SAND2022-7877C

A Natural Language Processing Tool to Support Stakeholder Engagements During a Consent-Based Siting Process

Thushara Gunda[*] and Matthew D. Sweitzer[†]

[*]*Sandia National Laboratories, Albuquerque, NM, tgunda@sandia.gov*
[†]*Sandia National Laboratories, Albuquerque, NM, msweitz@sandia.gov*

DOI: [To be inserted by ANS]

INTRODUCTION

Nuclear energy is one of the leading sources of low-carbon electricity across the world, providing up to 10% of the global electricity supply in 2018 [1]. Within the United States, the nuclear fleet generates approximately 20% of the nation's annual electricity [2]. However, the nuclear sector faces multiple challenges, including an aging fleet and management of waste (i.e., spent nuclear fuel) associated with the electricity generation process. Currently, nuclear waste is stored in temporary containers on-site of the nuclear plants – initially under water and then eventually moved into dry storage. However, it is widely recognized that deep geologic disposal would be the best long-term solution for these waste products [3].

Past siting attempts demonstrated that siting of nuclear waste management facilities is not just a question of technical suitability, but also associated public perceptions of risk and trust [4]. In recognition of the importance of the public acceptance, the Department of Energy – the management and implementation authority for nuclear waste in the United States – introduced the consent-based siting process [5]. This process recognizes that a community-driven perspective is required to ensure equitable and just management of nuclear waste. In particular, the process recognizes the close collaboration required between the government officials (across all levels) as well as the public and interested groups [6].

Requests for Information (in 2017 and 2021) have helped DOE identify a number of relevant organizations and the level of awareness of communities about nuclear waste management [7]. However, understanding specific concerns and priorities that a particular community may have with nuclear waste siting is challenging from such broad sampling methods. In an attempt to address this issue, this work aims to evaluate the potential for analyzing local journalism coverage using natural language techniques to understand community narratives associated with nuclear waste.

Natural language processing (NLP) falls into a broader category of computational methods developed in the social sciences, which can help identify valuable information from societal discourse [8]. NLP techniques have been successfully used in multiple domains to understand patterns within text. For example, network-based methods have been combined with text data to understand diversity of discourse as well as potential gaps and biases within the text [9]. Researchers have also used NLP to detect "fake" (or false) news reports and distinguish them from factual articles [10]. More recently, researchers have also used NLP on newspaper articles to compare and contrast narratives around resources within regional newspapers [11].

Using established techniques such as sentiment analysis, content analysis, and entity extraction, our aim is to evaluate spatial and temporal patterns in the public discourse, as well as identify the attitudes and entities engaged in discussions about nuclear waste issues. Findings from our work provide insight into whether these analysis methods can be used in real-time to shape an engagement strategy by incorporating community priorities and concerns into ongoing dialogues and enable mutual learning to inform siting-related decision making.

METHODS

Data for this study were collected from DataNews [12]. DataNews is a news content aggregation website that enables users to download articles through an application programming interface (API). Users can query the database to understand different sources, headlines of recent publications, as well as a 5-year archive of news articles. The query capabilities enable users to identify content of interest through keyword searches, geographic and language restrictions, and sorting (e.g., by relevance, date of publication, etc.). For this analysis, the DataNews "News" API was queried using the keyword 'nuclear' (case-insensitive), a United States source restriction, and reverse-ordered date sorting. Once duplicate entries were removed, we had a corpus (i.e., collection of text documents) that contained approximately 150,000 articles from 5,000 news sources. All data collection, cleaning, and analyses were performed using open-source software R, version 4.1.3 [13].

Figure 1 depicts the procedures we followed after data collection. Specifically, we implemented two cleaning procedures: 1) date restrictions and 2) removal of 'stop words'. Although DataNews is supposed to be a 5-year archive, there were a few articles in the corpus that dated back further. So for date restrictions, we explicitly excluded any articles that were published greater than 5 years prior to the data collection date; this resulted in a corpus that ranged in publication date from February 1, 2017 to January 31, 2022. Additionally, 'stop words' (e.g., "the", "an", "of", etc.) were removed from the corpus, since they convey little meaning in text and can influence similarity assessments of documents. After data cleaning procedures, the final corpus contained $n = 148,322$ articles from $4,813$ unique sources. These articles had $22,785$ unique bylines (i.e., author) of the article.

