2018

Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?

Loni Hagen, *University of South Florida*

Information Processing and Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?

Loni Hagen

School of Information, University of South Florida, Tampa, USA

ARTICLE INFO

Keywords:
Topic modeling
Content analysis
E-petitions
Computer-assisted content analysis

ABSTRACT

E-petitions have become a popular vehicle for political activism, but studying them has been difficult because efficient methods for analyzing their content are currently lacking. Researchers have used topic modeling for content analysis, but current practices carry some serious limitations. While modeling may be more efficient than manually reading each petition, it generally relies on unsupervised machine learning and so requires a dependable training and validation process. And so this paper describes a framework to train and validate Latent Dirichlet Allocation (LDA), the simplest and most popular topic modeling algorithm, using e-petition data. With rigorous training and evaluation, 87% of LDA-generated topics made sense to human judges. Topics also aligned well with results from an independent content analysis by the Pew Research Center, and were strongly associated with corresponding social events. Computer-assisted content analysts can benefit from our guidelines to supervise every process of training and evaluation of LDA. Software developers can benefit from learning the demands of social scientists when using LDA for content analysis. These findings have significant implications for developing LDA tools and assuring validity and interpretability of LDA content analysis. In addition, LDA topics can have some advantages over subjects extracted by manual content analysis by reflecting multiple themes expressed in texts, by extracting new themes that are not highlighted by human coders, and by being less prone to human bias.

1. Introduction

Governments around the world have implemented electronic petition (e-petition) platforms to hear the voices of the public regarding their policy suggestions or concerns to governments (Dumas et al., 2015). Scholars argue that e-petitions provide mechanisms to enable the public to express their views to the elected officials (Bochel, 2013) and further demonstrated that these systems have impacts on policy decisions (Bochel, 2012). Countries such as Scotland, Wales, Great Britain, Germany, South Korea, and Australia, as well as the United States use e-petition platforms, expecting to increase citizen participation and government transparency (Hagen, Harrison, & Dumas, 2018). Despite the increasing popularity of these platforms, content analysis of e-petitions is largely lacking.

Traditionally, content analysis is used to systematically evaluate the symbolic content of recorded communication manually (Kolbe & Burnett, 1991). However, the increasing availability of large electronic archives, along with increased computational capacities, has led to interest in computer-assisted content analysis. Topic modeling is such a popular and efficient, computer-assisted technique for analyzing textual content (Chuang et al., 2014; DiMaggio, 2015; J. Grimmer, 2016; Lucas et al., 2015; Baumer, Mimno, Guha, Quan, & Gay, 2017; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). Topic models are especially useful for identifying

E-mail address: lonihagen@usf.edu.

https://doi.org/10.1016/j.ipm.2018.05.006
Received 9 October 2017; Received in revised form 15 May 2018; Accepted 15 May 2018

hidden topics and themes based on the word co-occurrence for each document in the corpus. In addition, topic modeling is a good way to "let the text talk" because the identified topics do not depend on the evaluators' individual perspectives or experiences (Graneheim & Lundman, 2004; Mimno, Wallach, Talley, Leenders, & McCallum, 2011).

Although topic modeling has numerous advantages, it also has limitations. Loss of *interpretability* is a major limitation of complex learning algorithms such as topic models (Witten & Frank, 2005). Topics produced by complex algorithms are difficult to interpret because their outputs, which are numeric values, are produced based on mathematical properties, while interpretation of them depends on the goals of the analysis, the researcher's perspectives and domain knowledge (Croft, Metzler, & Strohman, 2010; Suominen, 2016; Yau, Porter, Newman, & Suominen, 2014). Topic modeling involves several steps, and each decision influences the outcome. Because topic modeling outputs (e.g., topics and document rankings) are fully data-driven, the extent to which they accurately reflect themes written in the text is questionable. There are few studies that provide guidance for how each of these decisions should be made to generate dependable topics. Although acknowledged by some computer scientists (Boyd-Graber, Mimno, & Newman, 2014), these issues have not yet been fully investigated and discussed in the literature for content analysis.

To maximize the potential of topic modeling for content analysis, it seems critical to develop a systematic approach focusing on the interpretability and validity of topic model outcomes. To fill the gap in the literature, we offer a framework for topic modeling to extract emerging topics from e-petition data, then evaluate that framework by training LDA models and assessing the generated topics from human interpretability perspectives. The results of this study indicate that LDA is an efficient and a valid content analysis tool for finding latent streams of thoughts expressed in e-petitions. The findings demonstrate that, when properly supervised under our framework, LDA can help practitioners and social scientists analyze topical contents from large volumes of texts.

This article makes a primarily methodological contributions in the area of computer-assisted content analysis. It advances our understanding on how to systematically supervise computer-assisted content analysis processes to achieve interpretable results when using LDA models. It also argues that computer-assisted content analysis can complement manual content analysis by extracting new themes and suggesting different ways of categorizing textual contents. This article concludes by making suggestions to increase the dependability of topic modeling.

2. Literature review

2.1. We the people e-petitions

The Obama Administration introduced an electronic petitioning platform called *We the People* (WtP) to increase government transparency and encourage citizen participation (Open Government Partnership, 2015). WtP is considered "one of the most prominent legacies of the open government initiatives" (Hitlin, 2016, p. 2). As of July 2015, over 19 million individuals had created accounts on WtP, and over 411,546 petitions generated in the system had accumulated over 27 million signatures (The White House, 2015b; a detailed introduction of WtP is provided in Hagen et al., 2016). WtP's popularity was driven, in part, by the Obama administration's pledge to respond to any petitions that can attract at least 100,000 signatures over a 30-day period. Petitioners can solicit support using social media applications (e.g., Twitter and Facebook) embedded in the platform, and/or personal network and websites. For that reason, signature counts reflect the petition's popularity.

WtP provides 39 subject categories, which include subject categories that are conventionally recognized and to some extent mirror the organizational division of the federal bureaucracy. As Fig. 1 shows, the WtP categories are broad and provide a high-level structure for the organization and categorization of petitions. Further, the nature of the categorization method (petitioners selecting up to three categories from the 39 categories provided) indicates the WtP-categorization scheme may not be useful for describing the petition data for the purpose of understanding the issues and topics petitioners express. For example, Fig. 1 shows the distribution of petitions created between September 22, 2011 and January 3rd, 2015 on each of the 39 WtP categories. The top seven most popular categories include: Civil Rights and Liberties, Human Rights, Criminal Justice and Law Enforcement, Foreign Policy, Health Care, Family, and Government Reform. These categories together account for 52% of all petitions, the largest of which, *Civil Rights and Liberties,* and *Human rights* together covers 26% of all petitions (see Fig. 1). WtP-categories may be helpful to understanding the federal agencies responsible for specific petitions, but they may not be suitable for understanding issues and concerns petitioners express to the federal government.

In contrast to the WtP subject categories, actual discussions on the platform include diverse subjects—the most frequently initiated petitions were about health care, military and veterans' issues, illness, immigration, animal rights, holidays and criminal investigations (Hitlin, 2016). WtP's primary goal was enhancing citizens' participation in government (The White House, 2015a), but there has been little attempt to systematically understand what kinds of voices have been raised on the platform. Lack of content description of e-petitions may be attributable to an absence of methods to efficiently analyze the contents of plentiful and unstructured e-petitions. Traditionally, textual analysis has relied on content analysis approaches.

2.2. Use of topic modeling for content analysis

Content analysis is defined as "the systematic, objective, quantitative analysis of message characteristics" (Neuendorf, 2017, p. 1). Text is one of the most popular ways for authors to communicate their ideas and convey the social situation (Stone, Dunphy, Smith, & Ogilvie, 1966, p. 6). By characterizing text data systematically, content analysis is expected to provide new insights in understanding the essential ideas and further guide practical actions (Krippendorff, 2012, p. 24). Content analysis' popularity has risen dramatically since 2006, which may be attributable to the availability of large volumes of text data (Neuendorf, 2017, p. 3). Computing power to

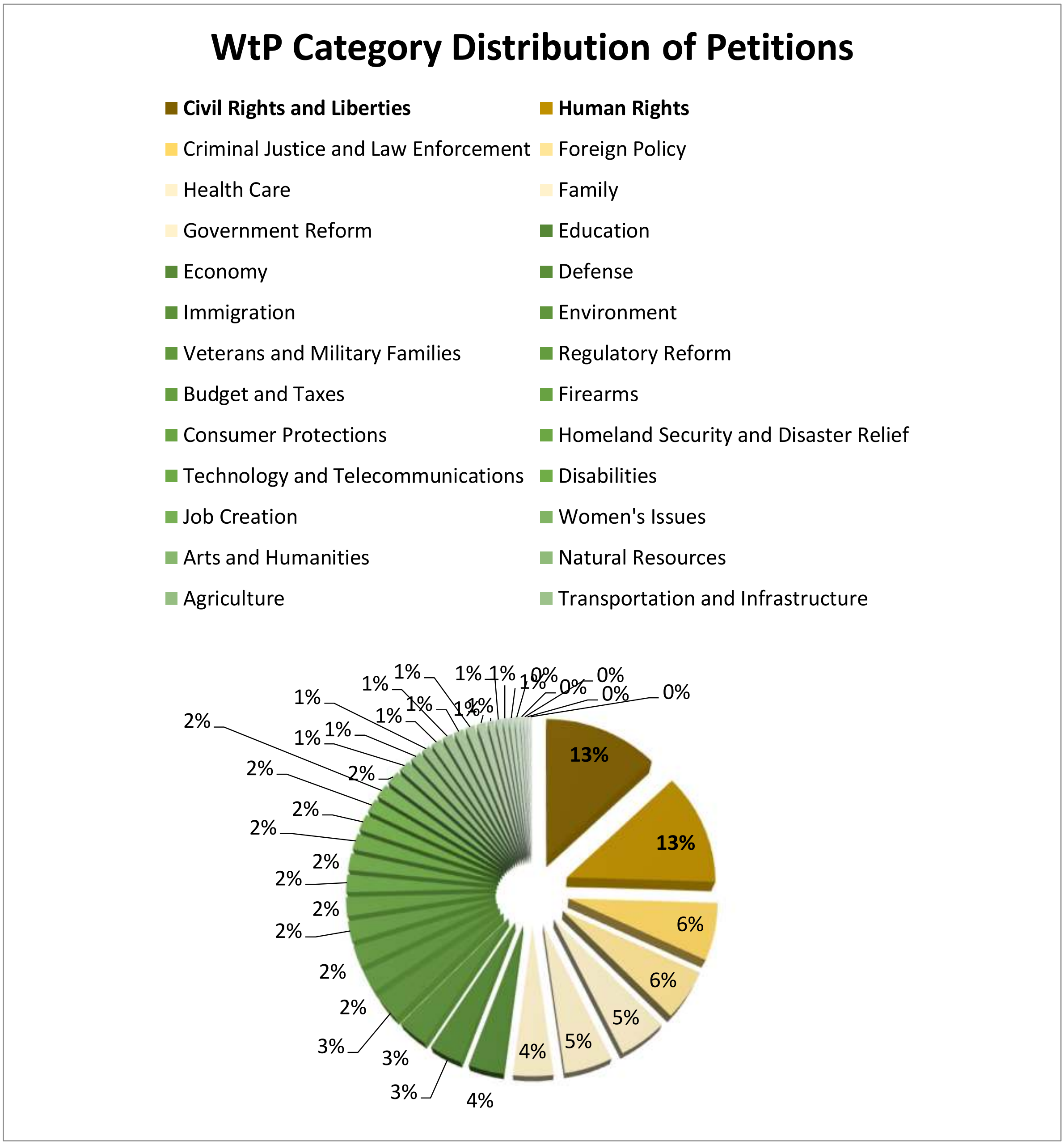


Fig. 1. Distribution of petitions on 39 WtP categories (Sept. 22, 2011-Jan. 3, 2015).

