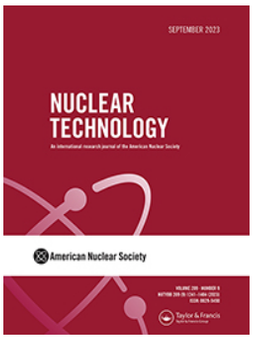


Nuclear Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/unct20

Spatiotemporal Analyses of News Media Coverage on "Nuclear Waste": A Natural Language Processing Approach

Matthew D. Sweitzer & Thushara Gunda

To cite this article: Matthew D. Sweitzer & Thushara Gunda (2023): Spatiotemporal Analyses of News Media Coverage on "Nuclear Waste": A Natural Language Processing Approach, Nuclear Technology, DOI: 10.1080/00295450.2023.2229566

To link to this article: https://doi.org/10.1080/00295450.2023.2229566

Published online: 02 Aug 2023.

NUCLEAR TECHNOLOGY

DOI: https://doi.org/10.1080/00295450.2023.2229566

Spatiotemporal Analyses of News Media Coverage on "Nuclear Waste": A Natural Language Processing Approach

Matthew D. Sweitzer* and Thushara Gunda

Sandia National Laboratories, Albuquerque, New Mexico 87185

Received April 3, 2023
Accepted for Publication June 10, 2023

Abstract — *The siting of nuclear waste is a process that requires consideration of concerns of the public. This report demonstrates the significant potential for natural language processing techniques to gain insights into public narratives around "nuclear waste." Specifically, the report highlights that the general discourse regarding "nuclear waste" within the news media has fluctuated in prevalence compared to "nuclear" topics broadly over recent years, with commonly mentioned entities reflecting a limited variety of geographies and stakeholders. General sentiments within the "nuclear waste" articles appear to use neutral language, suggesting that a scientific or "facts-only" framing of "waste"-related issues dominates coverage; however, the exact nuances should be further evaluated. The implications of a number of these insights about how nuclear waste is framed in traditional media (e.g., regarding emerging technologies, historical events, and specific organizations) are discussed. This report lays the groundwork for larger, more systematic research using, for example, transformer-based techniques and covariance analysis to better understand relationships among "nuclear waste" and other nuclear topics, sentiments of specific entities, and patterns across space and time (including in a particular region). By identifying priorities and knowledge needs, these data-driven methods can complement and inform engagement strategies that promote dialogue and mutual learning regarding nuclear waste.*

Keywords — *Natural language processing, nuclear waste, structural topic modeling, named entity recognition, sentiment analysis.*

Note — *Some figures may be in color only in the electronic version.*

I. INTRODUCTION

Nuclear energy is one of the leading sources of low-carbon electricity across the world. It provided up to 10% of the global electricity supply in 2018.[1] Within the United States, the existing nuclear power fleet generates approximately 20% of the nation's annual electricity.[2] Nuclear energy is also emerging as a key player for nations' climate goals, with some estimating the need to double power generation by 2050.[3] In the United States, nearly all of the nation's commercial spent nuclear fuel is currently stored at the reactor sites where it was generated, either submerged in pools of water (wet storage)[4] or in shielded casks (dry storage).[5] For the foreseeable future, the U.S. Nuclear Regulatory Commission has determined that the spent fuel can continue to be safely stored in licensed facilities.

Past siting attempts in the United States for consolidated storage facilities and a geologic repository for commercial spent nuclear fuel have demonstrated that

*E-mail: msweitz@sandia.gov

the siting of consolidated commercial spent nuclear fuel management storage facilities is a question not only of technical suitability but also of the associated public acceptance.[6] The U.S. Department of Energy (DOE) is responsible for the disposal of commercial spent nuclear fuel and high-level radioactive waste in the United States pursuant to the Nuclear Waste Policy Act of 1982, as amended. Congress has provided appropriations and directed the DOE to establish a program or more federal consolidated interim storage facilities for commercial spent nuclear fuel using a consent-based siting process.[7] Consent-based siting necessitates prioritizing the participation and needs of people and communities and centers equity and justice considerations.[8] The study that is the subject of this report considered "nuclear waste" broadly[a] in order to situate and understand public narratives in news media coverage of nuclear waste.

The study described in this report and the techniques presented here leverage text analytics-based methods to understand public narratives around "nuclear waste." Specifically, by leveraging recent advances in computational social-scientific methods, such as natural language processing (NLP), this report elucidates the many varied ways that "nuclear waste" is discussed in a large (approximately 150 000) corpus of news articles about nuclear topics (broadly construed). This dataset includes a variety of publication sources with both local and national reach across many different regions of the country and over time.

Natural language processing is a broad class of computational research methods that lies at the intersection of artificial intelligence, computer science, and linguistics. NLP has increased in popularity in recent years, particularly as the internet has made more text-based data available, to generate insights into individual preferences and societal discourse.[9] There are several different types of NLP analyses that can help an analyst or researcher understand themes, entities (such as people, places, or organizations), or other patterns within the text (e.g., sentiments conveyed). NLP has been used to successfully extend our understanding from a collection of words to associated narratives.[10] For example, network-based methods, such as community detection performed on the structure of topic co-occurrence across documents, have been combined with text data to understand the diversity of discourse as well as potential gaps and biases within the text.[11,12] The combination of topic modeling and sentiment analysis has further provided insights into specific narrative frames and flow within texts, for example, in analyses of literature as well as traffic incident reports.[13,14]

Researchers have also used NLP to detect "fake," or false, news reports and distinguish them from factual articles.[15] Finally, the computational efficiency of NLP has also enabled researchers to apply theories and compare and contrast narratives within regional newspapers across relatively large datasets.[16,17] Although many of these evaluations can be done by a team of human analysts, NLP methods can greatly expedite these processes and allow researchers to reliably analyze and summarize the content within large amounts of text data.[18] Importantly, when paired with qualitative interpretation, NLP can make obscure, abstract, or complex and interdependent relationships within and between texts (and between text and metadata) more discernible than a qualitative analysis technique, such as close reading (e.g., Ref. [19]) or conventional content analysis (e.g., Refs. [20] and [21]) on its own. While many of the applications and potential findings of NLP techniques overlap those of qualitative analytic approaches, quantitative techniques such as NLP can leverage more fine-grained detail at larger scales to generate novel insights as well. In this way, NLP can be thought of as an augmentation of or supplement to, rather than replacement of, qualitative analysis.

This report evaluates the utility of NLP for generating insights into public narratives around "nuclear waste." Using established techniques such as topic modeling, entity extraction, and sentiment analysis, our aim is to evaluate spatial and temporal patterns in the public discourse about "nuclear waste" issues. The paper is organized as follows. Section II provides greater details about the NLP methods utilized for this study. Section III steps through the results of the full corpus topic analysis, named entity recognition (NER) and sentiment analyses performed on the "waste"-related subset, and the secondary subtopic analyses. The authors conclude in Sec. IV with a discussion of the relevance of this work to ongoing consent-based siting efforts and future directions for this line of work using NLP to examine public discourse around nuclear waste. In addition to introducing a new methodology to support consent-based siting-related activities, this study highlights key nuances to consider when leveraging NLP techniques (from data collection to analysis techniques).

[a] News media articles often use blanket terms such as "nuclear waste" or "waste" to refer to material such as spent nuclear fuel, high-level radioactive waste, transuranic waste, and/or low-level radioactive waste, which may or may not match U.S. statutory definitions for different nuclear and radioactive materials and/or wastes. The authors opted to use "nuclear waste" and "waste" interchangeably here to refer to these materials in a similar manner to the news articles under study.

II. METHODS

Data for this study were collected from DataNews.[22] DataNews was (at the time of data collection)[b] a news content aggregation website that enabled users to download articles through an application programming interface (API). Users can query the database to gather different sources, headlines of recent publications, and 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 U.S. source restriction, and reverse-ordered date sorting; API queries were conducted using the *R* package *rjson*.[23] The authors elected to use a broad keyword search (e.g., as opposed to "nuclear waste") in order to minimize the likelihood that the query would exclude relevant articles that used different terminology (e.g., "spent nuclear fuel"). Moreover, this approach allows analysts to assess how coverage rates of "waste"-related issues compare to other nuclear topics, such as energy production, weapons development, etc. Once duplicate entries were removed, the corpus (i.e., collection of text documents) contained approximately 150 000 articles from nearly 5000 news sources and written by nearly 23 000 unique bylines.[c] All data collection, cleaning, and analyses were performed using open-source software *R*, version 4.1.3.[24]

Figure 1 depicts the procedures that this study followed after data collection. Specifically, the authors 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 listed publication dates prior to 5 years before the first date of data collection (January 31, 2022). Since these articles were beyond the scope of our intended sample, and because there was a greater likelihood that these listed publication dates were erroneous, this study explicitly excluded articles that were published greater than 5 years prior to the data collection date; this dropped $n = 274$, or 0.18% of articles, from the analysis and resulted in a corpus that ranged in publication date from February 1, 2017, to January 31, 2022. Figure A.1 in the Appendix depicts the frequency of articles published across this date range. Additionally, "stop words" (e.g., "the," "an," "of," etc.) were removed from the corpus using the "textProcessor" function[d] in the *stm R* package since they convey little meaning in text and can influence similarity assessments of documents. After the data cleaning procedures were performed, the final corpus contained $n = 148\ 322$ articles from 4813 unique sources and written by 22 785 unique bylines.

