Leveraging AI to Analyze Social Media and Local News to Understand Community Perspectives on Radioactive Waste Management - 26535

Matthew Weber[1], Mingwei Shen[1], Rachel Acosta[1], Steven Krahn[2], Megan Harkema[2]
[1]School of Communication and Information, Rutgers University
[2]Vanderbilt University

ABSTRACT

The present study reports on findings from an analysis of community data extracted from social media and local news media to understand community perspectives on radioactive waste management, and nuclear energy more broadly. Specifically, this work addresses different approaches to automatically analyzing and evaluating news media and social media content. As part of that effort, the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) is focused on developing procedures for assisting the DOE with its process for siting a federal consolidated storage facility (CSF) for spent nuclear fuel (SNF) storage. This work draws on two aspects of that work—as part of the Department of Energy's (DOE's) Collaboration-Based Siting (CBS) Consortia process and through a project funded by the U.S. Department of Energy, Office of Nuclear Energy (DOE-NE). Both projects leverage social and news media data and advanced analytical tools and techniques, including artificial intelligence (AI), to understand community sentiments on nuclear energy and waste.

A core concern in the long-term planning for the storage of SNF is the need for engagement and interaction with external and community stakeholders. Thus, this aspect of the work focuses on understanding and analyzing community perspectives based on existing community interactions documented in social and news media; our goal is to use insights from these data to develop approaches that complement other tools, such as focus groups, interviews and surveys, to gain a holistic understanding of community perspectives. Moreover, this work specifically tests the capacity to develop an AI-based pipeline to generate meaningful topic models representing community concerns.

This paper reports on findings from one of the focal communities examined as part of CRESP's efforts: Aiken, SC proximal to the Savannah River Site. In addition, this work reports on research from one commercial nuclear facility: Red Wing, MN around the Prairie Island Nuclear Generating Plant. The combination of these two locations represents a broad set of discussions and community concerns for analysis and provides a base for testing analytical approaches. Indeed, the growth of AI technology presents an opportunity to advance topic modeling technology, and to more rapidly model and understand community perspectives through existing digital traces. Questions remain, however, regarding the efficacy of these models. This work tests AI and non-AI based workflows for examining community media and compares the outcomes of these models to inform future tool development.

MODELING TO UNDERSTAND NUCLEAR WASTE DISCUSSIONS

Prior research examining collaboration-based siting (CBS) has largely focused on technical design approaches when engaging with external stakeholders [4], such as integrated design process(es) [5]. Recently, however, practitioners and researchers alike have recognized the need to focus on public perceptions and attitudes toward nuclear waste management and CBS [6]. Social media and news media data provide a robust source of information on public perceptions and attitudes, but the breadth and depth of data sources often mean that rapid large-scale analysis is challenging and time consuming.

In recent years, much has been made of the potential for topic modeling to be used to advance understanding of public discussions regarding a wide range of controversial or important topics. These approaches have

the potential to enable rapid monitoring of public attitudes and perceptions regarding the siting process spent nuclear fuel (SNF) storage facilities. In both news media and social media, a variety of topic modeling approaches have been leveraged to better understand the themes emerging in media coverages and discussion – as well as to understand the nature of discussion on social media platforms.

For example, Sweitzer and Gunda [1] demonstrated the potential for utilizing natural language processing to facilitate topic modeling and to help identify conversations most relevant to CBS processes. In an analysis of social media content, focused on Twitter data, researchers were able to develop an AI-based autolabelling approach [2]; in their work, the authors used 126 million tweets to tune their modeling efforts and develop a more accurate large-language model approach to topic identification. The study was then applied to data representing all 50 states to show that there is generally neutral or positive sentiment towards nuclear power at the time of the analysis. The ramifications of this type of research are notable. For instance, in Japan in the wake of the Fukushima disaster, there has been a renewed emphasis on the need for clear communication with the public grounded in science and reflective of public communication priorities. Researchers found that a co-occurrence analysis of word pairs used in YouTube videos on nuclear energy could effectively shape understand of public sentiment and information communication strategies [3]. The same type of analysis can be conducted in the context of analyzing questions asked in public comment periods [4]; classic approaches to topic modeling such as STM (Structure Topic Model) can help to summarize large batches of comments. More advanced approaches such as large-language models have the potential to reveal significant insight into the types of information being sought through these question-asking periods.

