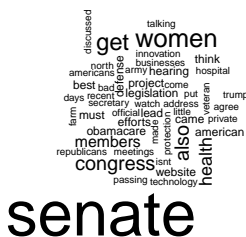
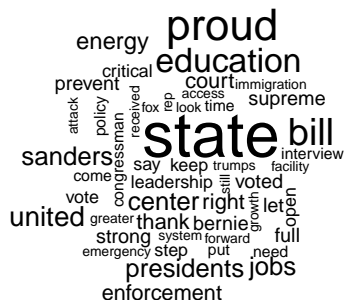


Figure SI7: Word Cloud: Six topics from **model 4** (the 20-topic model, one iteration)



For the label validation tasks, we asked the two careful coders to read carefully read the representative document. **Table SI3** shows the top five documents for each economic topic with which they come up with the labels.

Table SI3: Representative Documents

Label (Optimal)	Document
Raising the Debt Ceiling	<p>(1) What is most imppt part of debt ceiling deal to POTUS? Getting past his reelect? Raising taxes? Blame? How about avoiding a downgrade for our debt by dealing with spending?</p> <p>(2) POTUS regrets vote against raising the debt limit as Senator</p> <p>(3) Voted NO on raising debt ceiling without addressing spending problem. Need to both prevent a debt limit crisis today &amp; debt crisis tomorrow.</p> <p>(4) Rather than gamesmanship over shutdowns, how about solving the problems of spending and debt? That is what the people want.</p> <p>(5) An increase in the debt limit should be accompanied by structural reforms that deal with excess spending. This measure did not.</p>
Income equality	<p>(1) Rich Get Richer, Poor Get Poorer: A study by economist Emmanuel Saez shows that from 2002 to 2012 the income for the bottom 90 percent of income-earners decreased by 10.7 percent, while the income for those in the top 0.01 percent increased by 76.2 percent. Read the study here: [Link Omitted]</p> <p>(2) “The 7.6% unemployment figure so common in headlines these days is utterly misleading. An estimated 22 million Americans are unemployed or underemployed; they are virtually invisible and mostly excluded from unemployment calculations that garner headlines.” [Link Omitted]</p> <p>(3) Middle Class Collapse: The middle class is receiving less of America’s total income as median wages stagnate and wealth concentrates at the top, according to a new study by the Pew Research Center released on Wednesday. Since 2000, the middle class has shrunk in size and fallen backward in income and wealth. Pew’s key findings: -In 2011, 51 percent of adult Americans were in the middle class, down from 61 percent in 1971. -From 2001-2010, the median income of the middle-class fell 5 percent, but median wealth declined by 28 percent, to \$93,150 from \$129,582. -The median wealth of upper-income tier Americans was essentially unchanged-it rose by 1 percent to \$574,788 from \$569,905. Meantime, the wealth of lower-class Americans plunged by 45 percent, to \$10,151 from \$18,421. Continue reading here: [Link Omitted]</p>



(4) The Rich Get Richer: Between 2009 and 2010, 93 percent of all new income went to the top 1 percent, according to a study published by University of California economist Emmanuel Saez. The income chasm in America is growing wider. In the recovery after the early 1990s downturn, the wealthiest 1 percent captured 45 percent of the nation's income growth. After the dot-com collapse a decade ago, Saez said, 65 percent of the income growth went to the top 1 percent. The income gap is so broad today that it "shakes the foundations of the entire American project," Harold Meyerson wrote in The Washington Post. Writing in The New York Times, Steven Rattner said that "even in a country that sometimes seems inured to income inequality, these takeaways are truly stunning." [Link Omitted]

(5) This recession was caused by the greed, the recklessness and the illegal behavior on Wall Street. [Link Omitted]

---

Equal Pay for Women

(1) 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!

(2) If there's no pay gap, why are women's earnings less than men's in 264 of 265 major occupations? Four-in-10 families rely on women as equal or main breadwinners, yet New Mexico women earn an average of \$8,789 less per year than men. Women deserve equal pay, and the Paycheck Fairness Act would help to close the gap.

(3) 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.

(4) It's astonishing and saddening to think that in 2012, women still get paid less than men for doing the same work. In Delaware, on average, women are paid 81 cents for every dollar paid to men, amounting to a yearly gap of \$9,505 between full-time working men and women. On Equal Pay Day, let's recommit ourselves to fighting to end pay discrimination.

(5) On the 50th anniversary of the Equal Pay Act, we still have work to do to ensure equal pay for equal work. Women in Virginia are paid 78 cents for every dollar paid to men, which jeopardizes the economic security of women and their families. I'm proud to co-sponsor the Paycheck Fairness Act to close loopholes in the Equal Pay Act, strengthen workplace protections and help end wage discrimination.