II.A. NLP 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, three NLP methods were utilized: (1) topic modeling, (2) entity recognition, and (3) sentiment analysis (Fig. 1).

Topic modeling was used at two points: first, to identify the subset of news articles that are focused on "waste"-related issues and second, to identify subtopics within the "waste"-related subset of articles. Topic modeling is a method that bins documents into groups pertaining to similar subject matter. The method typically involves parsing a matrix of documents and the tokens, or unique words, contained within them to identify common terms within a group of documents and to separate different groups. The authors estimated structural topic models (STMs) using the *stm R* package[25] for this analysis, which implements FREX words—or words ordered by the harmonic mean of their FRequency within a group of

[b] DataNews was subsequently purchased by, and its data/API rolled into the services of, Perigon (https://www.goperigon.com).

[c] In the rare instance that byline information was missing for an observation, the authors decided to encode a unique random string for the byline. This is relevant in the structural topic model, described below, where the byline field was used as a topic covariate. The *stm* package used for this analysis is not equipped to account for missing data, and so, rather than throw out potentially useful content, the study authors decided to allow these few observations to be considered as though they had unique bylines. Although it is unknown exactly to what extent this analytic decision shaped our results, it is unlikely to have dramatically shifted the topic fit values in such a way as to change the plurality topic assignment of any article.

[d] The "textProcessor" function performs a number of preprocessing or "cleaning" steps, such as "stop word" removal with the SMART dictionary of "stop words,"[53] as well as removal of punctuation and numbers. Importantly, by default, this function also attempts to stem words to their common root (e.g., both "communicating" and "communication" become "communicat"), a process that is closely related to lemmatization in the NLP literature. The authors elected not to stem or lemmatize the articles prior to analyses because the computational complexity of the model estimation was not sufficiently high as to require text simplification. Furthermore, stemming and lemmatization can be detrimental when two words with the same root but different meanings in their respective contexts (e.g., "lead contamination" or "leading the industry") are confounded in the results. Moreover, these morphological changes to the text can impact the results of dictionary-based downstream analyses, such as sentiment analysis.[54,55]

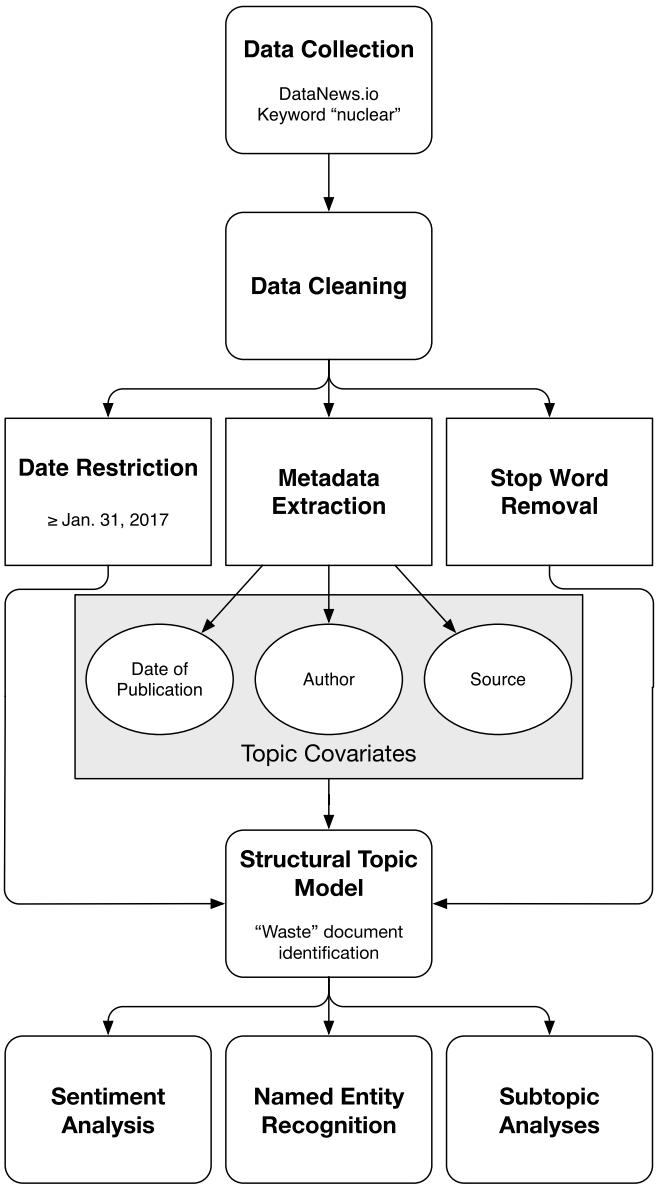


Fig. 1. Path diagram of data collection, cleaning procedure, and analysis strategy.

documents and relative EXclusivity to that group—to help analysts identify the subject matter of a topic. Importantly, STMs are mixture models, meaning that they do not assume that each document pertains to just a single topic; rather, documents are assigned a proportion "fit" score to each of the modeled topics. STMs differ from other topic modeling methods, such as Latent Dirichlet Allocation, in that they allow for the specification of document metadata as topic covariates; for more information about STMs, see Roberts et al.[26]

In other words, STMs estimate topic prevalence and content as both a function of document-term co-occurrence and as a function of features of the document, such as the byline (categorical) or date of publication (continuous). Using this information, one can both or separately account for the ways that the same topics can be written about using different words by, for example, different authors, and researchers can calculate model estimates for topic prevalence over different values of a particular covariate.

The STM approach of accounting for document-level metadata covariation has seen a wide variety of applications from open-ended survey responses,[27] to social media analysis[11] and regional responses to the COVID-19 pandemic.[28] For this newspaper study, an article's date of publication, byline, and source were used as prevalence covariates in the model specification.[e] Typically, k, or the number of topics one would like the model to identify, is specified by the analyst a priori. However, because there was no expectation of the number of groups of documents that might be uncovered in the broad "nuclear" corpus, the authors opted to utilize an algorithm for rapid topic number identification that is packaged with the *stm* software.[29] The model produced 71 topics of varying size and topic quality (see Fig. A.2 in the Appendix) that were further evaluated to identify "waste"-related topic(s). The model generates topic fit scores for each document, which reflects the proportion of the article that corresponds to each topic. The authors generated subject matter labels for each topic by first evaluating a list of the top ten FREX words for each topic, and then those labels were verified or changed by conducting a close read of a sample of the top-scoring articles for each topic. The remaining analyses use a subset of the corpus that contains articles for which a "waste"-related topic was a plurality of the document proportion (i.e., a "waste"-related topic was that article's highest fit score).

After subsetting to "waste"-related topics, NER was implemented to identify key entities (such as organizations, persons, etc.) that are referred to in these articles. NER was performed using the *R* package *spacyr*.[30] This software uses a language model[f] that was pretrained on general 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

[e] Date of publication, byline, and source were not included as content covariates in the model because the report authors were simply interested in extracting model information about how topic prevalence varied across values of the included covariates. Future work is needed to assess how topic content would shift if these features were included as content covariates as well.

[f] Specifically, the "en_core_web_lg" version 3.5.0 model was used,[56] which purports to have high precision [.85], recall [.86], and F-score [.85] for NER tasks.

identities, and more. For example, in the sentence "Contrary to their myopic moves, the Union of Concerned Scientists reported 'newly built reactors must be demonstrably safer and more secure,'" the organization "Union of Concerned Scientists" is identified by NER techniques. Such analyses not only can help shed light on which groups are mentioned in the context of "nuclear waste" but also could identify key opinion leaders. The rates at which States (geopolitical entities) appear in the text were further examined to assess regional differences in the frequency of "nuclear waste" coverage.

Next, sentiment analysis was implemented to evaluate the use of affective language (i.e., moods, feelings, and attitudes) within the "waste"-related articles. Sentiment analysis was performed using the *R* package *sentimentr*.[31] This package uses a built-in dictionary of valenced and weighted 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." The package author demonstrates that the algorithm accurately classifies 75% of texts that were labeled manually by human coders, which is an ~15% increase over the valenced word dictionary alone.[32] 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,[33,34] these insights could serve as a proxy for public opinions as expressed within newspapers.

Finally, the study authors estimated a second STM on the "waste"-related subcorpus. This second model also utilized each document's byline, source, and date of publication as topic covariates and leveraged the algorithm put forth by Mimno and Lee[29] to derive an appropriate number of topics, or k-value. The resulting 66-topic model allowed for the identification of various subtopics within the "nuclear waste" domain. This report delves into two of these subtopic areas to demonstrate different nuances that emerge across nuclear "waste"-related dimensions.

