Analyzing Public Feedback on Environmental Policies Using Natural Language Processing: The PHATE of Climate Narratives

A. Bhatt, E. DeMott, E. Fountain, A. Tulpule, Z. Wang, S. Dork, M. Berijanian, and N. Brugnone

Abstract
Public feedback on environmental policies plays a crucial role in shaping regulatory decisions. This paper presents an analysis of public comments submitted to the Federal Register regarding environmental policies using Natural Language Processing (NLP) techniques. Our approach includes text scraping and the application of PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding), originally developed for visualizing high-dimensional biological data, to natural language tasks. We leverage PHATE for visualization and hierarchical clustering to support exploratory data analysis. This work provides insights into the public's perception of environmental policies and their suggested alternatives, offering valuable guidance for policymakers.

1. Introduction
Environmental policies are increasingly influenced by public opinion, as regulatory bodies seek to balance economic, social, and ecological considerations. However, analyzing large volumes of public comments presents a significant challenge. In this study, we employ NLP techniques to extract and analyze thematic trends from public feedback on climate change adaptation and mitigation measures. By adapting PHATE from biological data analysis to textual data, we introduce a novel approach to uncover hierarchical structures and latent relationships in public comments, offering policymakers a new perspective on citizen input.

1.1 Data Source

The Federal Register serves as the official daily publication of the U.S. government, containing proposed and final regulations, public notices, and executive orders. It provides a comprehensive record of all governmental actions, making it an invaluable resource for monitoring and understanding policy developments. One of its most notable features is the public comment period, during which citizens, organizations, and experts can provide feedback on proposed rules and regulations. This public comment process allows anyone to post, which can influence policy decisions especially in areas such as environmental protection. With millions of public comments available on a wide range of issues, the Federal Register offers a rich dataset for analysis. The comments reflect not only public sentiment but also detailed insights into the concerns, priorities, and perspectives of different stakeholders. This wealth of raw, unstructured data presents the opportunity to extract valuable trends and patterns in public opinion on climate related issues.

2. Related Work

Natural Language Processing has been increasingly leveraged in the analysis of environmental policy and public discourse. Multiple studies have highlighted the potential of NLP pipelines for policy research, particularly in reducing the traditionally laborious process of policy analysis. Planas et al. [5] introduced a modular NLP framework for environmental policy analysis that employs BERT models to extract topics and construct a knowledge graph from scraped federal documents. This mirrors our effort in building a scalable pipeline to analyze public comments on climate policy, though our work diverges by integrating dimension-reducing visualizations and hierarchical clustering.

A noteworthy contribution by Swarnakar and Modi highlights the potential of NLP in shaping comprehensive climate policies by synthesizing diverse public narratives and expert discourses from both structured and unstructured data sources [2]. Their work underscores the transformative power of NLP in addressing climate change, leveraging key techniques such as topic modeling, opinion mining, discourse network analysis, and knowledge graphs. These methods provide valuable insights that can inform policymakers on public sentiment and policy directions. While their focus is broader and more policy-actor-centric, their emphasis on topic modeling and sentiment extraction aligns closely with our objective of analyzing public feedback trends.

Previous research has also demonstrated the viability of topic modeling approaches, such as Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP), for parsing ecological policy documents. Altaweel et al. applied these methods to identify evolving themes in governmental responses to mountain pine beetle outbreaks, thereby establishing the effectiveness of unsupervised models in environmental policy analysis [1]. Similarly, Hagen et al. employed topic modeling to identify emergent citizen-driven themes from petitions on the "We the People" e-petition platform, underlining the utility of automated methods in extracting structured insight from unstructured feedback [3].

Before this project, sentiment analysis has already demonstrated its value in the context of government policy. Corallo et al. presented an optimized sentiment classification approach for analyzing public opinions on government initiatives expressed on Twitter. Using a simpler model—Naive Bayes—compared to the more complex methods in this project, the study achieved 78% accuracy with a dataset containing thousands of tweets [2]. This highlights the importance of integrating emotional components into the analysis of public comment datasets. It also shows the potential of this project, incorporating millions of federal comments and using more complex models.

On the visualization front, the original PHATE method by Moon et al. [5] has been shown to preserve both local and global nonlinear structures in high-dimensional biological data. Its application to transcriptomic and microbiome datasets inspired our adaptation of PHATE to

textual data, enabling nuanced visual analysis of latent comment structures. Our study extends PHATE's utility into the NLP domain, offering policymakers a new tool for visual exploratory data analysis of public sentiment.

Lastly, the broad applicability of machine learning to climate change-related efforts was emphasized in Rolnick et al.'s call to action for the AI community [6]. While their work is more strategic and conceptual, it stresses the urgency of developing data-driven tools like ours to enhance environmental governance.

3. Methods

3.1 Data Collection

Public comments were sourced from the Federal Register via Regulations.gov, which serves as the centralized platform for accessing U.S. federal rulemaking documents and corresponding public input. The data on Regulations.gov is organized hierarchically: each federal agency maintains a collection of *dockets*, with each docket containing related policy documents, proposed rules, and associated public comments.

The platform hosts over 22 million public comments, with approximately 2.2 million pertaining specifically to climate-related topics—this subset formed the basis of our study. Initially, a custom Python-based API client was developed to retrieve and store comments in a MongoDB database. However, the API's rate limit of 1,000 requests per hour per API key presented a significant bottleneck, making it infeasible to gather a large enough dataset within the project's timeline.