systematically and reliably process "big data," merged with content analysis, has invigorated computer-assisted text analysis. Some content analysis scholars used computers to implement statistical natural language processing (NLP) and text mining techniques. One of the most popular techniques for analyzing text is topic modeling (Aggarwal & Zhai, 2012, p. 107; Wei & Croft, 2006), which refers to statistical methods to discover themes running through data by analyzing the words appearing in the texts (Blei, 2012). One of the simplest and popular topic models is Latent Dirichlet Allocation (LDA) (Blei, 2012; Blei, Ng, & Jordan, 2003), which is an unsupervised, nonparametric and generative method that treats words in documents as if they were generated by probabilistic sampling based on latent variables (topics). A given document may contain multiple topics, and each topic represents a distribution over a fixed vocabulary. An LDA model provides two initial outputs: (1) an estimated probability of a document being generated by a topic (referred to as an *affinity score)* and (2) the probability of a word being used to represent a topic (referred to as a *topic*).

Social scientists have used topic modeling for automatic content analysis of textual big data to discover themes from political texts (Grimmer & Stewart, 2013; Quinn et al., 2010; Roberts, Stewart, Tingley, & Airoldi, 2013). These studies have shown that topic modeling can identify new themes from texts, *independent of potentially biased perspectives* (Hopkins & King, 2010; Jelveh, Kogut, &

Naidu, 2015). Quinn et al. (2010) extracted topics from legislative speech data (a total of 118,065 speeches) to infer the relative amount of legislative attention paid to specific topics (Quinn et al., 2010, p. 224). Roberts and colleagues have used topic models to identify framing differences between Chinese news sources and those in other countries (Roberts et al., 2014; Roberts et al., 2013). And, by using topic modeling on academic writings by economists, Jelveh and colleagues found that political ideology influences the results of economic research (Jelveh et al., 2015).

In addition, topic models made *new discovery* possible by enabling large volumes of text to be analyzed efficiently and effectively Grimmer (2010) found that the press releases of senators representing the same state were more similar than those of senators representing different states (Grimmer, 2010). That finding contradicts Schiller's (2000) prior study showing that senators representing the same state tended to express dissimilar priorities in order to compete for more media and public attention. Grimmer (2010) underscored that prior research was limited to a relatively small number of newspaper statements put out by a handful of legislators, and, further, stressed that his *new discovery* was made possible by topic modeling and the use of big textual data. LDA was also adopted for qualitative interpretation, leading to production of specific topics, which led to understanding different political styles in the women's movement (Nelson, 2014). Hagen (2016) found that white nationalism was strongly expressed in electronic petitions, which reflects middle class resentment of immigrants and racial minorities (Hagen, 2016).

Information science scholars have applied topic modeling for content analysis to identify authors of unsigned documents (Krippendorff, 2012, p. 18); to analyze overarching themes from survey data (Baumer et al., 2017); to improve performance of topic modeling for better analysis of large volumes of text data (Liu & Xu, 2017; Nikolakakou, Bothos, & Gregoris, 2015; Seshadri, Mercy Shalinie, & Kollengode, 2015; Song et al., 2016; Zhao, Jin, & Yue, 2015); to measure research popularity and dissemination channels based on research topics (Yan, 2015). Baumer et al. (2017) specifically noted that themes extracted using topic modeling are highly similar to human experts' analysis, but with only a small fraction of the time commitment.

Although LDA has demonstrated its value for content analysis in multiple studies, several challenge still remain when applying LDA to content analysis. First, the lack of reliable and generalizable processes is common problem in unstructured and big open data analysis (Zamith & Lewis, 2015). For example, one of the most difficult challenges in using LDA for content analysis has been determining the optimal number of topics to harmonize with human sense making (Grimmer, 2016; Suominen & Toivanen, 2016).

Second, reliably interpreting LDA outcome is another major challenge for content analysis. Traditionally, topic modeling research (Blei & Lafferty, 2009; Blei et al., 2003) has used "held-out likelihood" to estimate a model's validity when applied to a set of unknown documents. Probability-based observations, such as held-out likelihood, however, do not measure the extent to which topic model outcomes represent *internal* topics (Boyd-Graber et al., 2014). To ensure reliable and valid interpretations, LDA results need to be evaluated against standards of comprehensibility, not judged solely on probability-based observations.

In what follows, we investigate a method to produce generalizable and reliable LDA outcomes by training the model and evaluating its final iteration. The following subsections describe the process.

2.2.1. *Training topic models*

Training LDA models involves two main tasks: stemming and determining the optimal number of topics to produce (i.e., optimal K). Stemming consolidates words with a common stem. For example, "fail", "failed" and "failing" can all be consolidated to the stem "fail." Stemming can enhance topic quality by increasing word frequency and reducing the vocabulary required for modeling. Full stemming has been shown to increase topic quality (Hopkins & King, 2010; Quinn et al., 2010). Similarly, Hagen, Uzuner, Kotfila, Harrison, and Lamanna (2015) found that full-stemming (via the Porter stemmer—a popular algorithmic stemmer) out performed both no-stemming and minimal stemming[1] (pluralization and tense suffix removal). Hopkins and King (2010) recommend full-stemming since the trade-off between losing information by full stemming and reduced complexity is "well worth it" (Hopkins & King, 2010). In order to discern the best stemming strategy and interpretability of topics, we devised an experiment to answer the following research question:

RQ1: What is the better stemming approach between full-stemming and no-stemming?

In LDA, the number of topics (referred to as K) and the text data itself are the two initial inputs required for fitting models. K is specified by the user. Selecting the number of topics is regarded as one of the most challenging issues in topic modeling, "since there is really no good solution" (Croft et al., 2010, p. 387). No agreed-upon formula exists to predict the optimal number of topics. Instead, the best choice of K largely depends on the task and the size of data set (Baumer et al., 2017, p. 1403; Croft et al., 2010, p. 387; Leydesdorff & Nerghes, 2017; Nikolenko, Koltcov, & Koltsova, 2017, p. 94).

The optimal number of topics can be derived by probability (perplexity) or human judgment. Perplexity is a probability-based estimate of how well a model will fit a sample. It measures the effectiveness of a given set of parameters (calculated using the training set data) on a set of unknown data (Croft et al., 2010; Jurafsky & Martin, 2009). Statistically, perplexity is "equivalent to the inverse of the geometric mean per-word likelihood" (Blei et al., 2003, p. 1008). By convention, language modeling uses perplexity as the preferred measure for model evaluation (Blei et al., 2003; Wallach, Murray, Salakhutdinov, & Mimno, 2009). In order to avoid issues using one fixed test set for measuring perplexity, a preferred approach is cross validation. If models are stable, we should get similar effects by dividing the training data into sub-sets by using the training set "both as initial training data and as held-out data" (Manning & Schütze, 1999, p. 210). We can randomly partition the training data to k-subsets where all subsets are mutually exclusive and approximately equal in size. Ten-fold cross validation ($k = 10$) is the most popular.

Although perplexity is useful for evaluating a predictive model, a study conducted by Chang, Gerrish, Wang, Boyd-Graber, and

[1] Minimal stemming strips –ed and -ing from verbs, and –s from nouns.

Blei (2009) found that the best-fitting model and human judgment are negatively correlated. That is, models that had lower perplexity often had poor interpretability (Chang et al., 2009). So the authors recommend that practitioners of topic modeling should use evaluation methods that depend on "real-world task performance" rather than optimizing likelihood-based fitting measures (Chang et al., 2009, p. 8).

Alternatively, expert human judgment has been used to evaluate topic modeling outputs, especially when the modeling outputs are supposed to be used for human interpretation (Mimno et al., 2011; Mimno et al., 2011). Although human judgment is effective for judging the interpretability of topic modeling outputs, it is labor and time intensive (Baumer et al., 2017).

Because probabilistic and judgment based approaches each has its own strengths and weaknesses, a two-step process incorporating both seemed to be a reasonable approach for deriving an optimal K. In the first step, we would train the model using different values for K, running cross validation and measuring the performance (such as perplexity) for each iteration. Cross validation results would suggest a range of numbers for further investigation. In the second step, human judges would manually assess the *quality* of topics, and the optimal number would be determined by the K that produces the best quality topics within the range in the first step (Quinn et al., 2010).

The optimal number of topics, K, "must be large enough to generate interpretable categories that have not been over-aggregated and small enough to be usable at all" (Quinn et al., 2010, p. 216). If set K too low, the topics will be over-aggregated. Setting it too high—for example, fifty or more topics—will cause user fatigue and overload human effort, making the results uninterpretable (Yau et al., 2014, p. 777). Accordingly, our study sought to answer the following research question:

RQ2: When the goal is to interpret the topics, how does one decide the optimal number of topics (K) to produce?

2.2.2. *Evaluation of topic models*

As mentioned in Section 2.2.1, the outcomes of topic models have been "traditionally" evaluated by probability-based metrics such as perplexity, which yield a measure of held-out likelihood (Blei, 2012; Blei et al., 2003; Boyd-Graber et al., 2014). Held-out likelihood can address reliability (Trochim & Donnelly, 2006, p. 80). That is, a model with high held-out likelihood is considered to be reliable since the trained model (using the training set) can perform well on *unknown* data.

