SAND2022-10567C

Tag It: A Natural Language Processing Approach to Enhancing Delineation of Public Comments Regarding the Consent-Based Siting Process

Danielle N. Sanchez*, Kamille Hackett*, Matthew D. Sweitzer*, and Thushara Gunda*

*Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185, dnsanc@sandia.gov, khacket@sandia.gov, msweitz@sandia.gov, tgunda@sandia.gov

INTRODUCTION

Consent-based siting is an approach that considers public needs and concerns regarding the siting of nuclear waste management facilities [1]. This approach needs to be adaptable and flexible to help ensure that nuclear waste siting is carried out in a responsible manner that is sensitive to communities' needs. Engagement with the public is a key component of consent-based siting efforts.

One method that the U.S. Department of Energy (DOE) uses to engage the public is through a request for public comments [2]. In 2017, DOE released a request for public comments in which the public was asked relevant questions about a draft consent-based siting process for storage and disposal facilities [1]. Comments from the public were received in multiple ways, including email and through the regulations.gov website [1]; all received comments can be viewed on the DOE website [2].

Once a public comment period ends, human analysts review and bin received comments using a prescribed categorization guidance to inform the generation of responses. The process of binning comments can be resource-intensive due to the volume of comments collected and the need for consistency on the part of analysts. While no automated tool can replace the analyst, text-based analytical methods, such as natural language processing (NLP), have demonstrated successes that could complement existing approaches. NLP has been used, for example, to analyze frequent terms within research literature and identify patterns of keyword co-occurrence to inform priorities for stakeholder engagement [3]. Using linguistic tools – such as part-of-speech tagging, grouping of words, and sentiment analysis – NLP methods can analyze a body of text (i.e., corpus) to gain insights into patterns of content [4]. NLP has also been combined with machine learning techniques to detect specific types of comments [5].

Our analysis explores the potential for using NLP methods to aid analysis of public comments for consent-based siting. Specifically, we focus on using NLP for categorization of public comments in an efficient and consistent manner. To evaluate the value of NLP-based methods, we conduct a comparison with human-led comment processing and categorization to assess the degree of correspondence between the two methods. This analysis will be one of the first demonstrations of leveraging human-machine teaming to strengthen the comment review categorization in support of consent-based siting activities.

METHODS

This analysis evaluated 303 unique submissions from the public to the DOE's 2017 request regarding the consent-based siting process; duplicate versions of comments (i.e., form letters) were removed prior to analysis. These comments were read into the open source language, R (Software version 4.1.1) for processing and categorization of comments. Categorization was done using the structural topic modeling (STM) technique within R [6], which has been successfully implemented to understand narratives within unstructured text [7].

Comments were organized into tables within a Portable Document Format (or 'pdf') file; the table columns captured the raw comment as well as metadata associated with the comment (i.e., site of collection, analyst-coded labels, etc.). This data was processed into a machine-readable format using functions within the pdftools package [8]. Information about the location of the words on the pdf pages was used to extract details from individual columns of the table (i.e., x- and y-value coordinates served as the left/right and top/bottom bounding values, respectively). Once the comments were read in, the data was processed to remove special characters (e.g., "/") and converted all text to lowercase.

After processing, the STM algorithm was implemented to group comments into "topics" based on the frequent and exclusive (i.e., FREX) co-occurrence of words within the text. To determine the number of topics (i.e., k) that best categorize the comment data, a diagnostic procedure was run on 10 to 50 topics. The topic number with both high semantic coherence and exclusivity was then selected. Semantic coherence refers to the likelihood that each topic contains words that occur together in a document [9, 10], while exclusivity refers to the likelihood that the words in one topic are unlikely to occur as words in a different topic [10]. Topic assignments are not mutually exclusive for a given comment. In other words, a single comment can be assigned to multiple topics (e.g., 0.6 for topic X and 0.4 for topic Y) within STM; the proportions across topics sum to 1 for each comment.

With the selected k, STM was executed to group the comments into the associated number of topics. FREX words were then used to assign a label to each topic that captures the general subject matter. Sankey visualizations were used to compare the STM-generated topics to existing analyst-generated labels. Sankey visualizations enable comparison of different categories by mapping the proportion of one label (e.g., human-generated label) onto another label (e.g., STM-generated topic). The width of the bars connecting the two labels indicates general agreement between labels; wider bars indicate higher alignment. Sankey visualizations are particularly helpful for evaluating comparisons between the STM-vs-analyst labeling, since topic assignments are not mutually exclusive for a given comment (as noted above).