---

Resources  
for Local  
Communities

(1) North Dakota will receive more than \$190,000 in federal funds to help local communities improve emergency preparedness and support training programs to help first responders handle potential hazmat situations. Our emergency personnel need to have the resources and training to stay safe as they protect us, and these funds will help our state be ready for, mitigate, and respond to possible hazmat incidents. Safety should be everyone’s top priority — that’s what my RESPONSE Act aims to accomplish by giving first responders the resources and training they need, and I will continue to make sure we protect North Dakota’s communities and families.

(2) Survivors of violent crimes need our support and I will continue working to ensure they have the resources they need to get back on their feet.

(3) While we hope that flooding won’t affect our communities this spring, we need to make sure that we are prepared. In Washington, I am taking steps to ensure that homeowners are protected and our communities have the necessary infrastructure in place to combat floods. I encourage you to reach out to my District Office at [Number Omitted] if you have any questions or concerns.

(4) We need to do all we can to ensure that growing communities across North Dakota, and especially in western North Dakota, have the support they need to maintain the high quality of life we expect. We’re working to ensure that we have affordable, quality housing for our families and that our law enforcement officials have the resources they need to keep our communities safe.

(5) The FBI continues to make positive steps to address the safety needs of our state by building up the law enforcement infrastructure to support our growing communities. While the energy boom has delivered undeniably great economic benefits throughout North Dakota, it has also resulted in new law enforcement and social challenges our communities have never seen before. That’s why the FBI’s commitment to begin the process of opening a permanent office in Williston is such a needed development. The FBI and I have been working closely on this issue for about a year, and I’ll continue to provide as much assistance and guidance as possible as the agency works through the necessary steps to make that office a reality. We must provide the best possible resources to make sure that families and communities across our state are strong and safe.

---

Jobs &  
Keystone  
Pipeline

(1) 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.

(2) 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.

(3) Continuing her efforts to support programs that help New Hampshire workers develop the skills, knowledge and innovative thinking necessary for jobs in the economy of the 21st century, Governor Hassan announced today eight companies will receive job training grants to help them train 324 workers in new skills. The job training grants total more than \$148,000. The eight companies contributed matching funds, bringing the total funds for training workers to more than \$297,000.

(4) As part of her efforts to support programs that help New Hampshire workers develop the skills needed for jobs in the 21st century economy, Governor Maggie Hassan announced today four companies will receive job training grants to help them train 47 workers in new skills. The job training grants total \$14,585.50. The four companies contributed matching funds, bringing the total funds for training workers to \$29,171.

(5) Continuing her efforts to help New Hampshire workers develop the skills and innovative thinking needed for success in the good jobs of the 21st century economy, Governor Maggie Hassan announced today that seven New Hampshire companies have been awarded job training grants to help them train 220 workers in new skills. The job training grants total \$58,141, and the companies contributed the same in matching funds to bring the total amount for training workers to \$116,282. “I am excited to announce this latest round of important grants that will help New Hampshire workers develop new skills for high-quality jobs at our growing businesses,” Governor Hassan said. “New Hampshire’s Job Training Fund is a valuable resource that has helped provide thousands of workers with skills essential for success in the innovation economy. We must maintain our commitment to higher education and job training in order to attract innovative businesses, help existing companies grow, support the creation of good jobs and expand middle class opportunity.”

---

Trade

(1) American manufacturers already operate at a tremendous disadvantage due to China’s undervalued currency. Therefore, it is imperative that we crack down on Chinese currency manipulation before consideration of any of the free trade agreements.

(2) Read Ben’s op-ed on what Congress can do to bolster the U.S-#Africa trade relationship, elevate #humanrights and fight global poverty. “The U.N. projects that by 2035 Africa’s working-age population will be larger than that of China or India. The potential importance of this emerging demographic, both as producers and potential consumers, is difficult to overstate. #AGOA allows the U.S. and Africa both to take advantage of this dynamism.”

(3) I'm pleased the Senate voted for free & fair trade — American labor & American products win every time we have a fair chance to compete. Completing Trade Promotion Authority (TPA) legislation is a prerequisite to publicly debating trade agreements, such as the Trans-Pacific Partnership (TPP). The President has the Constitutional authority to negotiate trade deals unilaterally, but TPA allows the American people, through Congress, to have input on the final deal. Oklahoma & our entire national economy will greatly benefit from opportunities in more trade markets around the globe. The value of Oklahoma's exported goods & services already equals \$10.6 billion, according to the Oklahoma Council of Public Affairs. American manufacturers have a \$50 billion trade surplus with the 20 countries with which the United States has free trade agreements. Exports made up 13% of the economy in 2014 — up from 10% in 2000 & 7.2% in 1990. TPA will continue the surplus, & create many more opportunities for America's farmers, ranchers, manufacturers, & families. When the American worker is able to fairly compete on the world market, we win.