When used by themselves, likelihood-based comparisons, however, provide limited information for interpretability because they do not measure the ideas expressed in text (Boyd-Graber et al., 2014). Content analysis metrics should assess the meanings expressed in the text, but perplexity only measures the likelihood that the model will perform well on unknown data. Therefore, using only perplexity measures, it is impossible to discern whether the inferred topic-words adequately capture the ideas contained in the texts. In the following, we discuss three possible evaluation methods to address this limitation: direct human reading and judgment, comparing computer–human annotation results, and external validity.

When human judges are used to interpret topic model outcomes, the most effective approach is "careful reading of the underlying texts and of the model output by domain experts" (M. E. Roberts et al., 2015, p. 19). Alignment between LDA outputs and human understanding of them provides a measure of validity; that the outputs are measuring what they are supposed to measure, the meaning expressed in e-petitions. Humans can read a number of randomly selected documents that are assigned to each topic cluster and rate topics (top-x topic words) on a three-point scale based on the topic coherence (Newman, Lau, Grieser, & Baldwin, 2010). Adopting the manual reading approach, Quinn et al. (2010) found that information in the topic keywords "did an excellent job describing the documents assigned to each (substantive) topic" (Quinn et al., 2010, p. 218). Reading assigned documents can also provide additional context for the underlying meanings behind the topics (Quinn et al., 2010, p. 218). Therefore, our study sought to answer the following research question:

RQ 3: To what extent are LDA-generated topics interpretable by human judges?

Alternative evaluation approach is to compare the LDA-generated topic assignments with human-generated topic assignments. Once a model is trained, the model can automatically assign topics to new documents by holding the model parameters constant (Top right of Fig. 2). Measuring a model's reliability with new, unknown data assures that the learned model's performance can generalize beyond its initial training dataset (Creswell, 2013; Quinn et al., 2010). Human judges can assign each of the new documents to one of

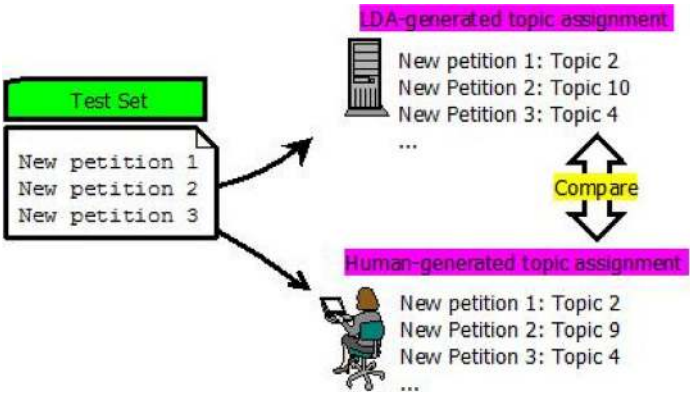


Fig. 2. An illustration of computer-human IRR measurement.

Table 1
A framework: content analysis assisted by LDA.

Phase	Task	Analysis strategy
Training (RQ 1 and 2)	Stemming	• Ten-fold cross validation
	Deciding the optimal K	• Human judgment
Evaluation (RQ 3, 4, 5, and 6)	Human reading	• Internal coherence between topics and documents • Generate an agreed-set of topics with topic quality and labels attached
	Computer–human comparison	• IRR between human and LDA topic assignment
	External validity	• Mapping social events with topics expressed in WtP petitions • Validate LDA topics against manual analysis of the same dataset by independent scholars

the topics (Bottom right of Fig. 2), then the human generated topic list can be compared with topic-assignment results from the trained LDA model. The level of agreement indicates the final model's reliability and gauges the interpretability of LDA-generated topics, and we posed the following research question:

RQ 4: To what extent do LDA-generated topic assignments correspond to human-generated topic assignment?

External validity suggests that findings from a given study can be generalized to different settings, persons, or times (Steckler & McLeroy, 2008). Since e-petitioning is a way for citizens to express their interests and concerns, we assume that the content of WtP petitions will reflect the information to which the citizens are exposed. If that assumption is correct, WtP topics should correspond to exogenous events (Quinn et al., 2010, p. 222). If the topics align with significant political events in the right period, it endorses LDA results as a valid means of measuring public opinions expressed in e-petitions. Our study examined that question:

RQ 5: To what extent do LDA-generated topics correspond to political/social events that are important to citizens?

The external validity of LDA topics can also be measured by comparing them with topic categorization results, using the same data, conducted by people who are not exposed to our LDA topics and analyses. This corresponding research question for this study is:

RQ 6: Do LDA-generated topics correspond to topics generated by an independent content analysis, using the same dataset?

2.2.3. Framework for content analysis using LDA

After reviewing the current literature, we developed a framework for content analysis using LDA that is suitable for analyzing topics to characterize e-petition data (see Table 1). The framework includes training LDA models to generate topics, and evaluating the generated topics.

3. Methods and findings

We conducted two empirical studies to train and evaluate the LDA models and outcomes following the framework devised in Section 2.2.3. The purpose of Study 1 was to demonstrate the process of stemming and deriving the optimal topic numbers by comparing automatic approaches with human interpretation. Study 2 assessed the validity and reliability of the LDA-generated topics. Both of the studies are evaluated from human interpretability perspectives.

3.1. Data

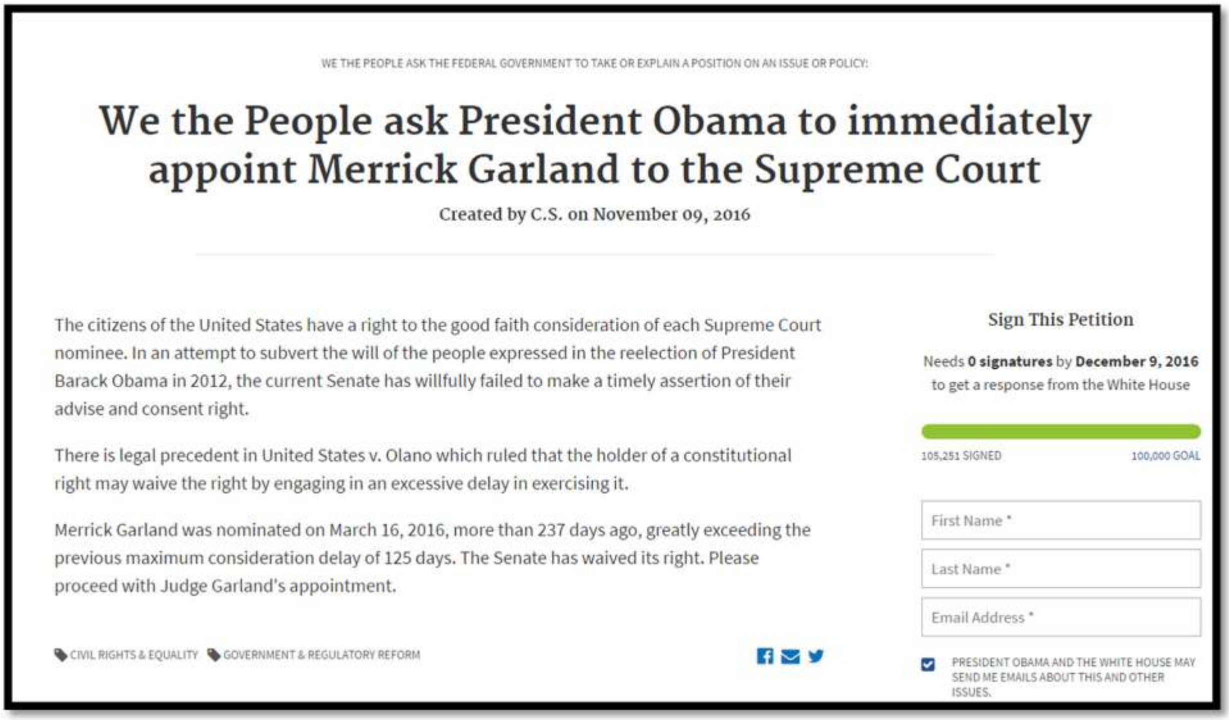
The data used for this study were obtained from a publicly available White House application program interface (API) containing information about all petitions displayed on the WtP website between September 22, 2011, when WtP was formally inaugurated, and January 3, 2015, when the initial data analysis took place. This corpus contained 324,594 tokens and 37,663 unique words in 3344 petition documents. Each petition is composed of a title and a rationale. We combined each title with its corresponding rationale into one document, which forms the basic unit for this analysis. Fig. 3 is an example of a WtP petition.

Using English petitions, we converted all text to lower case, normalized white space, eliminated punctuation, non-alphanumeric characters, and removed short words of only one or two characters using the Natural Language Toolkit (Bird, Klein, & Loper, 2009). We used unigram tokenization and minimal preprocessing because topic modeling prefers the rawest state with the simplest tokenization (Yau et al., 2014). After the preprocessing, the average length of a petition document—with combined title and rationale—was 97.07 words. The shortest petition contained zero tokens. We used the MALLET topic modeling package (McCallum, 2002) for training, thus, used MALLET's 524-word, built-in stopword dictionary for stopword removal.

Using the preprocessed set, we set aside a test set by randomly sampling 30% of the corpus. Using the remaining data as a training set, we ran and fit various models. Then, we assigned to each document a single topic—one with the highest probability of belonging to the document. We used the R programming language for analysis and visualization of LDA outcomes. In what follows, we report methods and findings from the two experiments: Training LDA Models and Evaluating the LDA Models.

3.2. Study 1: training LDA models

Although used to understand the major topics in large volumes of documents, it is not clear whether topics actually provide a semantically meaningful summary of the corpus. As discussed in 2.2, popularly used performance measures are based on statistical properties of texts, and, in fact, there is no theory-based reason to believe that probability-based measures are good indicator of



WE THE PEOPLE ASK THE FEDERAL GOVERNMENT TO TAKE OR EXPLAIN A POSITION ON AN ISSUE OR POLICY:

We the People ask President Obama to immediately appoint Merrick Garland to the Supreme Court

Created by C.S. on November 09, 2016

The citizens of the United States have a right to the good faith consideration of each Supreme Court nominee. In an attempt to subvert the will of the people expressed in the reelection of President Barack Obama in 2012, the current Senate has willfully failed to make a timely assertion of their advise and consent right.

There is legal precedent in United States v. Olano which ruled that the holder of a constitutional right may waive the right by engaging in an excessive delay in exercising it.

Merrick Garland was nominated on March 16, 2016, more than 237 days ago, greatly exceeding the previous maximum consideration delay of 125 days. The Senate has waived its right. Please proceed with Judge Garland's appointment.