Testing Approaches to Modeling Discussions of Nuclear Energy

In general, modeling topic-specific issues related to scientific or complex information can be challenging. Traditionally, approaches such as latent Dirichlet allocation (LDA) have been used to generate coherent topic models as output [11]; this type of approach uses probabilistic models to identify coherent subgroups. But LDA models can struggle with complex language, especially when certain terms are utilized with a relatively low frequency in a text [5]. More advanced models today can overcome prior limitations by better understanding underlying sentence structure. For instance, prior work on legal texts note that the complexity of legal judgements and the specificity of language means that topic models often need to be tuned in order for researchers to effectively identify salient topics [6]; advanced models are more capable of handling this type of tuning.

Thus, this research study set out to evaluate emerging approaches to topic modeling of news and discussion (via social media content) related to nuclear energy and nuclear waste storage. Building on prior research, we focused on two leading approaches to topic modeling that are generally accepted as best practices in computational social science domains: BERTopic and Large language models (LLMs) models. BERTopic models focus on identifying clusters in the text of documents to identify common themes. Unlike previous approaches to topic modeling, BERTopic does not rely on counts of words alone, but leverages natural language processing to look at terms in relationship to one another. In recent years, advances with BERTopic allow for the incorporation of some aspects of LLMs to improve the identification of clusters – this approach is generally referenced as BERTopic AI [7].

The rapid development of large language models and AI-assisted data analysis has opened new avenues of analysis with regards to topic modeling. A recent comparative study found that large language topic models perform particularly well in terms of being able to produce meaningful and diverse topics from large textual datasets [8]. LLMs for topic modeling rely on self-supervised machine learning models that ingest large amounts of text to develop advanced models of sentence structure and meaning. LLMs used for topic modeling can thus be more effective in terms of extracting clusters of related terms and identifying underlying sentence structure; in turn, this enables these types of models to be used to more effectively identifying topic clusters.

Therefore, this work specifically compares BERTopic against more recent large-language models for topic modeling. Some prior research suggests that there may be gains made by using newer LLM models for analyzing complex information [9], but there is no research to date applying this work specifically to the discussion of nuclear energy or nuclear waste. Better understanding the appropriateness of model application to discussions of nuclear energy amongst stakeholders has clear implications for future research and practice. Effectively being able to summarize large corpuses of news data or social media content can help practitioners to rapidly assess shifts in public sentiment, or to identify emerging discussion topics.

DATA

Data were extracted from Facebook groups where discussions occur regarding community issues, with a particular emphasis on discussions regarding radioactive waste and nuclear energy. In addition, local news articles were retrieved from media outlets within a radius of each of the focal communities included in this research.

For the purposes of this analysis, two subsets of data were utilized. The first dataset comes from a Nuclear Energy University Program (NEUP) project focused on commercial SNF storage. A set of coherent news articles from 2024 were used as a test set for understanding coverage of the Prairie Island Nuclear Generating Plant in Red Wing, MN. This facility was selected as a commercial nuclear facility located in a community where there is a complex set of engagement issues, including an active Tribal population, concerns regarding water rights, and long-term questions about ongoing nuclear operations.

To create a contained test set, news articles were retrieved from Newsbank, an online aggregator and database of full-text news articles in the United States and globally. There were around 40 relevant articles retrieved covering 2024. These were found by searching the broad keywords of "Prairie Island" and "nuclear" and limiting results to news outlets within a 150-mile radius covering Minnesota and two neighboring states (Iowa and Wisconsin).

The second set of data comes from work performed in support of the DOE's CBS process. A second dataset was used to test topic modeling of social media data. The dataset for social media data came from an ongoing research project examining stakeholder engagement in the context of CBS processes regarding the storage of SNF. In the radius around the DOE's Savannah River Site, 51 Facebook groups were identified with topics or discussions directly relevant to nuclear energy or the Savannah River Site, and 112,066 individual content posts were retrieved. Data were retrieved via a third-part application, Apify, and stored in a database structure.