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

RESULTS & DISCUSSION

Figure 1 shows the semantic exclusivity for STM models with 28-30 topics. Diagnostics indicated that $k = 29$ would result in topics with both high semantic coherence and exclusivity. An execution of the STM model with $k = 29$ identified a number of interesting themes within the comments.

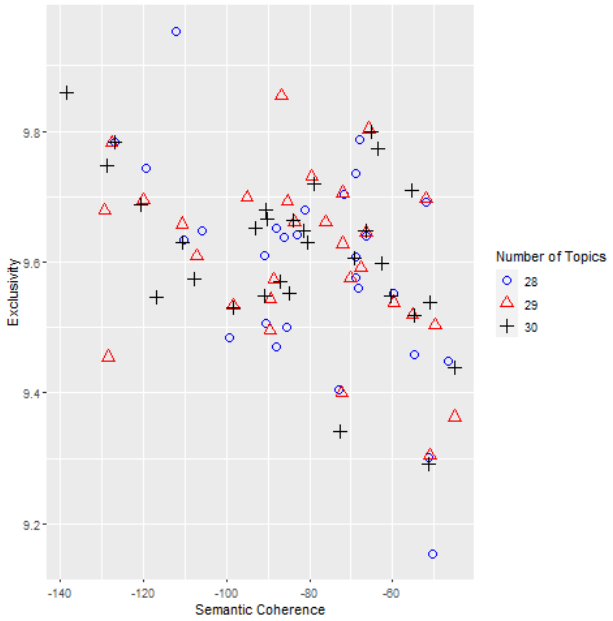

Fig. 1. Semantic Coherence and Exclusivity Models with $k = 28 - 30$ Topics.

The most prevalent topic in the corpus of comments, was topic 11 (Exit Process), which refers to a community's ability to withdraw from the consent-based process. This was followed by topic 5 (Texas Interim), the interim facility in Andrews, Texas and topic 21 (State Veto), which refers to a state's ability to veto the siting process. This may indicate that examples of existing nuclear waste facilities, such as the one in Andrews County, Texas helps foster community trust and gives the populace an idea of what to expect when hosting a facility in their region. Second, an option for a state veto and the ability to exit the consent-based process may help engender a greater sense of trust with the process, since communities will have more control regarding whether a nuclear waste site will be hosted. Exemplar comment excerpts from topics 11, 5, and 21 can be found in (Table I).

Surprisingly, safety and environmental concerns, including transportation of waste (topic 22), radiation fallout concerns (topic 28), adverse societal impacts (topic 9), and environmental impacts of a facility (topic 1) were the least prevalent topics, present in less than 5% of comments each, respectively (Figure 2). While safety and environmental concerns are still critical aspects to the siting process, these results indicate that a community's sense of control over the siting process and examples of successful facilities were more prevalent about the consent-based siting process during the 2017 comment

Topic No.	Comment Excerpt	Assigned Label
11	"various off ramps should be allowed at appropriate stages...people must be assured that should they change their minds, they do have a way to exit the siting process"	Exit Process
5	"on april 28, 2016, waste control specialists (wcs) submitted to the nrc a license application to construct and operate a consolidated interim storage facility at its 14,000 acre facility in andrews county, texas...wcs's history of safe disposal operations has been a significant factor in obtaining consent for the facility."	Texas Interim
21	"the nuclear waste policy act did address the issue of the role of the states in the decision making process. section 116(b)(2) of the nuclear waste policy act includes provisions for a notice of disapproval..."	State Veto

TABLE I. Representative Comments from Top 3 Topics.

period. Finally, the identification of multiple topics related to tribes (e.g., tribal land, tribal rights, and history preservation) indicates that there may be substantive contextual differences between each of these topics. Additional research is warranted to further explore the nuances identified in computational models and associated implications for CBS process development. Figure 2 illustrates the proportion of the corpus represented by each of the 29 topics, as well as the top 5 FREX words associated with each topic and the assigned label each topic was given by our team.

Next we compared the topic-model generated labels to the labels analysts supplied for each comment. The top three analyst labels were 1) 'Siting Process,' which include comments related to design elements of the siting process and community assistance and engagement (23% of 303 comments); 2) 'General Process Comments' (23%), which incorporates statements regarding the rationale for moving forward with the consent-based siting process; and 3) 'Siting Considerations,' which include general comments on siting considerations excluding geologic repositories (17%) [1]. The full distribution of analyst-derived labels mapped to STM-generated topics is shown in Figure 3.