CIVIL RIGHTS & EQUALITY GOVERNMENT & REGULATORY REFORM

Sign This Petition

Needs **0 signatures** by **December 9, 2016** to get a response from the White House

105,251 SIGNED 100,000 GOAL

☑ PRESIDENT OBAMA AND THE WHITE HOUSE MAY SEND ME EMAILS ABOUT THIS AND OTHER ISSUES.

Fig. 3. An example of an WtP petition

Note: The first two lines (bold and large font) are the title of the petition, and the rest of the text is the rationale of this petition.

individual topic-quality (Boyd-Graber et al., 2014). Relying solely on these measurements, it is not possible to know whether generated topics adequately capture the ideas contained in the texts. In order to demonstrate that inferred topics are "meaningful" and "interpretable," we incorporated human judgment, in addition to automatic measures, in the process of topic training. Since initialization influences topic outcomes, we used random initialization to verify that topic variations are small when topic numbers are 50 or smaller. Using small numbers of topics (between 5 and 30), Baumer et al. (2017) likewise observed that different initialization does not have a significant impact on topic variation when the number of topics is small.

The following sections describe the methods and findings from our investigation of the two research questions regarding LDA model training: stemming and optimal number of topics (K).

3.2.1. *Stemming*

We experimented with both no-stemming and full-stemming using ten-fold cross validation and human judgment, to decide which of the two stemming procedure resulted in more coherent topics. More specifically, we compared average perplexities using ten-fold cross validation varying the number of topics from 5 to 105 in increments of five. We also conducted a first-round coding task (on the topic-words) to assess the quality of topics generated by each of the two stemming approaches. Two human judges read topic words and coded as "1" if topic words can be grouped together as a single topic, "2" if topic words contain two topics, and "3" if topic words contain three or more topics, or do not seem to have interpretable topics. The author and another human coder preformed multiple pre-tests in order to produce a dependable coding scheme and then conducted additional pretests to measure the reliability of the scheme. Lower mean values indicate that topics are more coherent and interpretable by human judges (we call this topic quality).

Full-stemming was done using the Porter Stemmer in the Natural Language Tool Kit (Bird, Klein, & Loper, 2009, p. 107). In producing a spreadsheet for the full-stemming set, we replaced all stems in the 20 topic word clusters with the original, most commonly used word-form. This is done because some stems (for example, "caus" "abus" "genit" and "davier") can make interpretation difficult and therefore may cause bias in coding. For example, we replaced "police, polices, policed, policing" with the most frequently used word-form, "police" in preparation for human coding.

Both ten-fold cross validation and human coding results agreed that full-stemming was a better modeling strategy. Fig. 4 shows that models from the full-stemming set produced better quality topics (low perplexity indicates better quality topics). Results from human judgments agree with the perplexity results, that is full-stemming (average quality of topics: 1.43) to out-perform no-stemming (average quality of topics: 1.46) in producing high quality topics for human interpretation.

3.2.2. *Deciding the optimal K*

As a first step to determine the optimal K for an LDA model, we conducted ten-fold cross validation to determine a range of K values. Then, we conducted the human coding experiments within that topic range, in increments of five, setting an upper limit of 50,

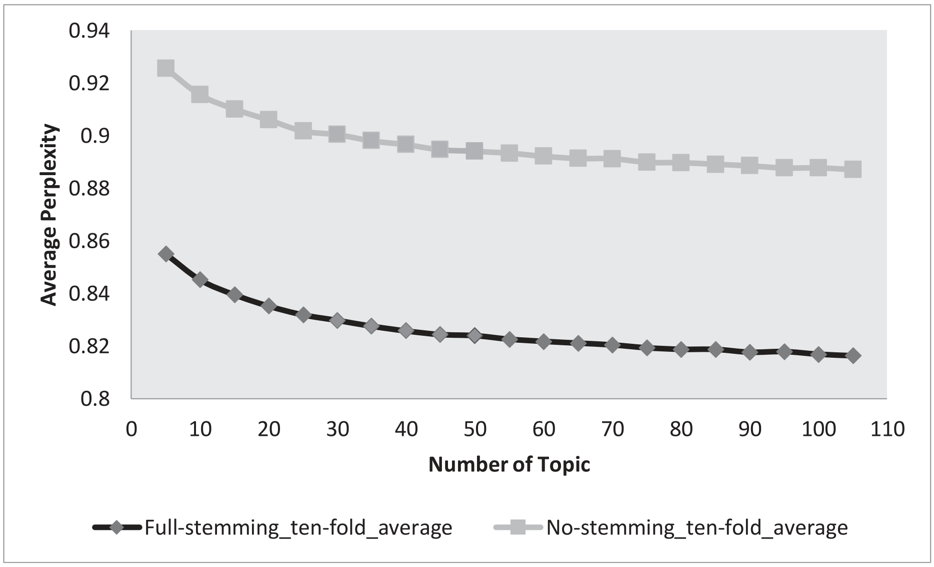
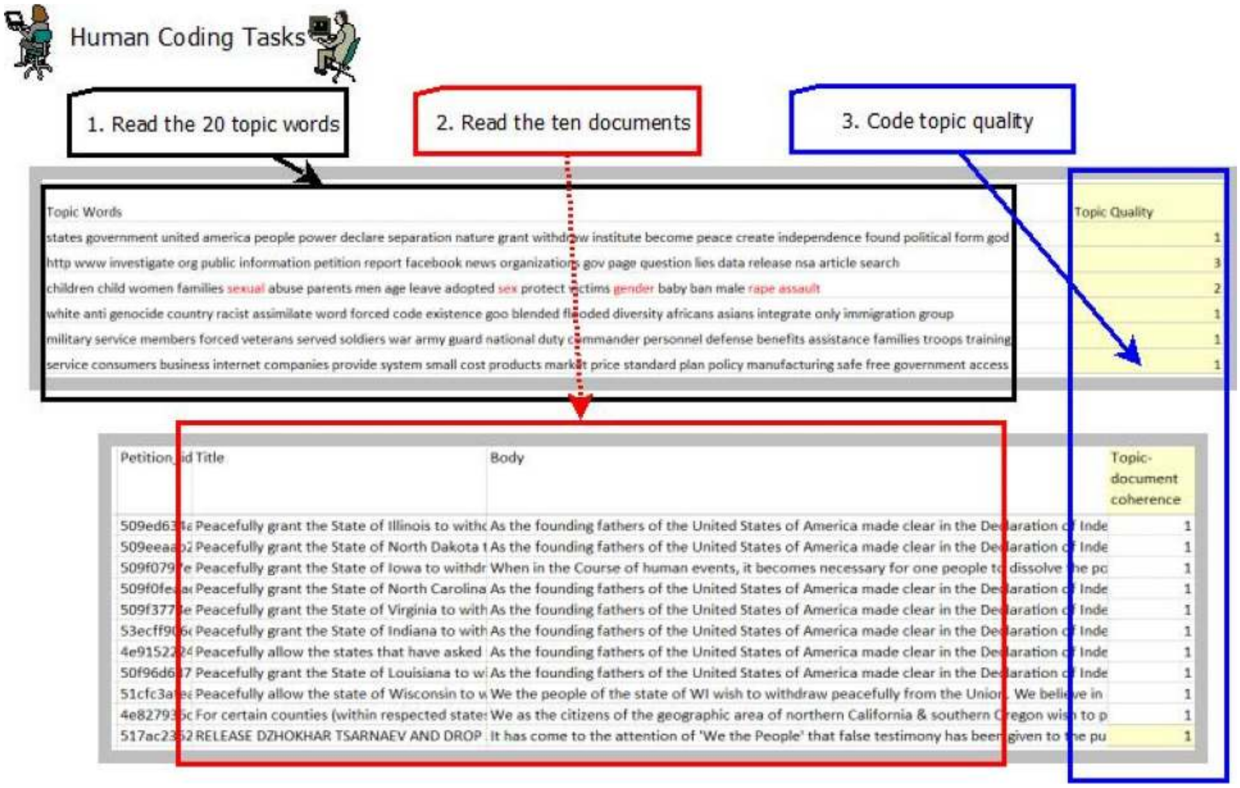


Fig. 4. Ten-fold cross validation using no-stemming and full-stemming sets *Note*: The black line is the average perplexity of ten-fold cross validation result from the full-stemming set, and the gray line is that from the no-stemming set.

as more than 50 topics is judged to be impractical to manually analyze. In this second-round human coding task, human judges were asked to evaluate: (1) *the quality of topic words*, and (2) *coherence between the topic words and the top-ten related documents*.

Because coding all models within a certain range of topics is not feasible, we selected only five LDA-models (i.e., 30, 35, 40, 45, and 50 topics) within the range (between 30 and 50) that was suggested by cross-validation. Human judges read 20 topic-words and ten topically-identified documents (see Fig. 5). Each ten-document group was composed of five documents with the highest affinity score for the topic, and other five documents randomly selected from the remaining documents assigned to that topic. This procedure



Topic Words	Topic Quality
states government united america people power declare separation nature grant withdraw institute become peace create independence found political form god	1
http www investigate org public information petition report facebook news organizations gov page question lies data release nsa article search	3
children child women families sexual abuse parents men age leave adopted sex protect victims gender baby ban male rape assault	2
white anti genocide country racist assimilate word forced code existence goo blended flooded diversity africans asians integrate only immigration group	1
military service members forced veterans served soldiers war army guard national duty commander personnel defense benefits assistance families troops training	1
service consumers business internet companies provide system small cost products market price standard plan policy manufacturing safe free government access	1

Petition_id	Title	Body	Topic-document coherence
509ed634a	Peacefully grant the State of Illinois to withc	As the founding fathers of the United States of America made clear in the Declaration of Inde	1
509eeab2	Peacefully grant the State of North Dakota t	As the founding fathers of the United States of America made clear in the Declaration of Inde	1
509f079'e	Peacefully grant the State of Iowa to withdr	When in the Course of human events, it becomes necessary for one people to dissolve the pc	1
509f0fe4ac	Peacefully grant the State of North Carolina	As the founding fathers of the United States of America made clear in the Declaration of Inde	1
509f3773e	Peacefully grant the State of Virginia to with	As the founding fathers of the United States of America made clear in the Declaration of Inde	1
53ecff906(	Peacefully grant the State of Indiana to with	As the founding fathers of the United States of America made clear in the Declaration of Inde	1
4e915224	Peacefully allow the states that have asked	As the founding fathers of the United States of America made clear in the Declaration of Inde	1
50f96d637	Peacefully grant the State of Louisiana to wi	As the founding fathers of the United States of America made clear in the Declaration of Inde	1
51cfc3ae	Peacefully allow the state of Wisconsin to w	We the people of the state of WI wish to withdraw peacefully from the Union. We believe in	1
4e82793c	For certain counties (within respected state:	We as the citizens of the geographic area of northern California & southern Oregon wish to p	1
517ac2352	RELEASE DZHOKHAR TSARNAEV AND DROP	It has come to the attention of 'We the People' that false testimony has been given to the pu	1

Fig. 5. Human coding process for deciding the optimal K.