Model Testing

The research team tested four different primary models; BERTopic, BERTopic+AI, GPT-4 and GPT-5. The model testing was conducted using a set of custom scripts written in Python. The scripts created two sets of topic models – the first using BERTopic and the second using OpenAI's GPT application programming interface (API). For BERTopic, two models were constructed – a baseline model and a second model using AI assistance to identify topics. For OpenAI, two models were constructed, the first using the GPT 4.1 model and the second using the GPT 5.1 model. These are essentially different iterations of large language modeling as OpenAI continues to refine their technology.

Dictionary Development

A dictionary is not required for modeling generally, especially for LLM. Nevertheless, dictionary development can help to guide model development. Thus, as part of the research effort, the team built a baseline dictionary for identifying domain-specific content related to SNF. The dictionary was built from 25 domain-specific PDF reports that focus on US SNF storage, management, and policy: this included reports from relevant stakeholder groups, including a DOE overview of DOE-managed SNF, the 2013

DOE strategy for SNF and high-level waste disposal, a Nuclear Waste Technical Review Board paper on DOE SNF management, the 2021 Government Accountability Office report on commercial SNF, and a 2023analysis of US SNF policy. Each PDF was read using Python scripting and converted to raw text using an automated PDF parser. All pages were combined into a single text stream per document so that term frequencies could be tracked at both the document and corpus levels.

Documents and relevant websites were pre-processed, text was lower-cased, non-alphanumeric characters removed, and text was tokenized. Common stop words were removed. Term extraction was then conducted to surface domain phrases such as "spent nuclear fuel," "geologic repository," and "collaboration-based siting." To keep the dictionary focused and manageable, the top 200 unigrams (single words), top 200 bigrams (word pairs), and top 150 trigrams (word triplets) were selected based on raw frequency. The resulting comma-separated values (CSV) provided a domain-specific dictionary for topic identification which was then used to support topic identification in the analysis.

Model Comparison

The final step was model comparison. The comparison was made based on topic coverage, entropy, topic diversity and topic coherence. UMass and C_v coherence scores are included but these are only available for BERTopic and cannot be calculated for LLM models. UMass measures topic coherence for each independent topic; optimal scores are closer to zero. C_v is a second measure of topic coherence, and a measure closer to one is optimal. UMass often performs better with larger datasets.

Broadly, coverage, entropy, diversity and coherence provide for the best comparison. Coverage measures how well the topics capture the full range of topics represented in the dataset. Entropy captures to what degree topics are distributed within a specific document. Topic diversity measures the degree to which topic models capture the breadth of a corpus of document while minimizing “junk” topics that lack meaning. Finally, coherence evaluate how semantically related words within a topic are – thus accounting for the degree to which a topic has meaning.

RESULTS

The following summarizes the results of the model comparison for the social media and local news media datasets related to discussions of nuclear energy and SNF. First, a summary of the model analysis using BERTopic models and GPT models is given, but the focus is placed on the comparison of BERTopic versus GPT output.

The results for the model analysis are broken up into a discussion regarding the news media data and the social media data. For each, we consider the most dominant topics; the goal here was the evaluate the effectiveness of each modeling approach, and thus this content provides context for understanding those results.

News Media Topics

We consider two specific sub-topics that were identified in the BERTopic modeling. The first topic discussed nuclear plant outages and incidents. The first topic was derived from six news articles. The language from BERTopic identified the topic using the keywords “nuclear plant” and “incident”; the BERTopic+AI variant refined this to “nuclear plant outages and incidents”. When run through the OpenAI GPT 4.1 and 5.1 models, the code instructed the API to provide a sentence summary of the topic. In this case, for OpenAI GPT 4.1, the output described the topic as “articles reporting about nuclear plant outages and incidents.” A similar discussion was provided by the GPT 5.1 variant. Discussions in the topic involved nuclear-related themes such as outage, incident, shutdown. Discussions often appear alongside such as cable, repair, Prairie Island. This topic highlights nuclear operations, safety, waste, or energy-system impacts.