Natural Language Processing Techniques

Multiple computational techniques from NLP were used to analyze the corpus. These tools leverage features of the written English language, such as the co-occurrence of words across documents or the capitalization of proper nouns, to identify key features of texts. Specifically, we utilized three

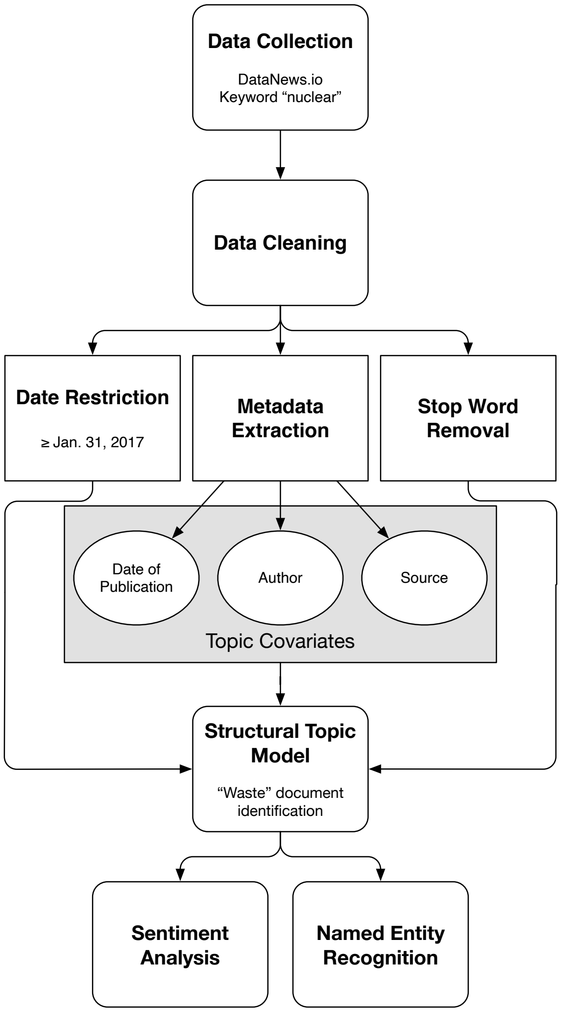

Fig. 1. Path diagram of data collection, cleaning procedure, and analysis strategy.

NLP methods: 1) a topic model, 2) entity recognition, and 3) sentiment analysis.

Topic modeling was used to identify the subset of news articles which are focused on 'waste'-related issues. Topic modeling is a method that bins documents into groups pertaining to similar subject matter. The technique typically involves parsing a matrix of documents and the tokens, or unique words, included in them to identify common terms within a group of documents and to separate different groups. We estimated a structural topic model using the *stm* R package [14] for this analysis, which implements FREX words – or words ordered by the harmonic mean of their FRequency within a group of documents and relative EXclusivity to that group – to help analysts identify the subject matter of a topic. Structural topic models differ from other topic modeling methods, such as Latent Dirichlet Allocation, in that they allow for the specification of document metadata as topic covariates [15].

For this analysis, date of publication, author, and source were used as covariates for the model. Typically, k, or the number of topics one would like the model to identify is specified by the analyst *a priori*. However, because we had no expectation of the number of groups of documents we might uncover, we opted to utilize the built-in algorithm for topic number identification within the *stm* package [16]. The model produced 71 topics of varying size and topic quality (figure not included) that were further evaluated to identify waste-related topic(s). The stm generates topic fit scores for each document, which reflects the proportion of the article that corresponds to each topic. For the remaining analyses, we subset the corpus to those articles for which a waste-related topic was a plurality of the document proportion.