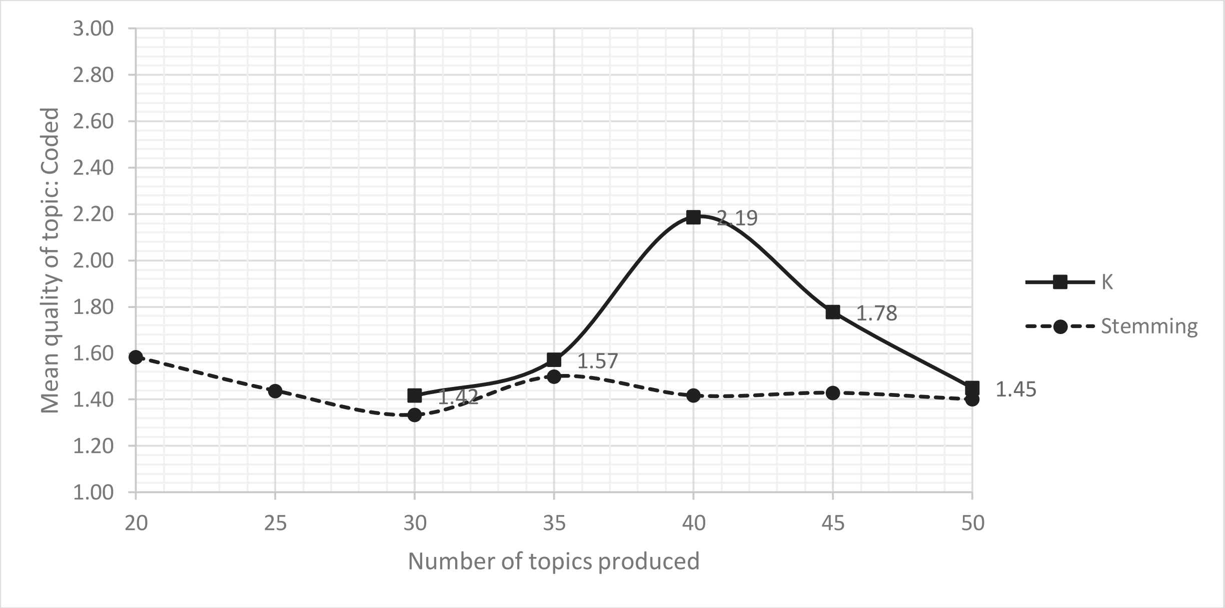


Fig. 6. Topic quality. *Note*: The line is the result from the K-coding task using models of 30, 35, 40, 45, and 50. The dotted line is the result from stemming-coding task using models of 20, 25, 30, 35, 40, 45, and 50. *Y*-axis value close to "1″ indicates the K-topic model produces high quality topics that are more human-interpretable.

was designed to mitigate possible bias in affinity score calculations, and to increase the breadth of document coverage, rather than just providing ten nearly identical documents.

Next, we randomly sampled 20% of topics for human coding (i.e., 6 topics from the 30-topic model, 8 topics from the 40-topic model). Two coders, each familiar with U.S. politics, independently rated the quality of topics and the coherence between the topics and the ten-related documents. As illustrated in Fig. 5, for each topic, each coder first (1) read the 20 topic words and (2) read the ten documents, and (3) independently coded topic quality following a prescribed coding scheme. Among the five topic-models (30, 35, 40, 45, and 50), we selected the model with the best topic quality to be the optimal topic number (K).

To minimize preprocessing, we did not eliminate proper nouns, although this made coding somewhat more complex. Instead of dropping all proper nouns as Chang et al. (2009) did, we allowed coders conduct a Web search when the relevance of a given word to the topic was unclear. This turned out to be useful. For example, in the *religion_gay* topic ("religious, church, gay, god, religion, christian, marriage, national, couples, anti, lgbt, beliefs, pledge, ***uganda*,** equal, sex, trust, muslim, word, abortion"), it was not clear whether "uganda" was relevant. Google search using "gay" and "uganda" showed that the Anti-Homosexuality Act was passed in 2014 by the Parliament of Uganda, which confirms that the proper noun "uganda" is a relevant topic-word for the *religion_gay* topic.

Results of the K-coding task are shown in Fig. 6. The line in Fig. 6 indicates that the 30-topic model produces the most coherent themes (average quality: 1.42) followed by the 50-topic-model (average quality: 1.45). For comparison, the dotted line in Fig. 6 shows the coding result from the stemming-coding task, which also indicates the 30-topic model produces the most coherent set of topics according to human judgment. We found that human-judged topic quality seems to be positively correlated with the perplexity before the perplexity significantly drops (i.e., 5, 10, 15, 20 topics). But, once the perplexity starts stabilizing (i.e., topics bigger than 25), human-judged topic quality is not correlated with the perplexity results. This is similar to the findings of Chang et al. (2009). Therefore, perplexity seems to be correlated with topic quality while perplexity significantly drops, but it is not correlated with human-judged topic quality when the perplexity starts stabilizing.

Although the human-based coding used 20 topic-words and 10 topically-identified documents, adding those 10 documents actually decreased topic quality. The documents may have provided a reference point for judging some ambiguous topic words to be irrelevant, whereas without documents such ambiguous cases were assumed to be relevant. Anecdotally, one coder stated that, on a few occasions, a topic seemed coherent but the documents did not quite support some of the topic words. Presumably, this occurred, in part, because half of the 10 documents were randomly selected.

3.2.3. Reliability of human coding results

To assess the reliability of human coding results, we calculated inter-rater reliability (IRR) (Krippendorff, 2004; Lombard, Snyder-Duch, & Bracken, 2002; Neuendorf, 2017). IRR measures agreement between coders. After experimenting with multiple measures of IRR, including percent agreement, Gwet's AC, Scott's pi (π) (Scott, 1955), Cohen's Kappa (κ) (Cohen, 1960), and Krippendorf's alpha (α) (Krippendorff, 2012), we report Gwet AC_1. This is because several studies (Girard, 2015; Gwet, 2002; Viswanathan & Berkman, 2011; Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013) have found that Gwet AC_1 is most appropriate for a highly skewed data sets such as ours (on average, 87% of topics are judged to be good quality and only 13% are judged to be bad quality). We

Table 2
LDA-topics with reconciled set of label and topic quality.

Label	10 topic words
marijuana*	marijuana drug legal cannabis medical federal substances control alcohol regulations
secession**	states government united america people power declare separation nature grant
http	http www investigate org public information petition report facebook news
energy**	energy jobs oil fuel american nuclear create pipeline project industry
mcclellan bill*	mcclellan veterans committee chemicals bill act health house toxic congress
election	day election national vote holiday voter america recognize language presidential
time	time people make many please years only families country states
animals**	animals dogs horses human slaughter pet killed ban wild trade
religion_gay**	religious church gay god religion christian marriage national couples anti
immigration_visa**	visa immigration residents process card application deported work green legal
east asian countries**	china japan south chinese korea human world korean government people
investigate*	investigate food justice prosecute criminal attorney labeling civil case violation
guns*	gun firearms ban weapons control safety carry faa shooting armed
terrorist_isis**	attack terrorist Israel war killed syria aid government group military
sexual_abuse**	children child women families sexual abuse parents men age leave
medical**	health cancer medical disease care treatment research patients disability funding
law_enforcement_abuse*	police office ebola law enforcement protect killed county department video
tax_fund**	tax funding federal years pay cuts budget dollars cost jobs
president_obama	president obama administration american house national petition support request states
russia_ukraine**	russian ukraine russia ukrainian sanctions government international armenian putin territory
education**	school students education public program test children research college teachers
criminal_sentences**	years sentenced prison murder charges crime pardon convicted man release
loans**	loans debt federal home reserve bank market payment interest financial
protect_wildlife**	protect water park national land species wolves endangered act wildlife
foreign_human_rights_and_liberty**	peace government protesters human people police violence political democracy freedom
white_ genocide**	white anti genocide country racist assimilate word forced code existence
military_issues**	military service members forced veterans served soldiers war army guard
internet_business**	service consumers business internet companies provide system small cost products
medal_of_honor_awareness**	awareness medal honor national house lives light award day raise
constitutional_issues**	states law act united government citizens federal protect amendment constitution

Note: The first column is the reconciled set of labels. Asterisks indicate topic quality judged by human judges.

** Indicates "good quality,"

* Indicates "fair quality," and no asterisk indicates "poor quality" topics. We replaced the stems in topic words with the original word-type.

implemented the reliability coefficients using the R function provided by Advanced Analytics, LLC (2010), and acquired 0.82 (Gwet AC$_1$). Considering (1) the exploratory nature of the study and, more importantly, (2) the challenge of coding topic words that require prior knowledge/expertise, the reliability coefficients 0.82 seems to be reasonably good for drawing reliable conclusions.

3.3. *Study 2: evaluating the LDA model*

Having determined the optimal $K = 30$ and having chosen the stemming method, we generated 30 topics using the training set. Then, we adopted three approaches—human reading, computer–human coding comparison, and external validity—to evaluate the generated topics and the final 30-topic LDA model. In the following section, we report the method and findings pertinent to the four research questions involved with evaluating the LDA model.

3.3.1. *Human interpretability*

For the 30 topics, we conducted a third-round human-coding task to evaluate topic quality. For the human coding task, we created two worksheets: the first contained 30-topics, each composed of 20 topic words, and the second contained 10 documents per topic (30*10 rows where each row contains a petition). For each topic, two coders independently read the 20 topic words and the ten documents, then (1) rated *topic quality*, (2) *internal coherence* between topics and assigned documents, and (3) *assigned labels* to topics.