The second topic discussed nuclear waste and storage expansion. The language from BERTopic identified the topic using the keywords "nuclear waste" and "storage"; the BERTopic+AI variant refined this to "nuclear waste and storage". For OpenAI GPT 4.1, the output described the topic as "articles reporting on nuclear waste and storage of nuclear waste." A similar discussion was provided by the GPT 5.1 variant. The topic was derived from five news articles. This topic involves nuclear-related themes such as waste, storage, expansion. It often appears in discussion alongside terms like spent fuel, environmental, Prairie Island.

Social Media Topics

We consider two specific sub-topics derived from the social media analysis. The first topic focused on broad discussions related to the Vogtle nuclear power plant. The first topic was derived from approximately 243 social media posts. The language from BERTopic identified the topic using the keywords "Vogtle", "nuclear power," and "plant"; the BERTopic+AI variant refined this to "nuclear power plant at Vogtle". For OpenAI GPT 4.1, the output described the topic as "discussions related to the Vogtle nuclear power plant." GPT 5.1's description was largely consistent. Discussions within the topic related to nuclear energy themes reflected in terms such as vogtle, plant vogtle, Georgia power. Discussions include vogtle unit and power.

The second topic was derived from approximately 34 social media posts. The language from BERTopic identified the topic using the keywords "nuclear apprenticeship"; the BERTopic+AI variant refined this to "a nuclear apprenticeship program". For OpenAI GPT 4.1, the output described the topic as "training and apprenticeship in nuclear operations." GPT 5.1 added Aiken to the topic sentence discussion. Discussions in this topic related to nuclear energy themes reflected terms such as production operator, apprenticeship and nuclear fundamentals. The topic appears in discussions involving Aiken, technical, operator and apprenticeship program. Overall, the topic clusters around nuclear-sector issues and training. The topic highlights the capacity of the modeling to identify significant but relatively niche topics.

Model Comparison

Figure 1 shows a comparison of key metrics for BERTopic, BERTopic AI, GTP-4 and GPT-5 for analysis of news documents. Overall, the model shows that the GPT variants perform significantly better with regards to coverage and diversity of topic. Further, GPT coherence was much stronger compared to the BERTopic models. Entropy was slightly lower for the GPT models but not significantly so. Overall, this suggests that there are notable gains in using artificial intelligence supported topic modeling to capture the breadth and depth of topics represented in news media content.