Taking the analyst label of 'Siting Process' as a case study

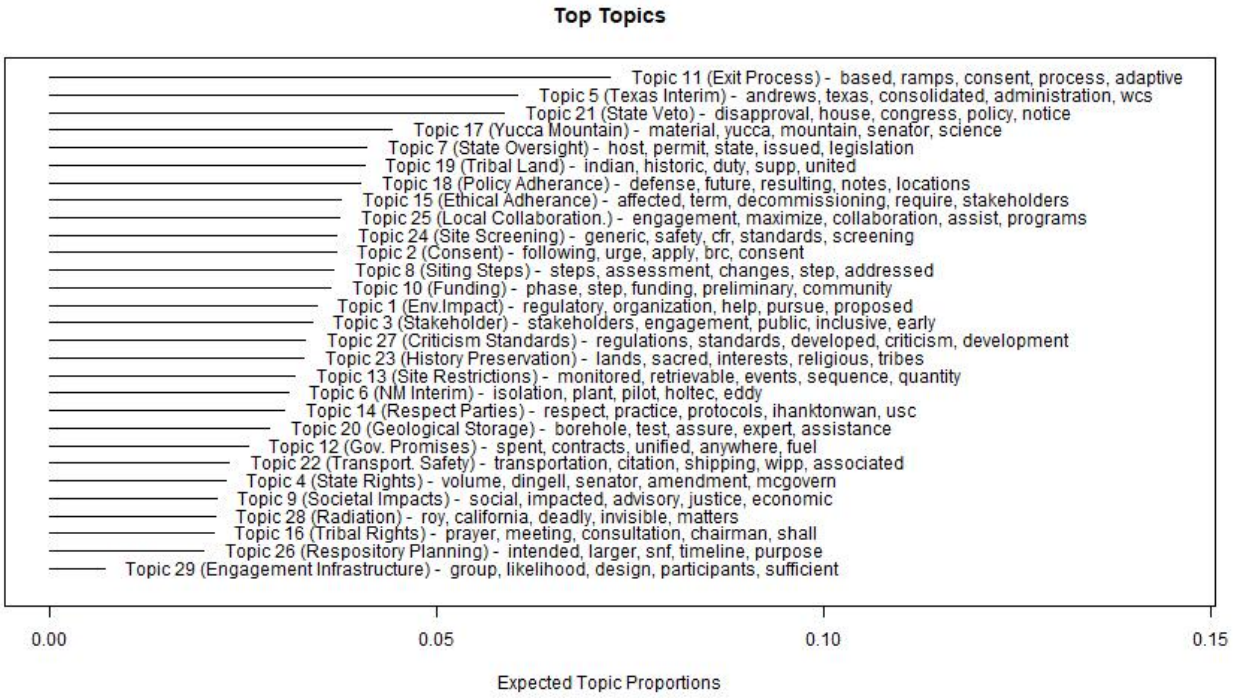

Fig. 2. Topic Prevalence for the $k = 29$ Comments. The length of the lines indicates the proportion of the corpus that corresponds to each topic. Topic labels are captured in parentheses.

(Figure 4), we see that two themes emerge within the topics, centered around: 1) community engagement and outreach and 2) design elements. First, community engagement and outreach is highly represented within this space - representing 30.4% (21 out of 69 comments) - and includes acquiring community consent (topic 2), working with local programs and stakeholders (topics 25 & 3), and community funding (topic 10). Additionally, the topic model also illuminated that the ability to exit the siting process (topic 11) and the ability to veto a site by a state (topics 4 & 21) may be an important consideration in the context of community relations; 16% of comments aligned with these topics (Figure 4). This may indicate that clear stopping points are an important part of community considerations regarding the nuclear consent-based siting process.

Second, the 'Siting Process' definition of design elements and implementation steps are represented within the STM topics. Patterns within the latter can be generally grouped around two sub-themes: 1) clarity of the process and 2) design elements. The first sub-theme is regarding clarity in the siting process, including the need for clear site development steps (topic 8), the need for clear planning steps and timelines (topic 26), and the impacts of perceived changing regulation standards (topic 27) - this sub-theme represented nine out of the 69 Siting Process comments (13%). The second sub-theme within design elements was regarding site consideration criteria, with state permits and oversight (topic 7), site screening criteria (topic 24), and geological testing (topic 20) representing 10 out of 69 comments (14.5%; Figure 3).