The *topic quality* coding followed the same process as 3.2. After independent coding, the two human judges created the final reconciled set, in which 21 topics (70%) were coded as "1" (good quality topic with one coherent theme), 5 topics (17%) were coded as "2" (intermediate quality topic with one or two themes with existence of irrelevant word), and 4 topics (13%) were coded as "3" (poor quality topic). In sum, a total of 87% were coded as interpretable, and Table 2 shows the topic quality of each topic in the final reconciled set. The percent agreement between the two raters was 0.83, and Gwet AC$_1$ of 0.78, suggest good inter-rater reliability (Creswell, 2013; Quinn et al., 2010; Trochim & Donnelly, 2006, p. 87). To assess *internal coherence* (between topics and documents), human judges rated each of the 10 documents as "1" if it was relevant to the given topic words, and "2" if it was not relevant. Two human judges coded 275 of the 300 documents (ten documents per topic)–the remaining 25 documents were not coded because their topic words were coded as "3" (poor quality topics). We found that the topic quality and the ten-assigned documents are highly correlated (Pearson's correlation coefficient of 0.70). In other words, although not perfect, topic words successfully represent the themes expressed in the assigned documents. The percent agreement of the two human judges was 0.88, and Gwet's AC1 was 0.83, thus suggests the coding results are reliable.

For human-labeling, two annotators read through the 20 topic words and ten documents assigned to the topic, and wrote a label (a word or phrases selected from topic words when possible) that reflected the meaning of the topic in a comprehensive way. After coding independently, the two human judges discussed the coding results and produced a reconciled set of labels. Table 2 shows the reconciled set of label, topic quality (indicated by number of asterisks), and ten topic words for each of the 30 topics.

3.3.2. Computer–Human IRR

We evaluated the LDA model performance against two criteria: (1) how well the 30 topic-model, fitted with the training set, successfully assigned topics on unknown petitions (the test set), and (2) how well assigned-topics in the test set, corresponded to human judgments (illustrated in Fig. 2). First, we assigned each document in the test set to the topic with the highest affinity score (we call this *computer-coding*). Then, we randomly sampled 5% (49) of the test set petitions and manually assigned each of them to a topic (we call this as *human-coding*). Then, we calculated IRR between computer-coding and human-coding results.

Initial coding consistency between humans and computers was quite low (Cohen's Kappa: 0.44) for the 49 petitions randomly sampled from the test set. That may be because we did not factor in an important LDA assumption—that a document is composed of multiple topics. Accordingly, we iterated to allow multiple membership (the top four topics with the highest affinity scores) and re-measured agreement between human and computer coding. We considered human–computer coding to "agree" when one of the human-coding topics corresponded to one of the top-four computer-assigned topics. With this method, Cohen's Kappa coefficient between computer and human was 0.76. This result suggests the learned 30-topic-model can predict topic assignment on unknown petitions, only when we allow multiple membership of a document. In the following sections, we report the results of the external validity checks: assessing perplexity, matching with social events, and comparing with Pew content analysis results.

3.3.3. External validity

Holding constant the learned 30-topic-model parameters, we estimated perplexity on the test set and compared it with the perplexity estimates from the training set's 10-fold cross validation. Perplexity estimates for the test set and training set were highly correlated (Pearson's correlation is 0.99 with p-value < 0.000). This suggests a high degree of confidence that the 30-topic model from the training set can be applied to unknown sets of WtP data, and is thus generalizable.

We also compared the LDA-generated topics with important social and political events expecting that WtP petitions may reflect the information to which the public pays attention. If the topics corresponded to significant political events in the right period, it can validate the LDA results to some extent. We decided to select the *guns* topic for an external validity check. Dumas et al. (2015) found that a network of people mobilized around gun control policies in WtP petitions after the Sandy Hook Elementary School shooting when gun policies were a prominent focus of public (Dumas et al., 2015). If the WtP petition platform correctly reflects the public's concerns, and LDA results are valid, then we would expect public attention to gun-related issues also to be reflected in media reports.

Using New York Times as a proxy for news media reports, we collected all New York Times news articles between December 23, 2012 and February 2, 2015 using ["gun" OR "firearm"] as the query from the LexisNexis database. We counted the number of appearances of the topic words (i.e. gun, firearms, weapons, ban, carry, control, law, armed, legislation, assault, amendment, shooting, violence, protect, bear, crime, free, prevent, capacity, any) in the 150 articles retrieved (we call this the New York Times corpus), and aggregated dates by week.

LDA-generated topics corresponded to media coverage of gun regulation debates (a significant social/political event). We visualized the number of topic-words for the "guns" topic in both NY Times corpus and WtP petitions in Fig. 7. The NY Times corpus shows a large increase in the frequency of the topic words in the *guns* topic shortly after December 14, 2012, when the Sandy Hook Elementary School shooting occurred. The first major spike on the WtP *guns* topic happened at approximately the same time (see

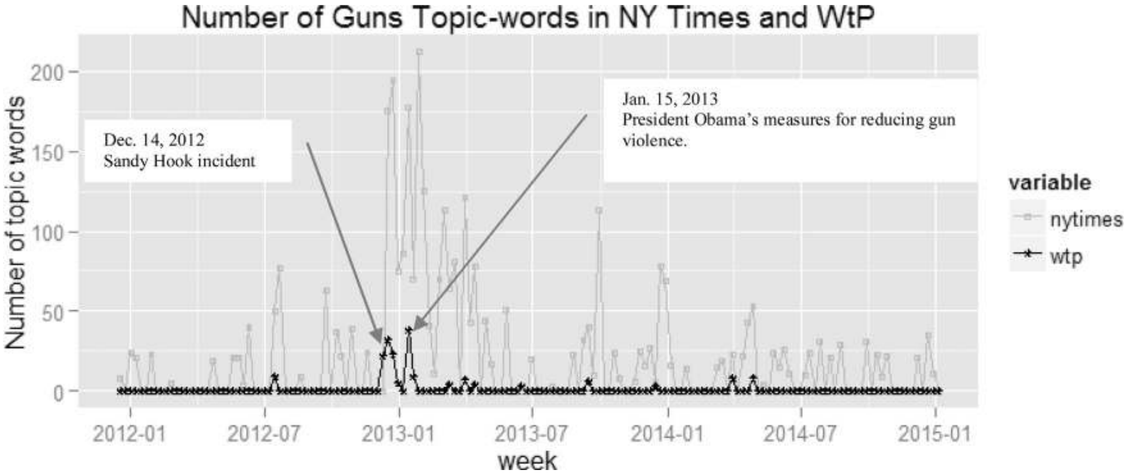


Fig. 7. Social events reflected in *guns* topic in WtP petitions and NY times corpus.

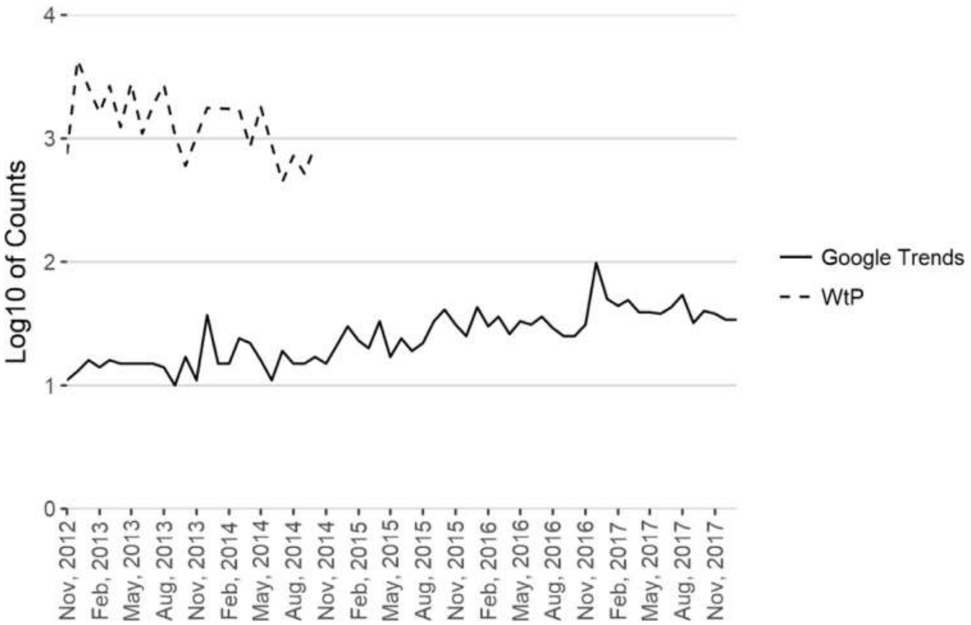


Fig. 8. Google Trends and WtP topics
Note: Y-axis is logarithm of number of google search query (Google Trends line) and number of signatures (WtP dotted line).

Fig. 7). Pearson's correlation between the *guns* topic in the NY Times corpus and in the WtP petitions is 0.6, which shows high correlation regarding the *guns* topic in these two platforms. Overall, according to the quantitative interpretation of topic-words, WtP activity seems to correspond to some important political events such as the Sandy Hook elementary school shooting and President Obama's measures—to include executive orders and legislative proposals—for reducing gun violence. This suggest that e-petition topics may sometimes be largely a reflection of media attention to major social events. However, as the next example shows, this does not always seem to be the case.

We also compared LDA topics with Google Trends search results (Fig. 8). Google Trends provide the relative number of search terms used in the Google search engine, and therefore can be used as a proxy to indicate the ups and downs of public attention to issues. Using "white genocide" as the search term, we extracted Google search frequencies between Dec. of 2011 and Jan. of 2018. When compared with LDA topics, the popularity of this search term increases over time in the Google search engine. In contrast, the *white_genocide* topic popularity in WtP platform was initially high in 2011 and 2012 but decreased in popularity thereafter. This suggests that e-petitions may reflect social information even before it become widely noticed.

3.3.4. Comparison with Pew content analysis

A good validation is to compare the LDA topics against manually coded results using the same data. Independently from our study, Pew conducted manual content analysis and published 25 subjects using 4799 WtP petitions created between Sept. 22, 2011 and July 3, 2016. Comparing LDA topics with Pew subjects shows the extent to which LDA topics correctly categorize subjects defined and coded by humans. Table 3 shows how those 25 subjects correspond to the 28 LDA-generated topics after dropping the two LDA topics (*http* and *times*) that are uninterpretable. It should be noted that the dataset used for the LDA analysis does not include the year 2015 and 2016, while those years are included in the Pew dataset. That difference likely explains why "2016 presidential campaign" (one of Pew's subjects) does not exist in the LDA topics. We report LDA topics in Table 3 under the headings "Identical to Pew" and "New themes."