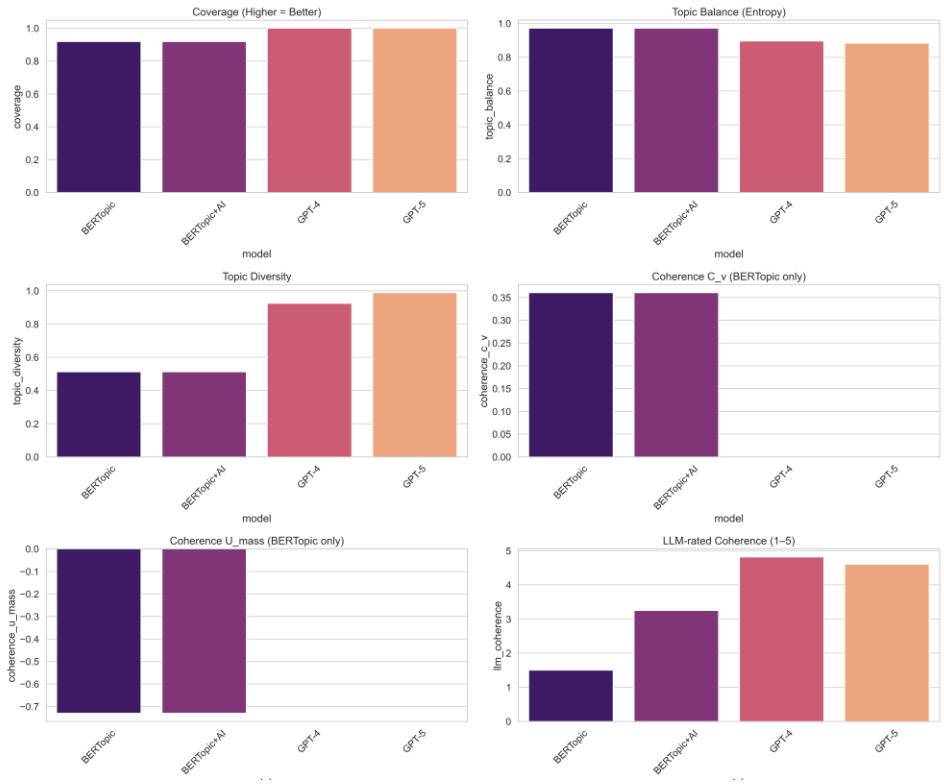

Figure 1. Summary of Model Comparison for Local News Analysis.

Figure 2 presents the same model comparison for the social media data collected and analyzed in this research. The data for this analysis covers a much larger breadth of topics, and the unit of analysis is much more fragmented (sentences and phrases shared on social media).

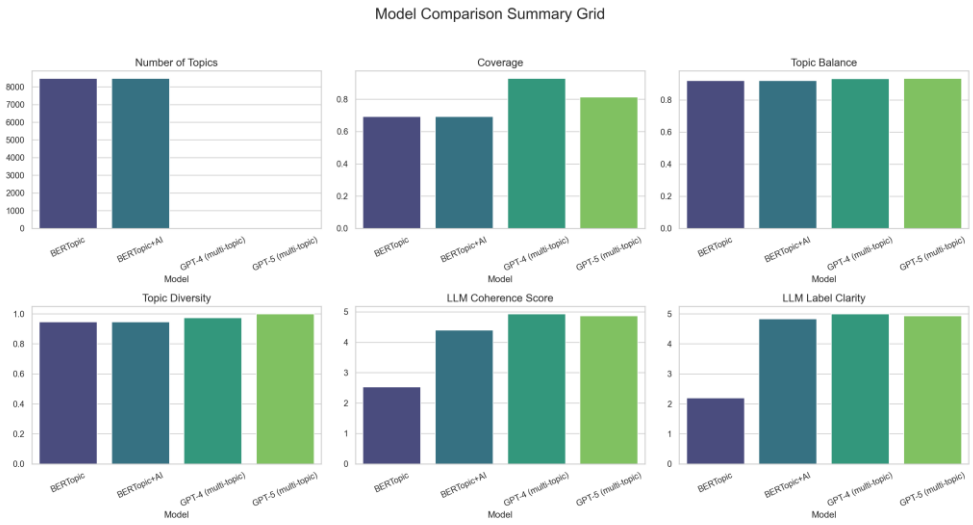

Figure 2. Summary of Model Comparison for Social Media Analysis.

The topic model shows the same scores for coverage, topic balance, diversity and coherence. The figures show that for social media content, the BERTopic models perform well in terms of topic balance and diversity, and generally well in terms of coverage. That said, the AI-based models (GPT 4.1 and 5.1)

perform much stronger in terms of coherence and coverage. In this case, the BERTopic models are strong but there are gains in using the GPT models for topic modeling and topic labelling.

DISCUSSION

The results show that in each case the GPT models can produce more robust and coherent interpretations using an Application Programming Interface (API) and fine-tuning of the models. Notably, while not discussed at length in this in this analysis, the addition of the dictionary text into the model framework did help with topic identification. Further, the results also point to the critical importance of using AI in tandem with additional data sources; if the AI model output is used in isolation it can easily lead to false positives or falsely attributed significance. Our findings suggest a workflow that incorporates multiple data streams with an AI interface to generate meaning from topic model outputs.

The results of this research demonstrate the potential utility of developing an AI pipeline for analyzing content extracted from social media channels and local news to understand community perspectives on radioactive waste management. Our work also shows that AI-supported models can be used to quickly and effectively generate insights into community attitudes towards issues related to nuclear energy and the storage of SNF. It is worth noting that topic modeling for social media content remains challenging due to the fragmented nature of the text and the variety of topics discussed in social media discussions. BERTopic performed well in this regard, but the AI-based models were somewhat more effective in identifying topics. There is a cost, however, to leveraging AI-based models. The BERTopic models can often be run locally or with comparatively less computing power. The AI-based models require API access and there is a cost associated with each model that is run (the cost varies based on the amount of content and the number of model iterations).