III. RESULTS

Two models will be discussed: (1) a 71-topic "nuclear" topic model and (2) a 66-topic "nuclear waste" topic model. The first model was used to identify "nuclear waste"–related topics; those articles that fit highest to "waste" topics were then reanalyzed in the second model to identify patterns within "nuclear waste"–relevant articles. For clarity, findings from these two nested STMs will be referred to as "Topics" (for 71-topic "nuclear" model) and "Subtopics" (for 66-topic "nuclear waste" model).

III.A. "Nuclear" Topic Model (71-Topic Model)

As mentioned in Sec. II, the full cleaned corpus (DataNews: n = 148 322) was first analyzed using a STM to identify "waste"-relevant news articles. In the first implementation, the authors set the k-value—or the number of topics into which to split the corpus—to zero; this allows the model to use an algorithm developed by Mimno and Lee[29] to identify an appropriate number of topics to model. Run on the DataNews corpus, this algorithm identified 71 topics. Their relative prevalence and the FREX words, or words that are ordered by the harmonic mean of their frequency within the topic and relative exclusivity to that topic, are shown in Fig. 2. In order to identify which topics pertained to "waste"-related subject matter, the authors leveraged the list of the 100 highest-scoring FREX words within each topic; if "waste" appeared among this list, the topic was labeled as a "nuclear waste"–related topic.[g] In the DataNews results, just one topic, number 70, was "waste"-related. Interestingly, "aquifer," "waste," "repository," and "cleanup" were included among the five most frequent and exclusive words used in this topic.

This "waste"-related 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 Fig. 3a. 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 be a by-product of either sparse coverage of the nuclear (broadly construed) topic in news media or limitations of the dataset

[g] The authors opted to use the term "waste" to identify relevant topics because they suspected that this would be the most common term and because the FREX algorithm is optimized to identify singular words instead of phrases. That said, to the extent that those synonyms are interchangeable with "waste" and to the extent that they appear in newspaper articles on the subject, the STM should still place those words or phrases into the same topic. And, indeed, words that are associated with "waste" in the media, such as "spent," "fuel," "canisters," and "disposal," are among Topic 70's 1000 highest-scoring FREX words; those same words are not associated with any other topic in the modeled results.

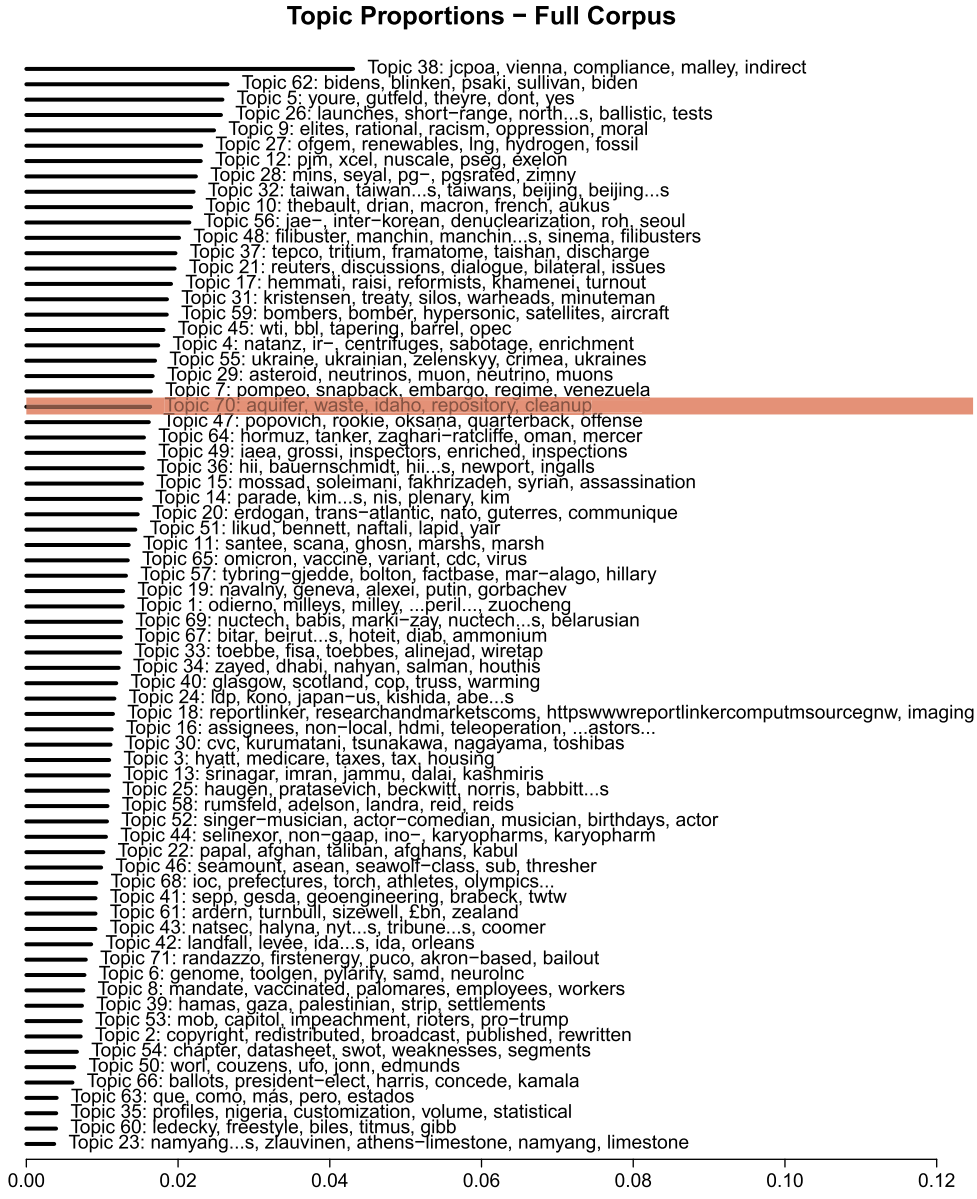


Fig. 2. Prevalence of 71 modeled topics across the full DataNews corpus ($n \approx 148\,000$ articles). Topic numbers are shown along with top five FREX words. The "waste"-related topic (those with the word "waste" among their top 100 most frequently occurring words) is highlighted in red.

itself (e.g., limited licensing agreements to republish older material[h]). Subsetting the corpus to contain only articles that predominantly discussed "nuclear waste" (i.e., documents with the highest fit score to Topic 70) resulted in a collection of $n = 2526$ articles from 372 sources and 527 unique bylines that were further analyzed for key entities, sentiments, and subtopics.

The entities identified with NER most frequently within "nuclear waste" articles highlight specific geographies (e.g., United States, Navajo Nation, and New Mexico; Fig. 4a). These could reflect either locations near nuclear-related facilities or entities that are involved with nuclear waste activities. Additionally, specific organizations also emerge, such as the U.S. Nuclear Regulatory Commission, the DOE, Congress, national laboratories, and private companies that are involved in nuclear waste activities. A couple of key persons, namely, recent U.S. presidents whose administrations shaped regulatory policy, also emerged in

[h] Unfortunately, since DataNews ceased operations in February of 2022, the authors have no way of knowing whether their licensing agreements allowed them to republish data at a consistent rate from each source over the past 5 years.

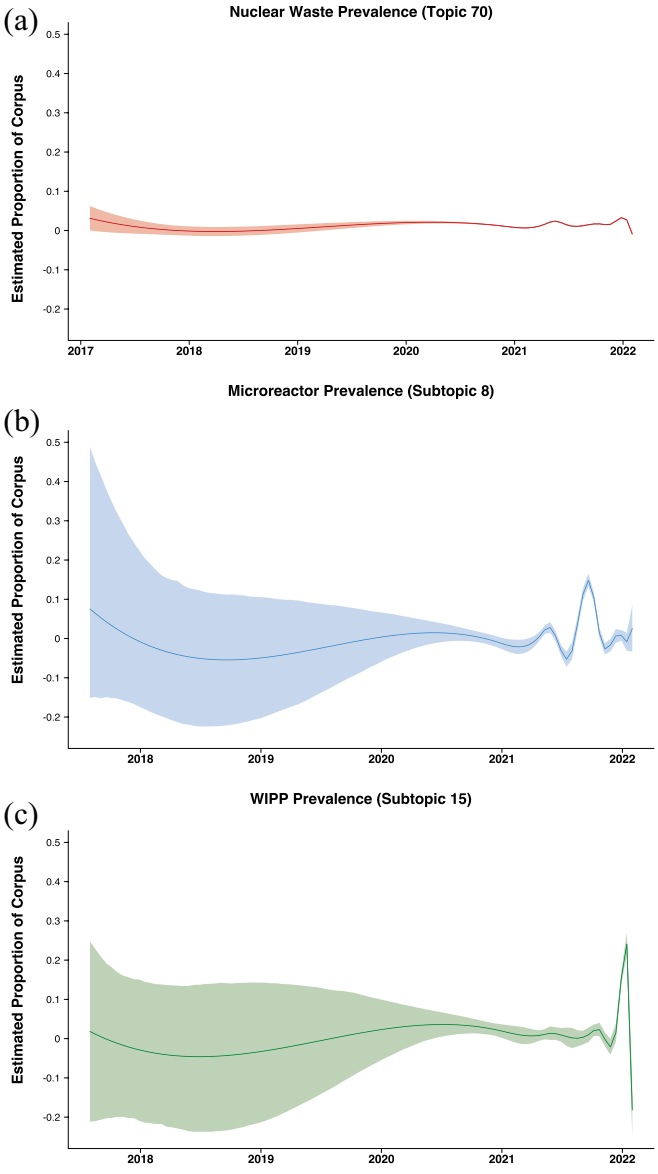


Fig. 3. Prevalence of (a) Topic 70, "Nuclear Waste," (b) Subtopic 8: "Microreactors," and (c) Subtopic 15: "Waste Isolation Pilot Plant" (WIPP) over time. The curvilinear lines indicate the estimated proportion of all news articles that pertain to the respective topic or subtopic while the shaded areas indicate the 95% confidence interval of those estimates.