Additionally, comments related to safety were also present within the dataset. For example, the potential for adverse societal impacts (topic 9), radiation fallout concerns (topic 28), and safe transportation of waste (topic 22) represented 10 out of 69 comments (14.5%). Although safety-related comments were relatively infrequent within the corpus (Figure 2), safety is a long-standing community-level concern when considering nuclear waste siting due to the pervasive imagery in popular media of the negative consequences of radiation contamination [11]. Therefore, having a clear outline of steps and procedures to help mitigate safety concerns are an influential aspect of the siting process.

Finally, it appears that previous examples of nuclear waste siting facilities were also present within comments regarding the Siting Process. For example, the Yucca Mountain repository; the Holtec Facility in Eddy-Lea County, New Mexico; and the Waste Control Specialists (WCS) facility in Andrews County, Texas represented 4% (3 out of 69) of comments in the Siting Process category. This implies that communities look to historical examples to inform their opinions regarding community engagement and design elements of the siting process.

CONCLUSION

This paper illustrates how NLP and associated text analytic methods can help analyze public comments received on a nuclear waste consent-based siting process to inform un-

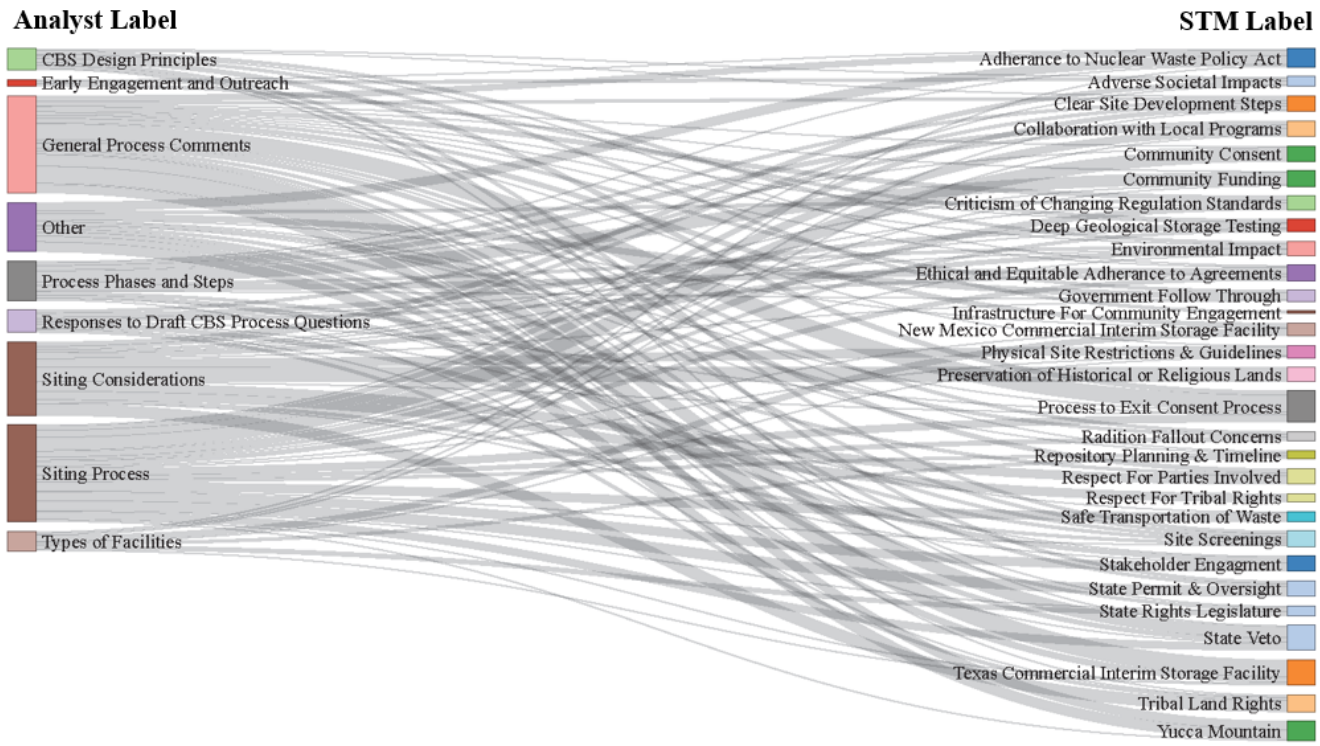

Fig. 3. Sankey Visualization of Analyst (left) to STM-Generated (right) Labels. The width of the bands represent the number of comments belonging to each Analyst-STM label pair.

derstanding of public perceptions and priorities. The STM trained as part of this work revealed 29 topics within the comments, whose subject matter ranged from having the ability to exit the consent-based siting process to infrastructure needed to support community engagement. These insights highlight key nuances in public considerations for nuclear waste siting activities.