After subsetting to waste-related topics, we implemented named entity recognition to identify key entities (such as organizations, persons, etc.) that are referred to in these articles. Named entity recognition was performed using the R package *spacyr* [17]. This software uses a language model that was pre-trained on text documents such as blog posts, news articles, and internet comments to tag parts of speech and identify whether proper nouns belong to persons, organizations, geopolitical identities, and more. Finally, sentiment analysis was implemented to evaluate affective language (i.e., moods, feelings, and attitudes) within the waste-related articles. Sentiment analysis was performed using the R package *sentimentr* [18]. This package uses a built-in dictionary of valenced words (e.g., negative: "unfavorable", "unsafe", etc.; positive: "favorable", "harmless", etc.) to quantify the sentiment contained in each sentence. Moreover, the algorithm adjusts sentiment values in accordance with the use of language-modifying words like "not" or "very". Sentiment values of each sentence were then summarized (for mean, M, and standard deviation, SD) to evaluate an article's overall stance on nuclear waste. Univariate t-tests were used to evaluate the significance of sentiment scores from zero (or neutral) scores for $n - 1$, where n is the sample size. Given the agenda setting functions of news media [19, 20], these insights could serve as a proxy for public opinions as expressed within newspapers.

RESULTS & DISCUSSION

After evaluating the FREX words for the $k = 71$ stm, we identified topic #70 grouped a number of terms that are most relevant to nuclear waste siting issues. For example, the term "waste" appears among the top 10 FREX words for this topic; additional frequent words include "Hanford", "environmental", "Holtec", and "cleanup". This topic comprised approximately 2% of the corpus downloaded from DataNews and ranked 23rd in prevalence compared to the other 70 topics. The evolution of topic 70 coverage over time is shown in Figure 2. Coverage of nuclear waste issues ebbed and flowed over the last 5 years, but reached a peak in late 2021. The large confidence interval in the period from 2017-2019 likely reflects the relatively few articles that were published at that time, which could reflect either sparse coverage of the topic within the news or limitations of the dataset itself. Subsetting the corpus to only contain articles that predominantly discussed nuclear waste resulted in a collection of $n = 2,526$ articles from 372 sources and 527 unique bylines that were further analyzed for key entities and sentiments.

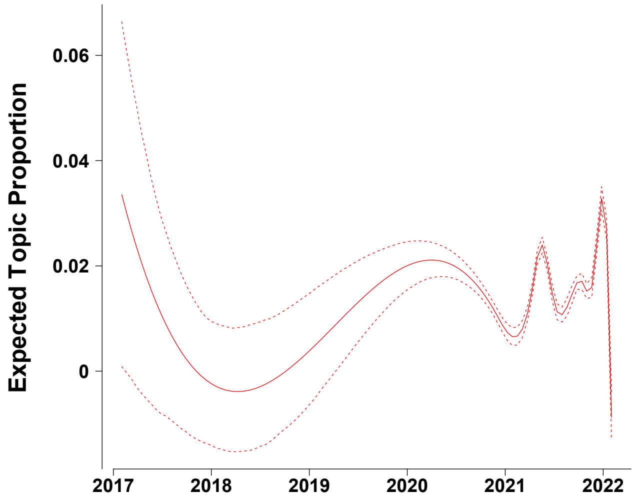

Fig. 2. Prevalence of topic 70 over time. The solid line represents the estimated proportion of all news articles that pertain to topic 70, while the dashed lines indicate the 95% confidence interval of that estimate.

The named entities identified most frequently within nuclear waste articles highlight specific geographies (e.g., United States, Navajo Nation, and New Mexico) (Figure 3). These could reflect either locations near nuclear-related facilities, or entities that are involved with nuclear waste siting decisions. Additionally, specific organizations also emerge, such as the Nuclear Regulatory Commission, the U.S. Department of Energy, Congress, national labs, and private companies that are involved in the nuclear waste siting process. A couple of key persons - namely recent U.S. presidents whose administrations shape regulatory policy - also emerged in these articles; additional evaluation is needed to understand regional and local individuals prevalent in nuclear waste siting discussions. Further evaluation is also needed to understand frequent references to historical events (such as World War II and the Cold War) in the context of nuclear waste. Future work will consider using covariance analysis between Topic 70 and the remaining topics to gain insights into how different nuclear topics are discussed in concert with nuclear waste issues.