these articles; additional evaluation would be needed to understand regional and local individuals prevalent in nuclear waste siting discussions. Further evaluation would also be needed to understand frequent references to historical events (such as World War II and the Cold War) in the context of "nuclear waste."

In order to understand regional variations in the "nuclear waste"–related discussions within the analyzed corpus, the study next examined the frequency with which different U.S. states were labeled by the NER procedure. Figure 5 shows a map of the relative prevalence of states. New Mexico—which is a state that does not currently host an active, large-scale nuclear energy reactor but is home to the Waste Isolation Pilot Plant (WIPP), a deep geological repository for defense-generated transuranic radioactive waste, and also the proposed location of a private commercial spent nuclear fuel storage facility—is the most frequently discussed state in the "waste" subset of articles. Other states that are frequently discussed include Washington, Idaho, Nevada, South Carolina, and Texas, all of which have DOE facilities that currently house nuclear waste.[35] Additionally, Washington, South Carolina, and Texas each host at least one large-scale nuclear energy reactor that is currently operating. Interestingly, several other current locations of commercial nuclear energy reactors, such as Illinois, Michigan, or New York, are not often discussed in the context of "nuclear waste" articles.

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 (i.e., no sentiment valence); $t(2525) = -12.12$, $p < 0.001$. The distribution of average sentence-level sentiments for each "waste"-related article is shown in Fig. 6a. The larger tail on the left (relative to the right) reflects the general tendency of news media to use negative valenced words more often compared to positive words. However, the vast majority of "nuclear waste" articles use generally neutral language; 90.06% of these articles used fewer than one affective word for every five sentences (± 0.2). Further, the range of average sentiment values was relatively constrained—range: $[-0.636, 0.524]$—indicating that even the most affective articles about "nuclear waste" tend to limit valenced words. By comparison, the average article in the full corpus tended to use positive valenced sentiments ($M = 0.022$, $SD = 0.115$), and the extremes at either end of the scale were much further apart than those articles in the "nuclear waste" topic—range: $[-1.028, 2.089]$.

III.B. Subtopics Within "Nuclear Waste" (66-Topic Model)

Finally, the authors ran a second STM on just the "waste"-related subset of articles. This secondary model identified many interesting subthemes within "nuclear waste" subject matter (see Fig. A.3 in the Appendix). To distinguish the second "nuclear waste"–only STM from the first nuclear (broadly) STM, later sections of

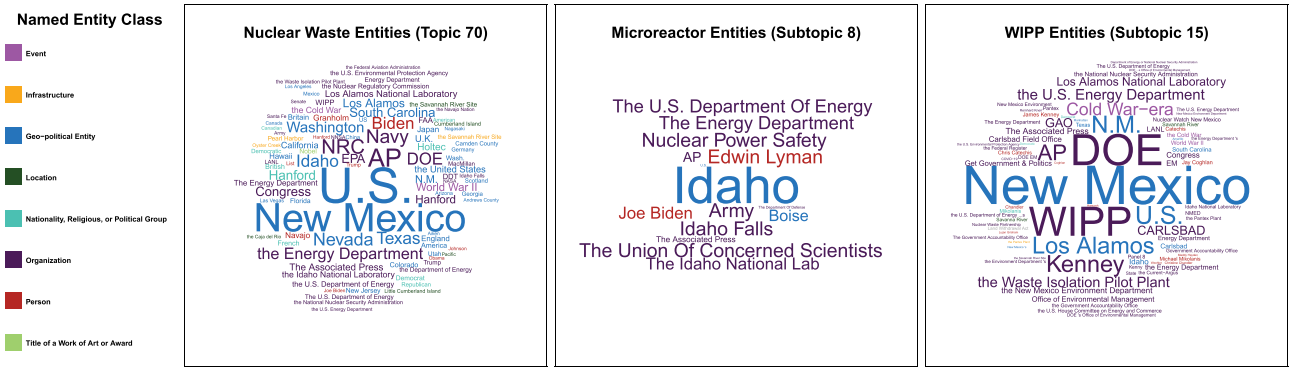


Fig. 4. Wordcloud of the 100 most frequently occurring named entities among all (a) "Nuclear Waste" articles, (b) "Microreactor" subtopic articles, and (c) "Waste Isolation Pilot Plant" (WIPP) subtopic articles. The microreactor subtopic contained just 23 unique named entities, so all are plotted here. Larger words appear more frequently in text than smaller ones. Note that the same entity may appear multiple times as a different entity type; for example, "Trump" appears as both a person (i.e., former President Donald Trump) and an organization (i.e., the Trump Administration).

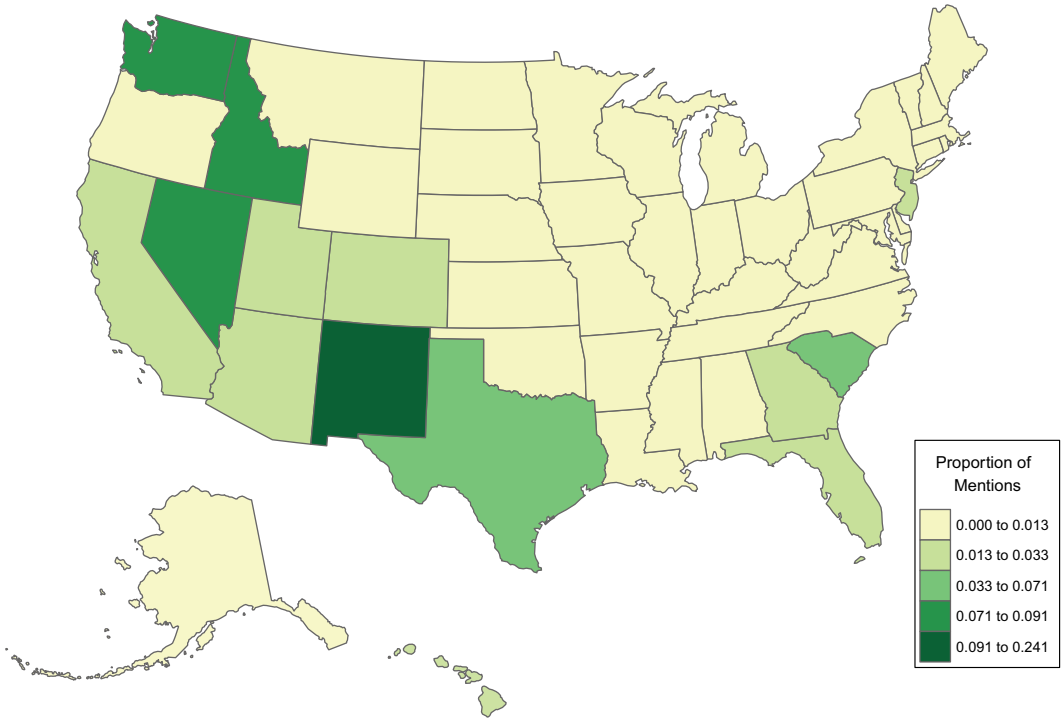
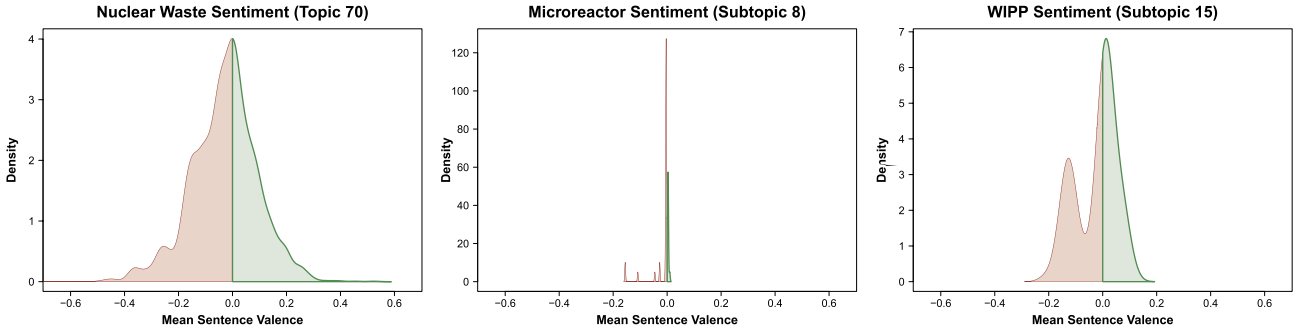