Beyond unpacking nuances present in the public comments, topic modeling methods also allowed us to compare alignment with human-generated groupings of comments. The case study comparison of STM-generated labels with the 'Siting Process' analyst label found high thematic correspondence with DOE's existing definition of the siting process. The STM-generated labels, however, illuminated additional nuances present in the comments that may warrant being included in future definitions (e.g., safety and existing siting facilities).

Ongoing analyses of these comments are needed to examine how topics within the STM may relate to one another through correlation analysis or distance-related metrics. Correlation analysis may indicate the strength and valence of relationships between topics (i.e., topic co-occurrence within comments) that can be helpful for understanding nuances within text (e.g., between tribal land vs. tribal rights). Distance metrics, on the other hand, can be used to reveal how similar or dissimilar topics are to one another [12, 13]. In addition, sentiment analyses could also be implemented to examine the tone of the comments and gain further insights into public opinions regarding nuclear waste siting. This analysis, for example, could indicate whether communities more favorably view consent-based siting when veto-like processes are present.

Future work could also examine additional years of public comments to investigate the generalizability of these topics over time (e.g., 2021). In addition, we will be able to elucidate if certain topic themes have become more prevalent or if new themes have emerged compared to the 2017 corpus of comments. Finally, insights from comments could be extended to compare with insights generated from social media and newspaper sources. Such comparisons could generate a more complete understanding of public discourse varies across communication platforms to further inform consent-based siting process activities.

ACKNOWLEDGMENTS

This work is funded by the Office of Nuclear Energy at U.S. Department of Energy. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly-owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