Sentiment analysis revealed that the authors of nuclear waste articles tended to favor the use of negative valenced words compared to positive ones ($M = -0.030$, $SD = 0.123$). Despite the large variance, a univariate t-test revealed that the distribution of sentiment scores did significantly differ from zero (no sentiment valence); $t(2,525) = -12.12$, $p < .001$. The distribution of average sentence sentiments for each article is shown in Figure 4. The larger tail on the left (relative to the right) reflects the general tendency of news media to use negative valenced words more often. However, the vast majority of nuclear waste articles use generally neutral language (i.e., fewer than one affective word for every 5 sentences). Further, the range of average sentiment values was constrained – range: $[-0.636, 0.524]$ – indicating that even the most affective articles about nuclear waste tend to limit valenced words. Future work could evaluate use of deep learning techniques and aspect-based sentiment analysis to better capture the specific entity or target of the sentiments as well as spatiotemporal variations in these patterns [21].

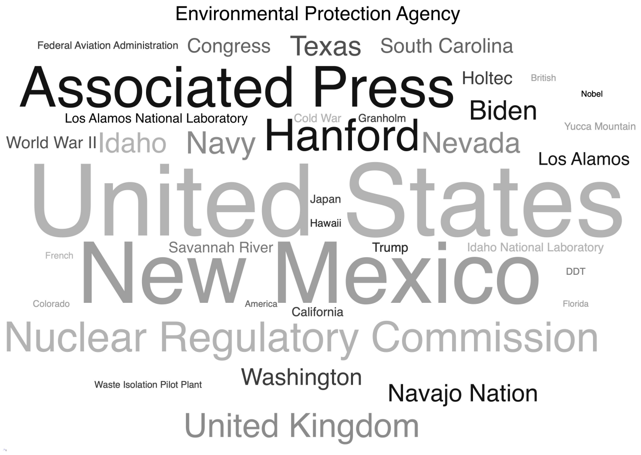

Fig. 3. Wordcloud of the 50 most frequently occurring named entities among nuclear waste articles. Larger words appear more frequently in text than smaller ones. *Note*: words or phrases which refer to the same entity (e.g., "U.S." and "United States") were combined.

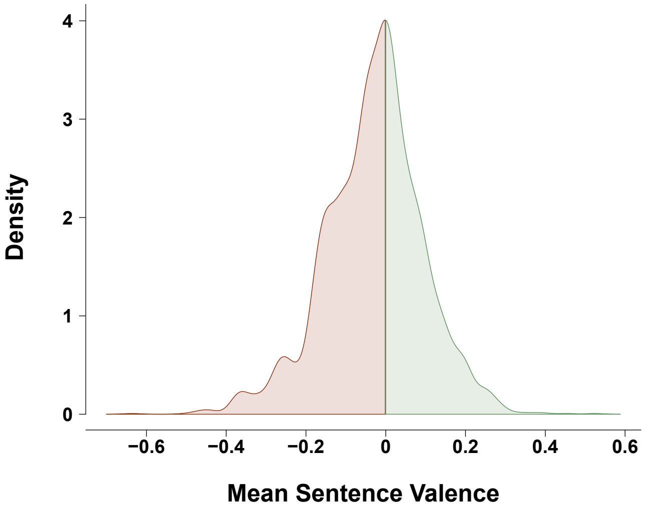

Fig. 4. Kernel density plot of the mean sentence sentiment among nuclear waste articles. Red indicates that more negatively valenced words were used in the article, while green indicates that the article expressed more positive sentiments.

CONCLUSION

This analysis demonstrates the significant potential for NLP techniques to gain insights into narratives around nuclear waste. Specifically, we highlighted that the general discourse regarding nuclear waste within the news media has been exhibiting a cyclical pattern with common entities reflecting a number of geographies and organizations. General sentiments within the nuclear waste articles appear to use neutral language, however the exact nuances need to be further evaluated. Use of deep learning techniques and covariance analysis will be explored in future analyses to better understand relationships between nuclear waste and other nuclear topics, sentiments of specific entities, and patterns across space and time (including in a particular region). These data-driven methods can complement and inform ongoing discussions about effective stakeholder engagement during consent-based siting for nuclear waste.