Fig. 5. Prevalence with which each state is named in the "nuclear waste" subset of articles. Darker shades of green indicate that the state is mentioned with greater relative frequency. State names were identified in NER results.

this report will refer to the topics generated for the "waste"-specific set of articles as "subtopics." In the interest of parsimony, this report will focus on two of the 66 subtopics from this model: Subtopic 8, which contains FREX words "microreactor," "prototype," and "terawatt-hours," and Subtopic 15, which contains the FREX words "shipments," "WIPP," and "gloves." The authors qualitatively labeled the former subtopic "microreactors" and the latter subtopic "Waste Isolation Pilot Plant" for ease of interpretation.

These two subtopics were selected because they concern very specific subject matter within nuclear waste in U.S. contexts more generally. Microreactors represent an important technological advancement in the industry, and future applications promise to, among other things, offer transportable heat and electricity sources to areas recovering from natural disasters.[36] Additionally, the Waste Isolation Pilot Plant is a deep geological repository for defense-generated transuranic radioactive waste located in Eddy County, New Mexico.[37] Although it does not



(a) (b) (c)

Fig. 6. Kernel density plots of the mean sentence sentiment among all (a) "Nuclear Waste" articles, (b) "Microreactor" subtopic articles, and (c) "Waste Isolation Pilot Plant" (WIPP) subtopic articles. The red shaded area indicates that more negatively valenced words were used in those articles while the green shaded area indicates that the articles expressed more positive sentiments.

house spent nuclear fuel from energy production, the site offers an important point of reference for the public and other stakeholders to illustrate a potential deep geological repository for spent fuel and the effects of a repository on its surrounding communities. Among the "waste"-related news articles, $n = 47$ stories had microreactors as their highest-scoring subtopic, while $n = 88$ documents had a plurality of their content fit to the Waste Isolation Pilot Plant subtopic. The authors verified that the articles assigned the highest-fitting scores to these subtopics were written about these subtopics by conducting a close read of a sample of the top-scoring articles from each data subset, or subcorpora. The subcorpora were further analyzed for sentiments and named entities.

Among "waste"-related news articles, the Waste Isolation Pilot Plant was the eighth most frequently covered subtopic, representing about 3.5% of the "waste" subject matter. As Fig. 3b shows, though, much of the attention to the Waste Isolation Pilot Plant took place over a relatively short period between October 2021 and the end of our data collection period in January 2022. By contrast, the microreactor subtopic appears less frequently in "waste"-related articles, ranking 16th among the identified subtopics and representing less than 2% of the "waste" subcorpus. Similar to the Waste Isolation Pilot Plant subtopic, though, coverage of microreactors is also more recent, peaking in late 2021 with announcements around the advancing technology (see Fig. 3c).

The Waste Isolation Pilot Plant and microreactor subtopics were further analyzed for sentiment. As shown in Figs. 6b and 6c, both subtopics used even more neutral language compared to the "nuclear waste" topic more broadly, as evidenced by their more limited range (Waste Isolation Pilot Plant = $[-0.206, 0.108]$; microreactors = $[-0.155, 0.010]$). News articles about microreactors tended to use more negative language (relative to zero or "balanced," $M = -0.012$, $SD = 0.035$), and a further univariate t-test revealed that the sentiment scores of these articles did in fact significantly differ from zero [$t(46) = -2.34$, $p < .05$]. Articles in the Waste Isolation Pilot Plant subtopic evinced a similar tendency to use negative words compared to positive ones ($M = -0.024$, $SD = 0.076$; $t(87) = -2.96$, $p < .01$). Taken together, these results indicate that although the range of sentiment values are substantially more constrained for the subtopics, the average microreactor or Waste Isolation Pilot Plant article is not affectively different, in terms of the amount and valence of sentiment, from the average "nuclear waste" article. Further analyses of other subtopics contained within "nuclear waste" would be necessary to explain the wider variance of sentiment in the broader "nuclear waste" topic. In other words, such analyses would help answer the question "why do discussions of 'nuclear waste' (broadly) vary more in their use of affective words compared to news articles about the microreactor and Waste Isolation Pilot Plant subtopics?"

For the last analysis, NER was conducted over the subcorpora of articles pertaining to microreactors and the Waste Isolation Pilot Plant. The most common entities are shown in Figs. 4b and 4c. Among the identified entities in the microreactors subtopic are "Idaho," "Boise," "Idaho Falls," and "The Idaho National Lab," suggesting that much of the discussion of microreactors in the news between 2017 and 2022 centers around that one state; no other regions of the country were identified. Similarly, the discussion of "waste" in the context of the Waste Isolation Pilot Plant was also primarily centered on one location in the country, as "New Mexico," "N.M.," "Los Alamos," and "Carlsbad" all appear among the most frequently named entities. Both subtopics mention specific persons with great frequency as well, suggesting that

these entities may be important stakeholders, opinion leaders, or have legislative or regulatory responsibilities with regard to their respective application areas.

IV. DISCUSSION

In the views of its authors, this analysis demonstrates the significant potential for NLP techniques to gain insights into narratives around "nuclear waste." Specifically, these results highlighted that the general discourse regarding "nuclear waste" within the news media has been fluctuating over recent years relative to "nuclear" subject matter more generally. Additionally, commonly mentioned entities reflect a limited number of geographies and stakeholders. General sentiments within the "nuclear waste" articles also appear to predominantly use neutral language, although the specific distribution of sentiment varies depending on whether you are looking at all or specific subtopics within "waste"-related articles (Fig. 6).

This analysis also underscores how people view current issues in the lens of related historical and ongoing events. For example, as noted above, the Waste Isolation Pilot Plant is a deep geological repository for defense-generated transuranic radioactive waste located in Eddy County, New Mexico. Although this site does not store spent fuel, the prevalence of this topic within the corpus indicates that work at the Waste Isolation Pilot Plant may be a reference point for the public regarding the siting of other nuclear waste management facilities. Continuing to be mindful of research and lessons learned from related events from the Waste Isolation Pilot Plant[38] and past siting attempts[6] to activities in other countries[39,40] will be important.

Relatedly, the finding that sentiment score distributions of "nuclear waste"–related articles (as well as those concerning Waste Isolation Pilot Plant and microreactors more specifically) are mostly neutral may or may not be surprising. On the one hand, the norm of journalistic objectivity and the scientific nature of this topic, compared to other unrelated topics covered in the news, ought to result in relatively little deviation from zero sentiment. On the other hand, given the history of negative perceptions about the risks involved around nuclear waste siting,[6,41,42] it would not be surprising to the authors if negative sentiments outnumbered positive ones, even if the extremity of sentiment scores remained close to zero. That the distribution of sentiment scores is so close to zero and nearly symmetrical indicates either that the norms of journalistic objectivity in scientific reporting around nuclear issues are quite strong or that news sources are beginning to shift away from negative coverage of nuclear issues in a way that is perhaps reflective of, or a cause of, shifting public sentiment.[43,44] A more historical analysis (i.e., greater than the 5-year archive offered by DataNews) is required to elucidate such a shifting trend in sentiment words used in newspaper articles about nuclear waste, if it exists.