Our LDA topics aligned surprisingly well with an independent content analysis by the Pew Research Center. For example, one of the most popular Pew subjects is "requests to honor individuals or create holidays" where petitioners request the White House "to honor an individual or create a national holiday" (Hitlin, 2016, p. 14). Our investigation on the LDA topic *medal_of_honor_awareness* reveals that the majority of these petitions request awarding the Medal of Freedom to certain people (i.e. police officers, firefighters, or teachers), and to proclaim a certain day as a national awareness day. Similarly, the majority of LDA topics, as presented in Table 3, closely resembles Pew subjects, which validates our framework and the final results.

4. Discussion

Citizens' access to political information and direct forms of political action such as boycotts, protests, and participation in citizen action groups has risen (Dalton, 2013). Scholars and practitioners rightly agree that we need to use diverse information sources and computational tools for understanding citizens' opinions expressed in large datasets related to those political actions. As a result, computer-assisted content analysis has arisen as a field of study to automatically develop categories and assign texts to categories.

Table 3
Pew categories and LDA topics.

25 Subjects (Pew: 4799 petitions)	28 Topics (LDA: 3,344 petitions)	
	Identical to Pew	New themes
Health care system	**medical**	
All specific illnesses combined		
Military and veterans' issues	military_issues	**mcclellan bill**
Immigration	immigration_visa	
Requests to investigate specific criminal cases	investigate	sexual_abuse
Request to issue pardons	criminal_sentences	
Requests to punish public officials (other than Obama)		
Requests to honor individuals or create holidays	**medal_of_honor_awareness**	
Animal rights	animals	**protect_wildlife**
Education policy	education	
Gun control/gun rights	guns	
Police and justice system	law_enforcement_abuse	constitutional_issues
China	**east_asian_countries**	foreign_human_rights_and_liberty
Russia and Ukraine	russia_ukraine	
Terrorism (domestic or international)	**terrorist_isis**	
Middle East issues combined		
Religious issues	**religion_gay**	
LGBT issues		
Technology	internet_business	
Marijuana/war on drugs	marijuana	
Obama administration	president_obama	**secession**
White nationalism	white_ genocide	
Energy and climate change	energy	
2016 presidential campaign	election;	
Taxes	tax_fund	**loans**

Note: Pew Research Center analysis of WtP petition is based on all petitions with a minimum of 150 signatures created between Sept. 22, 2011, and July 3, 2016. LDA-topics are generated based on the same petition data created between Sept. 22, 2011 and Jan. 3, 2015. Out of the 30 topics, *time* and *http* are excluded from this analysis because these two topics are uninterpretable.

Although low analysis costs using large dataset is the clear benefit, the field has lacked frameworks that facilitate the interactions between computational tools and human experts (Zamith & Lewis, 2015; Gill et al., 2017). Our work is an attempt to find and harmonize the best of the two worlds (computational tools and human judgments) by incorporating close supervision of human judgment in the process of LDA training, evaluation, and interpretation.

We devised a framework for content analysis using LDA, and tested the interpretability and validity of the generated LDA topics from human perspectives. The results show that, when closely supervised according to our suggested guideline, which combines human judgments and automated measures, 87% of the final LDA topics are interpretable. In other words, when closely supervised, LDA topics correspond closely with topics based on manual coding. In addition, these topics are valid in comparison with a test set, external social environments, and an independent manual content analysis by Pew Research Center.

As a result of comparing with the Pew subjects, we found that the results of LDA topic modeling have three advantages over manual content analysis. First, LDA topics reveal the multi-dimensional nature of written political texts by extracting topics with multiple themes; *terrorist_isis* and *relition_gay* topics are the two examples. The *terrorist_isis* topic encompasses two Pew subjects ("terrorism" and "Middle East issues combined"). An investigation of the top 20 relevant petitions reveals that the majority of these petitions condemn terrorist activities in Middle East countries, which explains why the LDA topic encompasses the two seemingly separate topics. Similarly, the LDA topic, *religion_gay*, also comprises two Pew subjects ("religious issues" and "LGBT issues"). An investigation of the top 20 relevant petitions reveal that many petitions in this topic combine LGBT issues with religion. For example, a petition titled "Federally legalize gay marriage" argues that religious and civil marriages should be separated. Another petition titled "Replace anti-gay Pastor Louie Giglio for the benediction at the inauguration with a pro-LGBT member of the clergy" requests the White House to replace a clergy from the benediction at the inauguration because of his anti-gay positions in the past. Other petitions request to take away the tax exempt status of specific churches due to their anti-gay marriage positions. These LDA topics interestingly highlight the difficulties of content analysis using political text because one text often includes multiple themes. And, at the same time, they show possible contribution of LDA for content analysis that naturally reveals the multi-faceted nature of contents in written texts.

Second, LDA topics extracted new themes that were not highlighted in the Pew subjects. For example, an LDA topic called *mcclellan bill* contains numerous petitions between 2011 and 2013 that request the White House to help pass a bill that contains some measures of support to the veterans who were exposed to toxic chemicals at Ft McClellan. Similarly, *secession* topic is not reported in the Pew subjects. We found over 35 petitions in the training set requesting secession from the US, which tended to be proposed right after the reelection of President Obama on November 6, 2012 (Kasperkevic, 2012). Significantly, a petition calling for the state of Texas to secede from the union attracted 125,746 signatures. The *loans* topic is another example. The majority of the petitions assigned to the *loans* topic request that the government revise federal loan policies regarding home or student loans. More specifically, several petitions request Fannie Mae to reduce principles for homeowners. Other petitions request the Federal Reserve to review questionable banking and mortgage lender activities. A petition titled “Allow all STUDENT LOAN interest to be deductible or a credit for Federal Tax purposes” request the federal government to make changes in their tax policies to benefit people with student loans. Finally, *protect_wildlife* is another example that was not identified in Pew subjects. This topic includes petitions requesting to protect wildlife by keeping national reserves, wilderness areas, and stop environmental hazards. Several petitions request the federal government to preserve gray wolf under the Endangered Species Act. These results demonstrates possible contributions of LDA topics by extracting topics that are honest expression of the public but somehow receiving less attention by human coders.

Finally, LDA topics can complement manual content analysis by extracting topics with significantly lower levels of human bias. The LDA topics are relatively independent of potentially biased perspectives because the computer does the coding. For example, an LDA topic *white_genocide* is constructed based on the most relevant topic words, whereas Pew named this as “White nationalism,” which does not reflect that fact that the term “white genocide” is a coded term used by white supremacy groups “as a way of suggesting that white people in the United States are under threat by virtue of shifting demographics or, more recently, political correctness” (Bump, 2016). This movement became so popular that the Twitter handle @WhiteGenocideTM (suspended as of Jan. 2018) was mentioned and retweeted by presidential candidate Donald Trump on 2016. Another LDA topic *law_enforcement_abuse* corresponds to the Pew issue subject of “Police and justice system.” The LDA topic captures petitioners’ resentment about brutality and abusive behaviors by law enforcements, which the Pew subject does not.

This study contributes to computer-assisted content analysis methodology by suggesting clear and generalizable guidelines. After empirically testing our guidelines, we make the following recommendations for the training and evaluation phases. First, during the training phase, we recommend minimal preprocessing. For example, we do not recommend dropping proper nouns since it may change outcome topics by omitting important information. We instead recommend using Google search for coding when the relevance of topic words is ambiguous. Second, full-stemming seems to produce better quality outcomes. In order to decide the number of topics to produce, we recommend using perplexity to narrow down the range of topics (to about five) for human judgment, the result of which will be used to decide the final K. Third, LDA-generated topics should be evaluated to ensure replicability, topic quality, internal coherence between documents and topics, and external validity. Distinctively from other topic modeling studies, we recommend supplying full documents in addition to topic-words for human judges’ topic quality coding, in order to provide important contextual information about the topic.

This study also informs topic modeling developers that computational content analysis is a field of a study that can utilize small number of topics using topic modeling, and it is important for tools to be developed that can assist policy analysts and social scientists. Some studies (Chang et al., 2009; Mimno et al., 2011; Newman et al., 2010) have focused on the evaluation of topic coherence compared to human judgment. In these studies, measures of topic quality are drawn using hundreds of topics. They tend to produce a large number of topics (≥ 200) because interpretation of topics is not the major goal of those studies. We purposefully chose a small number of topics to be feasible for understanding the content. When we produce a small number of topics (less than 50), topics tend to agglomerate, resulting in stable topics regardless of initialization. When the goal is interpretation of a corpus and the data size is small (thousands of documents), it makes more sense to produce a small number of topics from the outset rather than cherry picking a portion of good quality topics from a large number of topics produced, in which case topics will be extremely sensitive to initialization and sampling methods.

5. Conclusion

This paper provides LDA guidelines for computer assisted content analysis of e-petitions, which involves close human supervision throughout the training and evaluating process. We sought to automatically identify latent topics from WtP petitions. Our framework produced human-interpretable topics that closely corresponded to Pew's independent content analysis results. Our study demonstrates that LDA topics can (1) identify multi-faceted themes, (2) extract latent themes that may be overlooked, and (3) reflect themes in their raw state without bias.

Our findings have implications for a range of information professionals including social scientists and policy analysts who want to conduct content analysis with large volumes of text data. We provided a detailed guideline for training and evaluating LDA models using open-source tools for computer-assisted content analysis that is transparent and not dependent on a black-box approach. For computer scientists and software engineers, we highlighted the demands of social scientists and policy analysts, who are interested in interpreting a manageable number of topics. Social scientists can benefit from a well-engineered open source tool that enables human supervision of each step while making all the steps of training and evaluation transparent to avoid the black-box style approaches of commercial tools. Finally, for government policy decision makers, the procedures demonstrated here can help them conduct content analysis in a low cost and timely manner.

This article makes two major methodological contributions to content analysis. First, our guidelines for training and evaluating models were efforts to overcome one of the major problems of computer assisted content analysis using proprietary tools, which adopt “black-box” type approaches. We used open-source computer programs, and made the entire decision-making process transparent. Second, our LDA topics are validated from diverse perspectives—human supervision of automatic computer analysis, comparing with social events, and comparing with independent content analysis results. This supports efforts to establish LDA as a dependable content analysis methodology, rather than a mere algorithm.

Some limitations of the study should be mentioned. Our findings are based only on the WtP petitions that appeared on the website, which excludes petitions that failed to gather 150 signatures within the first 30 days. Although our LDA model performed well on a set of unknown data, our findings are only valid, at this point, for WtP petition data since we did not yet test our LDA framework on other datasets. The optimal K, for example, is very sensitive to writing style (one or two main arguments in each text) and the length of each text (average length of text was 98), as well as the number of texts used for training. In other words, our approach may not work on a corpus with texts that are too short (e.g. Twitter data) or include complex arguments (for example, texts about legal debates).