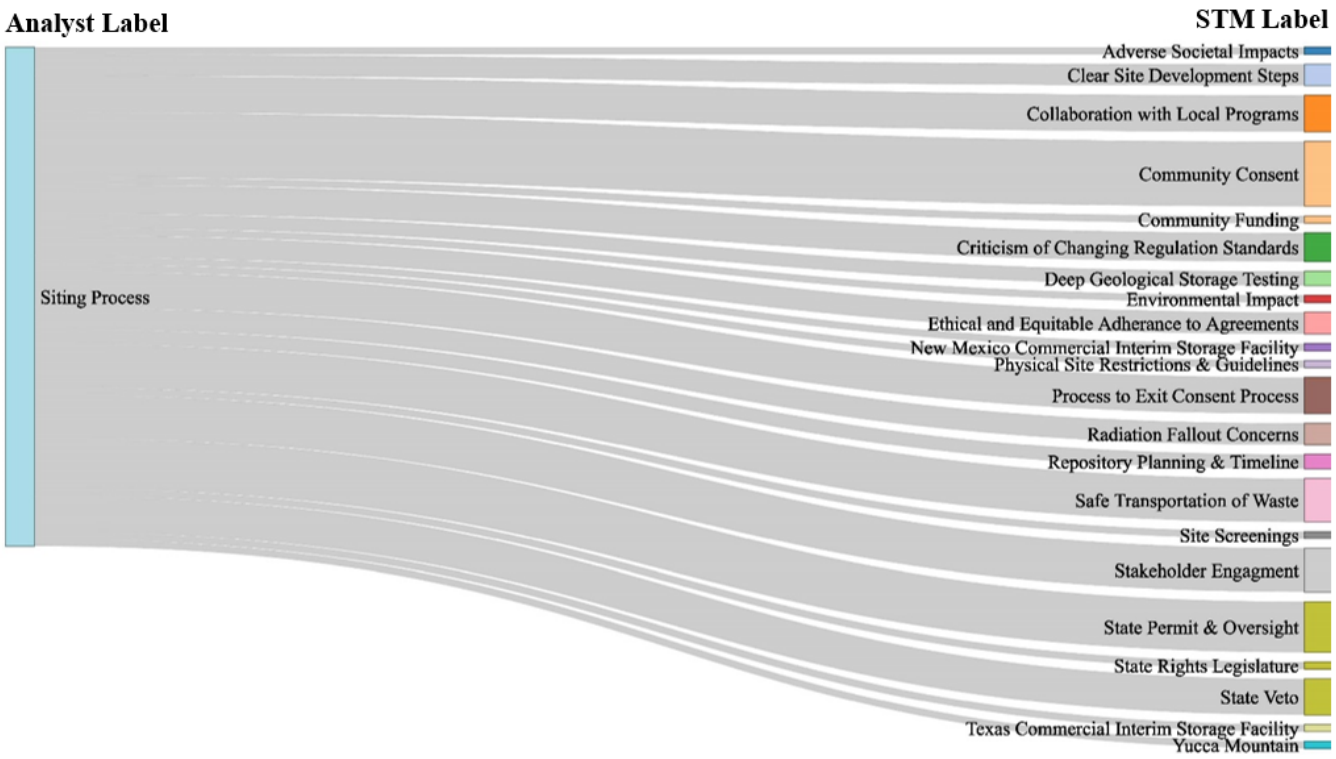

Fig. 4. Sankey Visualization of the Siting Process Labels. On the left is the analyst-generated label 'Siting Process' while the right represents topics generated from STM analysis. The width of the bands represent the number of comments assigned to each STM topic.

REFERENCES

1. U.S. DEPARTMENT OF ENERGY, "Draft Consent-Based Siting Process for Consolidated Storage and Disposal Facilities for Spent Nuclear Fuel and High-Level Radioactive Waste," https://www.energy.gov/sites/prod/files/2017/01/f34/Draft%20Consent-Based%20Siting%20Process%20and%20Siting%20Considerations.pdf (2017).
2. U.S. DEPARTMENT OF ENERGY, "Consent-based Siting," https://www.energy.gov/ne/consent-based-siting (2021).
3. S. R. SHAMS, D. VRONTIS, R. CHAUDHURI, G. CHAVAN, and M. R. CZINKOTA, "Stakeholder engagement for innovation management and entrepreneurial development: A meta-analysis," *Journal of Business Research*, **119**, 67–86 (2020).
4. F. S. GHAREHCHOPOGH and Z. A. KHALIFELU, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in "2011 5th International Conference on Application of Information and Communication Technologies (AICT)," (2011), pp. 1–4.
5. H. K. SHARMA, K. KSHITIZ, ET AL., "Nlp and machine learning techniques for detecting insulting comments on social networking platforms," in "2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)," IEEE (2018), pp. 265–272.
6. M. E. ROBERTS, B. M. STEWART, and D. TINGLEY, "Stm: An R package for structural topic models," *Journal of Statistical Software*, **91**, 1–40 (2019).
7. T. GUNDA, M. D. SWEITZER, K. T. COMER, C. FINN, S. MURILLO-SANDOVAL, and J. HUFF, "Evolution of Water Narratives in Local US Newspapers: A Case Study of Utah and Georgia." Tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (2018).
8. J. OOMS, *pdftools: Text Extraction, Rendering and Converting of PDF Documents* (2022), r package version 3.1.1.
9. D. MIMNO, H. WALLACH, E. TALLEY, M. LEENDERS, and A. MCCALLUM, "Optimizing semantic coherence in topic models," in "Proceedings of the 2011 conference on empirical methods in natural language processing," (2011), pp. 262–272.
10. M. E. ROBERTS, B. M. STEWART, D. TINGLEY, C. LUCAS, J. LEDER-LUIS, S. K. GADARIAN, B. ALBERTSON, and D. G. RAND, "Structural topic models for open-ended survey responses," *American journal of political science*, **58**, *4*, 1064–1082 (2014).
11. P. SLOVIC, J. H. FLYNN, and M. LAYMAN, "Perceived risk, trust, and the politics of nuclear waste," *Science*, **254**, *5038*, 1603–1607 (1991).

12. K. STEVENS, P. KEGELMEYER, D. ANDRZEJEWSKI, and D. BUTTLER, "Exploring topic coherence over many models and many topics," in "Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning," (2012), pp. 952–961.
13. T. GAO, B. CHENG, J. CHEN, and M. CHEN, "Enhancing collaborative filtering via topic model integrated uniform Euclidean distance," *China Communications*, **14**, *11*, 48–58 (2017).