The NER outputs can also serve as a basis for understanding equity nuances (i.e., a fair distribution of costs and benefits), by integrating sociodemographic information of regions discussed in the newspaper articles. Key distinctions emerge across the states, including percentages of persons with different racial or ethnic backgrounds, population density, poverty rates, and foreign-born persons.[45] For example, there are significantly more Native Americans[i] in New Mexico while the percentage of the population that is foreign born is much greater in Texas and Washington (Table A.I). These types of nuances can be important for designing communication strategies that ensure inclusive engagement of diverse groups, which is critical for achieving equity and justice objectives, which is a governmental priority for spurring economic growth as well as supporting climate objectives.[46,47] In particular, mobilization of local knowledge may be needed to ensure that more fair, just processes are considered for siting process activities.[48]

A number of research directions could be explored in future work, from expanding the data sources and NLP algorithms considered to augmenting the analysis techniques with mixed methods. For example, the data sources utilized in the analysis could be tailored to support focused regional research or evaluate multiple languages. The latter might be especially important in regions with notable foreign-born persons while the former will require metadata about the specific location and distribution of analyzed newspaper sources (not included in this analysis). Further research including newspaper articles with greater historical depth (i.e., greater than 5 years in the past) could provide additional context about how the language around "nuclear waste" issues has shifted, particularly concerning prior siting attempts. In such an analysis, it may also be prudent to include date of publication as a content covariate (as opposed to only a prevalence covariate, as the model in this report used) to account for the way that the words used to describe a particular subject matter may shift over time. The sources considered could also be expanded to incorporate social

[i] "Native American" is U.S. Census terminology, cited in Table A.I, that encompasses a variety of Tribal and Puebloan community members. The authors acknowledge the limitations of this terminology in describing the rich cultural, ethnic, and religious diversity that this term is applied to.

media discourse to understand real-time evolution of narratives. Combining source expansion with other data collection techniques, such as stakeholder interviews, could also address concerns regarding biases in newspaper coverage that have been identified in environmental studies.[49]

Future work could also 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. Transformer-based language models (e.g., BERT[50]) and aspect-based sentiment methods could be used to generate additional insights. For example, aspect-based sentiment analysis, which is a procedure where sentiments are assigned to named entities they refer to by parsing the sentence structure, could be used to assess whether "waste"-related issues are referred to more positively or negatively in association with a specific entity.[51] Similarly, training a nuclear-specific transformer model to assign topics or label entities would allow analysts to more readily apply the methods presented here to new data. This could be particularly useful in a live, updating data synthesis tool for which the corpus of news articles is constantly growing over time. The generation of such an interactive visualization tool (or tools) could also enable users to more easily explore the high-level and specific contexts around the mention of specific topics, sentiments, and entities.

Finally, NLP techniques could be augmented with mixed methods, such as qualitative analyses. For example, qualitative analyses could be used to evaluate articles expressing strong negative or positive sentiment and could help illuminate important nuances about specific topics (relative to nuclear at-large and "waste"-specific ones) for certain groups. Similar deep dives could be done for subtopics, such as micronuclear or modular nuclear reactors,[52] whose recent attention in the news media is reflective of the shifting public focus toward these capabilities and thus may be important touchpoints for the future of the nuclear energy industry. These findings can, in turn, be utilized to help support knowledge management (e.g., curation of resources) and community and group engagements that can help achieve broader goals of the consent-based siting process.

APPENDIX

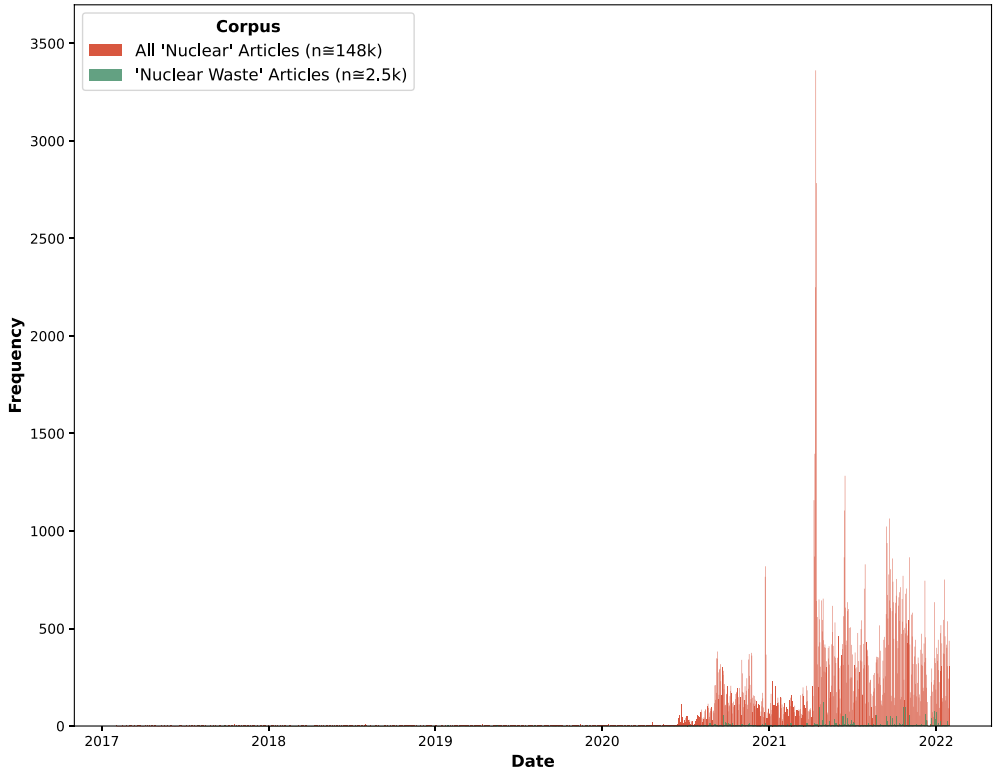


Fig A.1. Histogram of the number of articles by publication date for both the full "nuclear" DataNews corpus (red bars; $n \approx 148\,000$ articles) and the "waste"-related subset (green bars; $n = 2526$ articles).

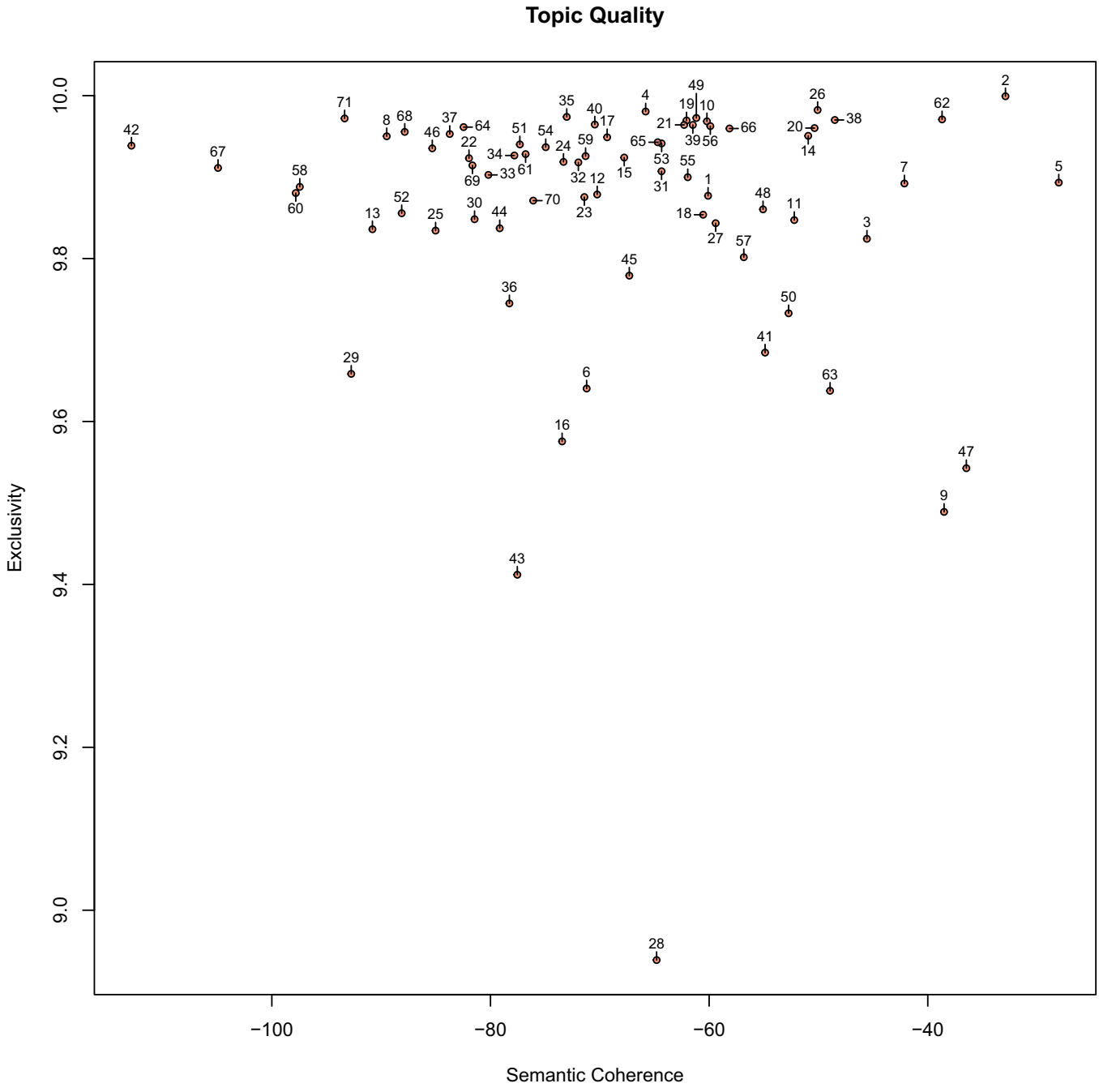


Fig A.2. Scatterplot of topic descriptive statistics from the initial STM ($n \approx 148\,000$ articles). "Semantic Coherence" refers to the likelihood that each topic contains words that occur more frequently together within documents. "Exclusivity," on the other hand, refers to the likelihood that the words that are assigned to one topic appear less frequently in other topics. In general, higher values of a topic's semantic coherence and exclusivity scores mean that the topic is more interpretable by the analyst. Additionally, higher clustering of these values across topics indicates that the model as a whole is more readily distinguishing topics from one another; in other words, the topics are of higher, comparable quality.