More future studies are needed validating LDA for content analysis to establish it as a dependable methodology. Although our procedure for topic modeling used e-petition data, an analogous procedure could be used to analyze different types of data. Testing LDA procedures on different data types and data sets—including, different snapshots of the same corpus, or different time intervals—would be needed to determine whether, and the extent to which, the approach is generalizable. In addition, variations of topic modeling such as structural topic models or dynamic topic modeling may increase efficiency of the content analysis by enabling topics' relationship to document metadata (Blei & Lafferty, 2006; Roberts et al., 2014).

Acknowledgments

I am grateful for support for this research from the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017S1A3A2066084). In addition, I would like to thank Teresa Harrison. Özlem Uzuner, S.S. Ravi, Will May, Xiaoyi Zhao, and Kurtis Hagen for their feedback and assistance on research that has been incorporated into this article. And, I especially thank Randy Borum for his valuable suggestions.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2018.05.006

References

Aggarwal, C. C., & Zhai, C. (2012). *A survey of text clustering algorithms. Mining text data*. Springer US77–128. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-3223-4_4.

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology, 68*(6), 1397–1410. https://doi.org/10.1002/asi.23786.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit.* Sebastopol, MA: O'Reilly Media.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). ACM. Retrieved from http://dl.acm.org/citation.cfm?id = 1143859.

Blei, D. M., & Lafferty, J. D. (2009). *Topic Models* Retrieved from https://www.cs.princeton.edu/ ~ blei/papers/BleiLafferty2009.pdf.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(4/5), 993–1022.

Bochel, C. (2012). Petitions: Different dimensions of voice and influence in the Scottish parliament and the national assembly for wales. *Social Policy & Administration, 46*(2), 142–160. https://doi.org/10.1111/j.1467-9515.2011.00828.x.

Bochel, C. (2013). Petitions Systems: Contributing to Representative Democracy? *Parliamentary Affairs, 66*(4), 798–815. https://doi.org/10.1093/pa/gss005.

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. In E. M. Airoldi, D. Blei, E. A. Erosheva, & S. E. Fienberg (Eds.). *Handbook of Mixed Membership Models and Their Applications* (pp. 225–255). Boca Raton, FL: CRC Press.

Bump, P (2016). Did a Donald Trump intern do this Retweet of '@WhiteGenocideTM?'. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/01/22/did-a-donald-trump-intern-do-this-retweet-of-whitegenocidetm/.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems* (pp. 288–296). . Retrieved from http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2009_0125.pdf.

Chuang, J., Wilkerson, J. D., Weiss, R., Tingley, D., Stewart, B. M., Roberts, M. E., et al. (2014). Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations. *Advances in neural information processing systems workshop on human-propelled machine learning.* Retrieved from http://www-dsp.rice.edu/files/Paper7.pdf.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20*(1), 37.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches.* SAGE.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice.* Boston, MA: Addison-Wesley.

Dalton, R. J. (2013). *Citizen politics. Public opinion and political parties in advanced industrial democracies*. Los Angeles: CQ Press.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2*(2), 2053951715602908 https://doi.org/10.1177/2053951715602908.

Dumas, C. L., LaManna, D., Harrison, T. M., Ravi, S. S., Kotfila, C., Gervais, N., et al. (2015). Examining political mobilization of online communities through e-petitioning behavior in We the People. *Big Data & Society, 2*(2), 2053951715598170 https://doi.org/10.1177/2053951715598170.

Gill, A. J., Hinrichs-Krapels, S., Blanke, T., Grant, J., Hedges, M., & Tanner, S. (2017). Insight workflow: Systematically combining human and computational methods to explore textual data. *Journal of the Association for Information Science and Technology, 68*(7), 1671–1686. https://doi.org/10.1002/asi.23767.

Girard, J. (2015). *Gwet's Agreement Coefficient (AC1/AC2) - File Exchange - MATLAB Central.* Retrieved February 14, 2016, from http://www.mathworks.com/matlabcentral/fileexchange/52361-gwet-s-agreement-coefficient–ac1-ac2-.

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today, 24*(2), 105–112.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis, 18*(1), 1–35.

Grimmer, J. (2016). *Measuring representational style in the house: The tea party, Obama and legislators' changing expressed priorities. Data analytics in social science, government, and industry*. Cambridge University Press.

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

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment, 1*(6), 1–6.

Hagen, L. (2016). *Topic modeling for e-petition analysis: Interpreting petitioners' policy priorities*. United StatesNew York: State University of New York at Albany.

Hagen, L., Uzuner, O., Kotfila, C., Harrison, T. M., & Lamanna, D. (2015). Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. *2015 48th Hawaii international conference on system sciences (HICSS)* (pp. 2134–2143). . https://doi.org/10.1109/HICSS.2015.257.

Hagen, L., Harrison, T. M., & Dumas, C. (2018). Data Analytics for policy informatics: the case of e-petitioning. Eds. In J. R. Gil-Garcia, T. Pardo, & L. Luna-Reyes (Eds.). *Policy analytics, modelling, and informatics: Innovative tools for solving complex social problems*. Springer.

Hagen, Loni, Harrison, T. M., Uzuner, Ö., May, W., Fake, T., & Katragadda, S. (2016). E-petition popularity: Do linguistic and semantic factors matter? *Government Information Quarterly, 33*(4), 783–795.

Hitlin, P. (2016). *"We the People": Five Years of Online Petitions.* Retrieved December 29, 2016, from http://www.pewinternet.org/2016/12/28/we-the-people-five-years-of-online-petitions/.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229–247.

Jelveh, Z., Kogut, B., & Naidu, S. (2015). *Political language in economics (SSRN scholarly paper No. ID 2535453).* Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2535453.

Kasperkevic, J. (2012). Texas secession petition grows on White House website. *Houston Chronicle*. Retrieved from http://www.chron.com/news/politics/article/Texas-secession-petition-grows-on-White-House-4030870.php.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18*(2), 243–250.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Krippendorff, K. H. (2012). *Content analysis: An introduction to its methodology* (Third edition). Los Angeles ; London: SAGE Publications, Inc.

Leydesdorff, L., & Nerghes, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora ($N < 1000$). *Journal of the Association for Information Science and Technology, 68*(4), 1024–1035. https://doi.org/10.1002/asi.23740.

Liu, Y., & Xu, S. (2017). A local context-aware LDA model for topic modeling in a document network. *Journal of the Association for Information Science and Technology, 68*(6), 1429–1448. https://doi.org/10.1002/asi.23822.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 4*, 587.

Lucas, C., Nielsen, R., Roberts, M., Stewart, B., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254–277.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (1 edition). Cambridge, Mass: The MIT Press.

McCallum, A. K. (2002). *MALLET: A machine learning for language toolkit.* Retrieved from http://mallet.cs.umass.edu/.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145462.

Nelson, L. K. (2014). *The power of place: Structure, culture, and continuities in U.S. women's movements*. Berkeley: University of California.

Neuendorf, K. A. (2017). *The content analysis guidebook* (Second edition). Los Angeles: SAGE.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 100–108). Association for Computational Linguistics.

Nikolakakou, N., Bothos, E., & Gregoris, M. (2015). Aggregating expectations to predict policy indices with information markets. *Electronic government and electronic participation: Joint proceedings of ongoing research and projects of IFIP WG 8.5 EGOV and ePart 2015* (pp. 63–71). . https://doi.org/10.3233/978-1-61499-570-8-63.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science, 43*(1), 88–102. https://doi.org/10.1177/0165551515617393.

Open Government Partnership (2015). What is the open government partnership? Retrieved May 16, 2015, from http://www.opengovpartnership.org/.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science, 54*(1), 209–228.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.

Roberts, M. E., Stewart, B., Tingley, D., & Airoldi, E. (2013). The structural topic model and applied social science. *Advances in neural information processing systems workshop on topic models: Computation, application, and evaluation*. Retrieved from http://scholar.harvard.edu/files/bstewart/files/stmnips2013.pdf.

Schiller, W. J. (2000). *Partners and rivals: Representation in U.S. senate delegations.* Princeton, NJ: Princeton University Press.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321–325.

Seshadri, K., Mercy Shalinie, S., & Kollengode, C. (2015). Design and evaluation of a parallel algorithm for inferring topic hierarchies. *Information Processing & Management, 51*(5), 662–676. https://doi.org/10.1016/j.ipm.2015.06.006.

Song, J., Huang, Y., Qi, X., Li, Y., Li, F., Fu, K., et al. (2016). Discovering hierarchical topic evolution in time-stamped documents. *Journal of the Association for Information Science and Technology, 67*(4), 915–927. https://doi.org/10.1002/asi.23439.

Steckler, A., & McLeroy, K. R. (2008). The importance of external validity. *American Journal of Public Health, 98*(1), 9–10. https://doi.org/10.2105/AJPH.2007.126847.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *General inquirer: A computer approach to content analysis.* The MIT Press.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464–2476. https://doi.org/10.1002/asi.23596.

The White House (2015). The open government partnership: third open government national action plan for the United States of America. Retrieved from http://www.whitehouse.gov/blog/2013/12/06/united-states-releases-its-second-open-government-national-action-plan.

The White House (2015). We the People, by the numbers. Retrieved October 25, 2015, from https://www.whitehouse.gov/share/we-people-numbers.

Trochim, W. M. K., & Donnelly, J. P. (2006). *The research methods knowledge base* (3rd Edition (3rd edition)). Mason, Ohio: Atomic Dog.

Viswanathan, M., & Berkman, N. D (2011). AC1 Statistic. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK82266/.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112). New York, NY: ACM. https://doi.org/10.1145/1553374.1553515.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 178–185). New York, NY: ACM. https://doi.org/10.1145/1148170.1148204.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann. Retrieved from http://www.mmds.org/.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*, 61. https://doi.org/10.1186/1471-2288-13-61.

Yan, E. (2015). Research dynamics, impact, and dissemination: A topic-level analysis. *Journal of the Association for Information Science and Technology, 66*(11), 2357–2372. https://doi.org/10.1002/asi.23324.

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics, 100*(3), 767–786. https://doi.org/10.1007/s11192-014-1321-8.

Zamith, R., & Lewis, S. C. (2015). Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The Annals of the American Academy of Political and Social Science, 659*(1), 307–318. https://doi.org/10.1177/0002716215570576.

Zhao, X., Jin, P., & Yue, L. (2015). Discovering topic time from web news. *Information Processing & Management, 51*(6), 869–890. https://doi.org/10.1016/j.ipm.2015.04.001.