Findings from these models can be used to inform one another; initial models from BERTopic can support refining later model development using AI-supported analyses [9]. As the workflow illustrated above shows, the initial BERTopic models were useful in identifying keywords and broad topic descriptions. The AI-based models refined the initial approach and provided additional explanation and context. Moreover, this work shows that AI-assisted topic modeling has the potential to assist in the quick detection of topics in social media and news media to better understand community concerns and topics of importance.

REFERENCES

[1] M. D. Sweitzer and T. Gunda, “Spatiotemporal Analyses of News Media Coverage on ‘Nuclear Waste’: A Natural Language Processing Approach,” *Nucl. Technol.*, vol. 210, no. 9, pp. 1706–1721, Sept. 2024, doi: 10.1080/00295450.2023.2229566.

[2] O. H. Kwon *et al.*, “Sentiment analysis of the United States public support of nuclear power on social media using large language models,” *Renew. Sustain. Energy Rev.*, vol. 200, p. 114570, Aug. 2024, doi: 10.1016/j.rser.2024.114570.

[3] Y. Sun, H. Tsuruta, M. Kumagai, and K. Kurosaki, “YouTube-based topic modeling and large language model sentiment analysis of Japanese online discourse on nuclear energy,” *J. Nucl. Sci. Technol.*, vol. 62, no. 11, pp. 1038–1050, Nov. 2025, doi: 10.1080/00223131.2025.2522354.

[4] M. D. Sweitzer, D. N. Sanchez, A. N. Rogers, T. Gunda, L. A. Drakopulos, and T. C. LaForce, “Thematic Shifts in Discussion and Questions About Consent-Based Siting,” Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2024.

[5] D. Maier *et al.*, “Applying LDA topic modeling in communication research: Toward a valid and reliable methodology,” in *Computational methods for communication science*, Routledge, 2021, pp. 13–38.

[6] A. Zadgaonkar and A. J. Agrawal, "An Approach for Analyzing Unstructured Text Data Using Topic Modeling Techniques for Efficient Information Extraction," *New Gener. Comput.*, vol. 42, no. 1, pp. 109–134, Mar. 2024, doi: 10.1007/s00354-023-00230-5.
[7] G. B. B. Gomes, R. Attux, and C. C. Cruz, "Enhancing Contributions to Brazilian Social Media Analysis Based on Topic Modeling with Native BERT Models," *J. Inf. Data Manag.*, vol. 16, no. 1, pp. 181–191, 2025.
[8] M. B. Meram, Ç. Kalkan, T. Çelikten, and A. Onan, "GPT vs. Other Large Language Models for Topic Modeling: A Comprehensive Comparison," *ICCK Trans. Emerg. Top. Artif. Intell.*, vol. 2, no. 3, pp. 116–130, 2025.
[9] O. El-Gayar, M. Al-Ramahi, A. Wahbeh, T. Nasralah, and A. Elnoshokaty, "A comparative analysis of the interpretability of lda and llm for topic modeling: The case of healthcare apps," 2024.
[10] N. Aamir *et al.*, "Topic Modeling Empowered by a Deep Learning Framework Integrating BERTopic, XLM-R, and GPT," *J. Comput. Biomed. Inform.*, vol. 8, no. 02, 2025.

ACKNOWLEDGMENTS

This material is based upon work supported under Department of Energy, Office of Nuclear Energy, Award DE-NE0009342, "Moving Toward Collaboration-Based Siting of a Federal CISF [CISF]: Improving Strategies for Community-Informed Decision-Making during Stakeholder and Tribal Engagements." Thus, this report was prepared as an account of work sponsored by an Agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof.

The report is also partially based on work supported by the U. S. Department of Energy, under Cooperative Agreement Number DE-FC01-06EW07053 entitled 'The Consortium for Risk Evaluation with Stakeholder Participation III' awarded to Vanderbilt University. The opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily represent the views of the Department of Energy or Vanderbilt University.