ACKNOWLEDGMENTS

This work is funded by the Office of Nuclear Energy at 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.

REFERENCES

1. INTERNATIONAL ENERGY AGENCY, "Nuclear Power in a Clean Energy System," `https://www.iea.org/reports/nuclear-power-in-a-clean-energy-system` (2019).
2. U.S. ENERGY INFORMATION ADMINISTRATION, "Nuclear explained," `https://www.eia.gov/energyexplained/nuclear/data-and-statistics.php` (2021).
3. R. C. EWING, R. A. WHITTLESTON, and B. W. YARDLEY, "Geological disposal of nuclear waste: A primer," *Elements*, **12**, *4*, 233–237 (2016).
4. P. SLOVIC, M. LAYMAN, and J. H. FLYNN, "Risk perception, trust, and nuclear waste: Lessons from Yucca Mountain," *Environment: Science and Policy for Sustainable Development*, **33**, *3*, 6–30 (1991).
5. P. LYONS, "Strategy for the Management and Disposal of Used Nuclear Fuel and High-Level Radioactive Waste," in "Presentation from the 2013 NARUC Winter Committee Meeting [2013-03-15]. http://www. naruc. org/meeting presentations. cfm," (2013), vol. 94.
6. U.S. DEPARTMENT OF ENERGY – OFFICE OF NUCLEAR ENERGY, "Consent-based siting," `https://www.energy.gov/ne/consent-based-siting` (*n.d.*).
7. U.S. DEPARTMENT OF ENERGY – OFFICE OF NUCLEAR ENERGY, "Responses to the Request for Information on Using a Consent-Based Siting Process to Identify Federal Interim Storage Facilities," `https://www.energy.gov/sites/default/files/2022-03/Responses%20to%20RFI%20on%20Consent-Based%20Siting%20and%20Interim%20Storage.pdf` (2022).
8. D. LAZER, A. PENTLAND, L. ADAMIC, S. ARAL, A.-L. BARABÁSI, D. BREWER, N. CHRISTAKIS, N. CONTRACTOR, J. FOWLER, M. GUTMANN, ET AL., "Computational social science," *Science*, **323**, *5915*, 721–723 (2009).
9. D. PARANYUSHKIN, "InfraNodus: Generating insight using text network analysis," in "The world wide web conference," (2019), pp. 3584–3589.
10. R. OSHIKAWA, J. QIAN, and W. Y. WANG, "A survey on natural language processing for fake news detection," *arXiv preprint arXiv:1811.00770* (2018).
11. 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).
12. "DataNews News API," `https://datanews.io`.
13. R CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2022).
14. M. E. ROBERTS, B. M. STEWART, and D. TINGLEY, "stm: An R Package for Structural Topic Models," *Journal of Statistical Software*, **91**, *2*, 1–40 (2019).
15. M. E. ROBERTS, B. M. STEWART, D. TINGLEY, and E. M. AIROLDI, "The structural topic model and applied social science," in "Advances in neural information processing systems workshop on topic models: computation, application, and evaluation," (2013), vol. 4, pp. 1–20.
16. D. MIMNO and M. LEE, "Low-dimensional embeddings for interpretable anchor-based topic inference," in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)," (2014), pp. 1319–1328.
17. K. BENOIT and A. MATSUO, *spacyr: Wrapper to the 'spaCy' 'NLP' Library* (2020), r package version 1.2.1.
18. T. W. RINKER, *sentimentr: Calculate Text Polarity Sentiment*, Buffalo, New York (2021), version 2.9.0.
19. M. E. MCCOMBS, D. L. SHAW, and D. H. WEAVER, "New directions in agenda-setting theory and research," *Mass communication and society*, **17**, *6*, 781–802 (2014).
20. W. RUSSELL NEUMAN, L. GUGGENHEIM, S. A. MO JANG, and S. Y. BAE, "The dynamics of public attention: Agenda-setting theory meets big data," *Journal of Communication*, **64**, *2*, 193–214 (2014).
21. H. H. DO, P. PRASAD, A. MAAG, and A. ALSADOON, "Deep learning for aspect-based sentiment analysis: a comparative review," *Expert systems with applications*, **118**, 272–299 (2019).