(4) As the Senate considers a 6-year Trade Promotion Authority (TPA) bill, to promote transparency, I am posting the entire bill publicly so Oklahomans can read it. Completing TPA legislation is a prerequisite to debating trade agreements, such as the Trans-Pacific Partnership (TPP). The President has the Constitutional authority to negotiate trade deals unilaterally, but TPA allows the American people, through Congress, to have input on the final deal. Feel free to read the bill here: [Link Omitted]

(5) 5,700 companies in Colorado export goods. Over 87% of those companies are small or medium-sized. Trade supports over 43,000 jobs in Colorado. Tearing down trade barriers helps Colorado companies grow, and helps the United States compete with other nations in the international market. That's why I support free trade deals like the Trans-Pacific Partnership.

---

Renewable  
Energy

(1) The winners & losers of ending America's long-standing crude oil export ban are clear [Link Omitted]

(2) Noatak, Alaska: Regular Unleaded gas at \$8.99 a gallon. Diesel/Heating Oil is the same price as Regular Unleaded. Thank you Don Fancher Jr.

(3) Why is Obama tapping oil reserves without adopting a national energy policy that will lessen America's dependence on foreign oil?

(4) Natural gas from American shale reduces reliance on imports and unreliable foreign suppliers.

(5) This will make us more dependent on imported oil as we transition to alternative energy sources.

---

Economic &  
Job Growth

(1) Today is #MFGDay15, a celebration of America's advanced manufacturing economy. Manufacturing has long been an engine of economic growth & innovation in West Virginia, and I am committed to expanding our manufacturing industry throughout our state.

(2) "Robust job growth requires boldness and risk-taking in the private sector. What we have now is boldness and risk-taking in the public sector."

(3) A #MadeInWI economy means innovative and exciting industries, from bioscience to digital technology manufacturing.

(4) I was glad to support the Rhode Island Textile Innovation Network this morning. I brought together this group of local textile manufacturers to foster partnerships and innovation in Rhode Island's advanced manufacturing sector, creating jobs and growing our economy.

(5) 1.6% growth of GDP anemic; we need policies that will get private sector off the sidelines to grow the economy and create jobs.

---

Celebrate  
Small Business  
messages

(1) Happy Small Business Saturday! Here I am with Minneapolis Mayor-elect Betsy Hodges & Creative Kidstuff owner Roberta Bonoff at Roberta's shop in Linden Hills. Small, local businesses like Creative Kidstuff are critical to supporting local economies. Last year, Small Business Saturday helped generate \$5.5 billion in consumer spending at small, local businesses like Roberta's. What are your favorite local shops?

(2) Locally owned small businesses are a cornerstone of communities across Indiana. When our small businesses succeed, Hoosiers succeed. #smallbiz #smallbizsat

(3) Shop locally today in support of our small businesses. Today is Small Business Saturday.

(4) Small businesses are the backbone of our communities — support the #MadeInWi businesses that are the heartbeat in your neighborhood! #ShopSmall for #SmallBizSat

(5) Today is Small Business Saturday! Support small businesses by shopping local.

---

Student Loan  
Rates & Minimum  
Wage

(1) 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.

(2) With federal student loan interest rates set to double on July 1, today I stood with student borrowers, graduates, Campus Progress and NYPIRG to announce the Federal Student Loan Refinancing Act, which would stop the rate from doubling and would allow borrowers to refinance their loans at a lower interest rate of 4%. Student loan debt has nearly tripled since 2004 and as a result, millions of families are burdened by higher levels of debt than ever before. To ease this burden, my bill would allow our young people to refinance in the same way that our businesses and homeowners do. You can read more about this legislation below: [Link Omitted]

(3) Raising the minimum wage to \$15 an hour is not a radical idea. It's supported by the majority of Americans. 1 in 3 New Jersey workers would benefit from increasing the minimum wage to \$15 an hour. We must raise the current starvation wage to a living wage.

(4) On July 1st, some student loan rates doubled from 3.4% to 6.8%. Tomorrow, the Senate will vote on Senator Jack Reed's legislation to retroactively freeze subsidized student loan rates at 3.4%. The House GOP is pushing for an alternative that would temporarily reduce the interest rate, but would eventually more than double the interest rate. We can't do that to the thousands of Rhode Islanders who have these loans and I look forward to casting my vote for Senator Reed's legislation tomorrow.

(5) The minimum wage is worth significantly less today than it was 50 years ago. The federal minimum wage of \$7.25 is a starvation wage. It's time to raise it to a living wage of \$15 an hour.

---