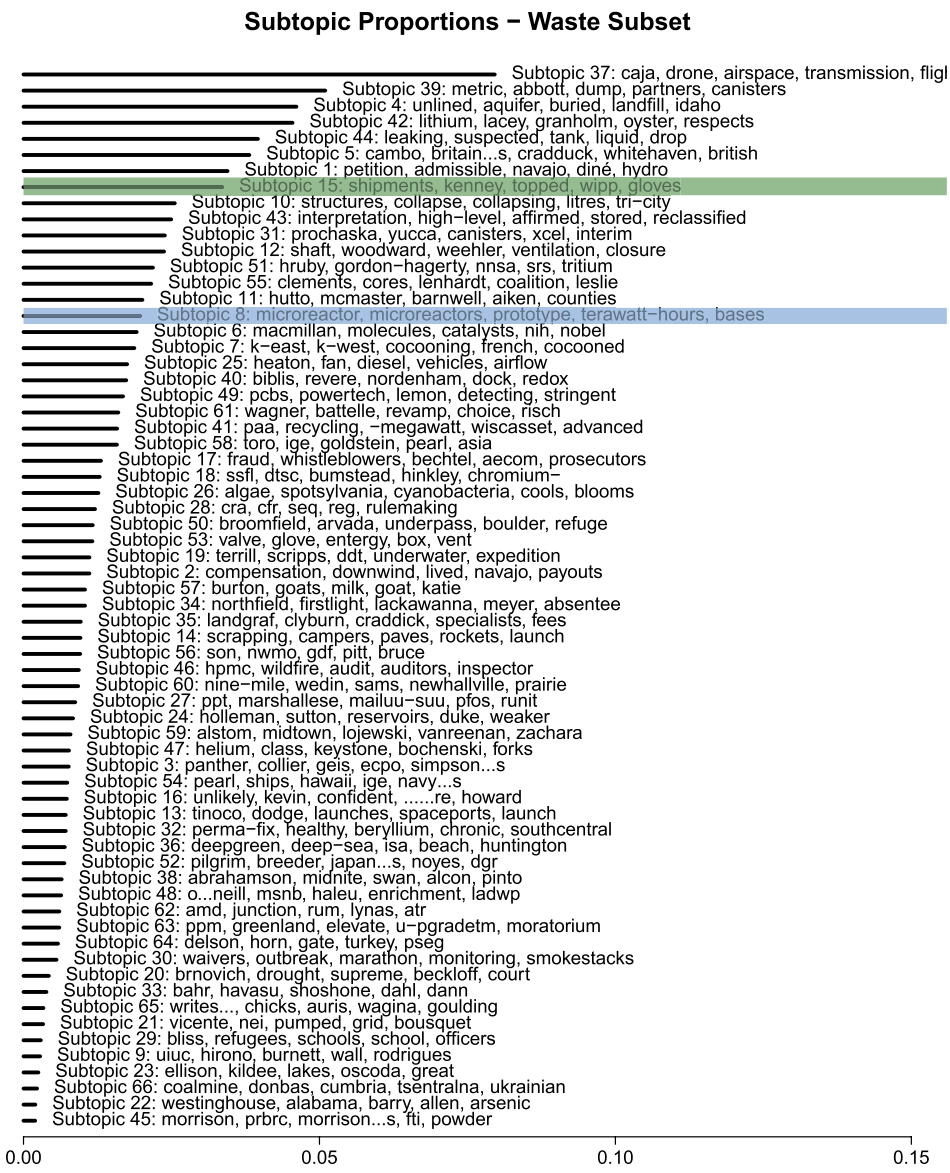


Fig A.3. Prevalence of 66 modeled subtopics across the "waste"-related subset of the DataNews corpus ($n = 2526$ articles). Subtopics shown with five FREX words; words are ordered by the harmonic mean of their frequency within the topic and exclusivity to that topic. The two subtopics this study explored further are highlighted; Subtopic 15 pertaining to the Waste Isolation Pilot Plant is shown in green, and Subtopic 8 concerning microcreactors is highlighted in blue.

TABLE A.I

Demographics of Key States Identified through NER

Region	White Alone (%)	Native American (%)	Veterans (%)	Foreign-Born Persons (%)	Persons in Poverty (%)	Mean Travel Time to Work (min)	Population per Square Mile
Idaho	92.8	1.7	6.08	5.9	10.1	21.2	22.3
New Mexico	81.3	11.2	6.69	9.2	16.8	22.7	17.5
Texas	77.9	1.1	4.86	16.8	13.4	26.6	111.6
Washington	77.5	2.0	6.69	14.5	9.5	28.0	115.9
United States	75.8	1.3	8.04	13.5	11.4	26.9	93.8

*"Native American" refers to "American Indian and Alaska Native Alone." "Foreign-Born Persons" are calculated across the period of 2016–2020. Data were sourced from the U.S. Census Bureau.[45]

Disclaimer

This is a technical report that does not take into account contractual limitations or obligations under the Standard Contract for Disposal of Spent Nuclear Fuel and/or High-Level Radioactive Waste (Standard Contract) (10 CFR Part 961).

To the extent discussions or recommendations in this report conflict with the provisions of the Standard Contract, the Standard Contract governs the obligations of the parties, and this report in no manner supersedes, overrides, or amends the Standard Contract.

This report reflects technical work that could support future decision making by DOE. No inferences should be drawn from this report regarding future actions by DOE, which are limited both by the terms of the Standard Contract and Congressional appropriations for the Department to fulfill its obligations under the Nuclear Waste Policy Act including licensing and construction of a spent nuclear fuel repository.

Acknowledgments

This work was funded by the Office of Nuclear Energy at the DOE. 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 DOE's National Nuclear Security Administration under contract DE-NA0003525. The views expressed in the article are those of the authors and do not necessarily represent the views of the DOE or the United States Government.

Disclosure Statement

No potential conflict of interest was reported by the authors.

ORCID

Matthew D. Sweitzer http://orcid.org/0000-0002-2075-6177
Thushara Gunda http://orcid.org/0000-0003-1945-4064

Data Availability Statement

The data are not publicly available due to licensing agreements between the data curator (datanews.io) and the respective newspaper publishers. The summary data that support the findings of this study are available on request from the corresponding author, MS.

References

1. "Nuclear Power in a Clean Energy System," International Atomic Energy Agency (2019); https://www.iea.org/reports/nuclear-power-in-a-clean-energy-system.
2. "Nuclear Explained," U.S. Energy Information Administration (2021); https://www.eia.gov/energyexplained/nuclear/data-and-statistics.php.
3. "Fission Vision: Doubling Nuclear Energy Production to Meet Clean Energy Needs," Nuclear Innovation Alliance (2022); https://nuclearinnovationalliance.org/fission-vision-doubling-nuclear-energy-production-meet-clean-energy-needs.
4. "Spent Fuel Pools," U.S. Nuclear Regulatory Commission (2020); https://www.nrc.gov/waste/spent-fuel-storage/pools.html.
5. "Dry Cask Storage," U.S. Nuclear Regulatory Commission (2023); https://www.nrc.gov/waste/spent-fuel-storage/dry-cask-storage.html.
6. P. SLOVIC, M. LAYMAN, and J. H. FLYNN, "Risk Perception, Trust, and Nuclear Waste: Lessons from Yucca Mountain," *Environment*, **33**, *3*, 6 (1991).
7. "Consent-Based Siting: Request for Information Comment Summary and Analysis," U.S. Department of Energy, Office of Nuclear Energy (2022); https://www.energy.gov/sites/default/files/2022-09/Consent-Based%20Siting%20RFI%20Summary%20Report%200915.pdf.
8. "Consent-Based Siting," U.S. Department of Energy, Office of Nuclear Energy; https://www.energy.gov/ne/consent-based-siting (n.d.).
9. D. LAZER et al., "Computational Social Science," *Science*, **323**, *5915*, 721 (2009); https://doi.org/10.1126/science.1167742.
10. E. CAMBRIA and B. WHITE, "Jumping NLP Curves: A Review of Natural Language Processing Research," *IEEE Comput. Intell. Mag.*, **9**, *2*, 48 (2014); https://doi.org/10.1109/MCI.2014.2307227.
11. J. B. BAYER et al., "Reimagining the Personal Network: The Case of Path," *Social Media Soc.*, **8**, *3* (2022); https://doi.org/10.1177/20563051221119475.
12. D. PARANYUSHKIN, "InfraNodus: Generating Insight Using Text Network Analysis," *World Wide Web Conference*, p. 3584 (2019).
13. S. MIN and J. PARK, "Modeling Narrative Structure and Dynamics with Networks, Sentiment Analysis, and Topic Modeling," *PloS One*, **14**, *12*, e0226025 (2019); https://doi.org/10.1371/journal.pone.0226025.
14. K. M. KWAYU et al., "Discovering Latent Themes in Traffic Fatal Crash Narratives Using Text Mining Analytics and Network Topology," *Accid. Anal. Prevent.*, **150**, 105899 (2021); https://doi.org/10.1016/j.aap.2020.105899.

15. R. OSHIKAWA, J. QIAN, and W. Y. WANG, "A Survey on Natural Language Processing for Fake News Detection," *arXiv preprint arXiv:1811.00770* (2018).

16. K. ISOAHO, D. GRITSENKO, and E. MÄKELÄ, "Topic Modeling and Text Analysis for Qualitative Policy Research," *Policy Stud. J.*, **49**, *1*, 300 (2021); https://doi.org/10.1111/psj.12343.

17. M. D. SWEITZER, T. GUNDA, and J. M. GILLIGAN, "Water Narratives in Local Newspapers Within the United States," *Front. Environ. Sci.*, **11**, (2023); https://doi.org/10.3389/fenvs.2023.1038904.

18. J. HIRSCHBERG and C. D. MANNING, "Advances in Natural Language Processing," *Science*, **349**, *6245*, 261 (2015); https://doi.org/10.1126/science.aaa8685.

19. J. STEFAN et al., "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges," *Proc. Eurographics Conf. Visualization (EuroVis 2015)*, N–A, R. BORGO, F. GANOVELLI, and I. VIOLA, Eds., Cagliari, Italy, 2015.

20. K. KRIPPENDORFF, "Content Analysis," *International Encyclopedia of Communications*, Vol. 1, p. 403, E. BARNOUW et al., Eds., Oxford University Press, New York.

21. S. E. STEMLER, "Content Analysis," *Emerging Trends* in the *Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource* (2015).

22. "DataNews News API," DATANEWS.IO; https://datanews.io (n.d.).

23. A. COUTURE-BEIL, rjson: JSON for R website, r package version 0.2.21 (2022); https://CRAN.R-project.org/package=rjson.

24. R CORE TEAM, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing (2022); https://www.R-project.org/.

25. M. E. ROBERTS, B. M. STEWART, and D. TINGLEY, "stm: An R Package for Structural Topic Models," *J. Stat. Software*, 91, 2, 1 (2019); https://doi.org/10.18637/jss.v091.i02.

26. M. E. ROBERTS et al., "The Structural Topic Model and Applied Social Science," *Proc. Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, Vol. 4, p. 1 (2013).

27. M. E. ROBERTS et al., "Structural Topic Models for Open-Ended Survey Responses," *Am. J. Polit. Sci.*, **58**, *4*, 1064 (2014); https://doi.org/10.1111/ajps.12103.

28. G. CAPANO et al., "Mobilizing Policy (In) Capacity to Fight COVID-19: Understanding Variations in State Responses," *Policy Soc.*, **39**, *3*, 285 (2020); https://doi.org/10.1080/14494035.2020.1787628.

29. D. MIMNO and M. LEE, "Low-Dimensional Embeddings for Interpretable Anchor-Based Topic Inference," *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014, p. 1319, Association for Computational Linguistics (2014).

30. K. BENOIT and A. MATSUO, "spacyr: Wrapper to the 'spaCy' 'NLP' Library," r package, version 1.2.1 (2020); https://CRAN.R-project.org/package=spacyr.

31. T. W. RINKER, "sentimentr: Calculate Text Polarity Sentiment," version 2.9.0 (2021); https://github.com/trinker/sentimentr.

32. T. W. RINKER, "sentimentr" (2021); https://cran.r-project.org/web/packages/sentimentr/readme/README.html.

33. M. E. MCCOMBS, D. L. SHAW, and D. H. WEAVER, "New Directions in Agenda-Setting Theory and Research," *Mass Commun. Soc.*, **17**, *6*, 781 (2014); https://doi.org/10.1080/15205436.2014.964871.

34. W. RUSSELL NEUMAN et al., "The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data," *J. Commun.*, **64**, *2*, 193 (2014); https://doi.org/10.1111/jcom.12088.

35. S. PETERS et al. "Spent Nuclear Fuel and Reprocessing Waste Inventory," U.S. Department of Energy, Office of Nuclear Energy Spent Fuel and Waste Disposition (2021); https://sti.srs.gov/fulltext/FCRD-NFST-2013-000263.pdf.

36. "What Is a Nuclear Microreactor?" U.S. Department of Energy, Office of Nuclear Energy (2021); https://www.energy.gov/ne/articles/what-nuclear-microreactor.

37. "U.S. Department of Energy's Waste Isolation Pilot Plant," U.S. Department of Energy, Office of Nuclear Energy; https://www.wipp.energy.gov/ (n.d.).

38. V. IALENTI, "Drum Breach: Operational Temporalities, Error Politics and WIPP's Kitty Litter Nuclear Waste Accident," *Soc. Stud. Sci.*, **51**, *3*, 364 (2021); https://doi.org/10.1177/0306312720986609.

39. Q. WANG and X. CHEN, "Regulatory Failures for Nuclear Safety—The Bad Example of Japan—Implication for the Rest of World," *Renewable Sustainable Energy Rev.*, **16**, *5*, 2610 (2012); https://doi.org/10.1016/j.rser.2012.01.033.

40. V. IALENTI, "Mankala Chronicles: Nuclear Energy Financing and Cooperative Corporate Form in Finland," *Nucl. Technol.*, **207**, *9*, 1377 (2021); https://doi.org/10.1080/00295450.2020.1868890.

41. H. KUNREUTHER, W. H. DESVOUSGES, and P. SLOVTO, "Nevada's Predicament Public Perceptions of Risk from the Proposed Nuclear Waste Repository," *Environment*, **30**, *8*, 16 (1988).

42. J. W. STOUTENBOROUGH, S. G. STURGESS, and A. VEDLITZ, "Knowledge, Risk, and Policy Support: Public Perceptions of Nuclear Power," *Energy Policy*, **62**, 176 (2013); https://doi.org/10.1016/j.enpol.2013.06.098.

43. A. S. BISCONTI, "Public Opinion on Nuclear Energy: Turning a Corner?" *NuclearNewswire* (2019); https://

www.ans.org/news/article-314/public-opinion-on-nuclear-energy-turning-a-corner/.

44. R. LEPPERT, "Americans Continue to Express Mixed Views About Nuclear Power," Pew Research Center (2022); https://www.pewresearch.org/short-reads/2022/03/23/americans-continue-to-express-mixed-views-about-nuclear-power/.
45. "Quickfacts: Idaho; New Mexico; Texas; Washington; United States," U.S. Census Bureau; https://www.census.gov/quickfacts/fact/table/ID,NM,TX,WA,US/PST045221 (n.d.).
46. "Executive Order on Tackling the Climate Crisis at Home and Abroad," Executive Order 14008 (2021).
47. "Advancing Racial Equity and Support for Underserved Communities Through the Federal Government," Executive Order 13985 (2021).
48. M. Z. BELL, "Spatialising Procedural Justice: Fairness and Local Knowledge Mobilisation in Nuclear Waste Siting," *Local Environ*., **26**, *1*, 165 (2021); https://doi.org/10.1080/13549839.2020.1867841.
49. M. D. CABALLERO, T. GUNDA, and Y. J. MCDONALD, "Pollution in the Press: Employing Text Analytics to Understand Regional Water Quality Narratives," *Front. Environ. Sci.*, 348 (2022).
50. J. DEVLIN et al., "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805* (2018).
51. H. H. DO et al., "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," *Expert Syst. Appl.*, **118**, 272 (2019); https://doi.org/10.1016/j.eswa.2018.10.003.
52. L. M. KRALL, A. M. MACFARLANE, and R. C. EWING, "Nuclear Waste from Small Modular Reactors," *Proc. National Acad. Sci.*, **119**, *23*, e2111833119 (2022); https://doi.org/10.1073/pnas.2111833119.
53. *The SMART Retrieval System: Experiments in Automatic Document Processing*, G. SALTON, Ed., Prentice-Hall (1971).
54. S. VIJAYARANI et al., "Preprocessing Techniques for Text Mining—An Overview," *Int. J. Comp. Sci. Commun. Net*., **5**, *1*, 7 (2015).
55. "When (Not) to Lemmatize or Remove Stop Words in Text Preprocessing," ModelOp (2019); https://www.modelop.com/blog/when-not-to-lemmatize-or-remove-stop-words-in-text-preprocessing/.
56. K. BENOIT and A. MATSUO, "spacy en_core_web_lg," version 3.5.0 (2023); https://spacy.io/models/ en.