Further complicating the process, a substantial number of public comments are submitted as file attachments (e.g., PDFs or Word documents), which require additional processing to extract readable text, adding considerable overhead to the pipeline.

To overcome these limitations, an open-source initiative called **Mirrulations** [4] was leveraged, which mirrors data from Regulations.gov. Mirrulations extracts and processes the content of attached documents, and makes the cleaned data publicly available via an Amazon Web Services (AWS) S3 bucket. Although S3 storage is not inherently searchable or optimized for direct access, it offers a scalable and efficient source of preprocessed data suitable for our needs. To integrate this data with our infrastructure, a Python script was created that downloaded records from the public S3 bucket and imported them into a MongoDB server hosted on the Michigan State University High Performance Computing Center (HPCC). This script was run multiple times to collect the full dataset. Once populated, the database allowed for efficient querying and streamlined downstream analysis.

3.2 Data Preprocessing

Once the raw comments were collected into the MongoDB database, a multi-stage cleaning and pre-processing pipeline to prepare for the subsequent natural language processing tasks was implemented. The original data collection in our MongoDB server consisted largely of noise. Duplicate entries, existence of html entities from the extracted text, and boilerplate messages directing to attachments to name a few. Thus, the raw text data underwent a two-stage cleaning process: initial filtering and text cleaning.

First usable records were extracted from a .json file of all the raw comments in our Mirrulations database in MongoDB server into a .csv file. A python script was created to skip empty or duplicated comment entries which retained metadata attached to comments. The columns, besides comments, were agency ID, posted date, docket ID, comment ID, document ID, and attachment URL. The extracted comments were further cleaned to improve consistency and reduce noise, shortlisted to include only the Environmental Protection Agency (EPA). For readability, HTML tags, special characters, URLs, Unicode characters, stopwords, excessive whitespace, retained alphanumeric text, and common punctuation were removed. Our threshold for filtering comments for better BERTopic clustering and analyses was to discard comments shorter than 10 characters and restrict our analysis to those submitted after the year 2000.

To support BERTopic modeling further and address the input length limitations of transformer-based models, long comments were transformed. Abundant in our dataset, a head-tail truncation method was utilized to deal with this. This approach is informed by the findings of Sun et al., who demonstrated that the method outperforms other truncation strategies (such as head-only and tail-only) in text classification tasks [9]. Their study suggests that critical information in documents often resides at both the beginning and end, making the aforementioned approach effective for preserving sentimental context within the shortened comment text data. The fully cleaned and processed dataset was then serialized into a compressed format to facilitate efficient loading for subsequent analyses.

3.3 BERT Embeddings and Sentiment Analysis

The next goal was to apply a modeling pipeline where semantic embedding generation and sentiment analysis were created. First, a dense vector embedding for each comment was generated using the Sentence-BERT model all-MiniLM-L6-v2. This is a lightweight pre-trained transformer model optimized for semantic similarity tasks and large amounts of data. Comments were tokenized, encoded in batches for memory efficiency, and stored alongside their respective processed comment.

From there, BERTopic was implemented. BERTopic is a modeling framework that clusters documents based on semantic embeddings. BERTopic uses UMAP for dimensionality reduction and HDBSCAN for clustering. This is useful as noisy high-dimensional language data can be

clustered into coherent topic groups. Each comment is assigned a topic label and topic name as well as a confidence probability of the model's certainty in its classification of the comment to a topic group.

To augment the topic modeling with emotional context, we conducted sentiment analysis using the pre-trained Hugging Face model cardiffnlp/twitter-roberta-base-sentiment. This model was selected for its robust performance on the public comments. Each comment's sentiment score was normalized to a continuous scale between -1 (strongly negative and disapproving) and +1 (strongly positive and affirmative). Discrete labels were also added to the comments, divided into three categories: -1 (negative), 0 (neutral), and +1 (positive).

Thus, the final dataset, in addition to containing processed comments and their metadata, also contains semantic embeddings, assigned topic and topic probability, sentiment score, and labels.

3.4 PHATE for Textual Data

Potential of Heat-diffusion for Affinity-based Transition Embedding, abbreviated PHATE, was primarily employed to visualize the relationships between themes, capturing complex structures in the data. By modeling high-dimensional textual data, PHATE provided an interpretable representation of latent structures and transitions in public discourse with the vision to allow policymakers to explore feedback dynamically. As referenced earlier, the original PHATE method efficiently maintained both local and global nonlinear structures in high-dimensional biological data. Motivated by its strengths, we adapted PHATE for textual data to facilitate visual analysis of latent comment structures. Although PHATE is a relatively new dimensionality reduction method, it proved more effective for our purposes than traditional techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (**t-**SNE), and Uniform Manifold Approximation and Projection (UMAP). As demonstrated below, PHATE offers a compelling alternative for visualizing complex textual relationships.

3.5 Visualization

The approaches taken to visualize the data consisted of generating embeddings for the comment text data, extracting the topics by clustering the comments, and finally using a variety of dimensionality-reducing visualization methods such as UMAP, PCA, t-SNE, and our primary tool of focus- PHATE- to observe the embeddings with topic information encoded with color. This produces visualizations that enable one to compare and contrast the way the same semantic information is captured and transformed by various visualization techniques. Additionally, supplementary visualizations techniques such as frequency of comments of a given topic over time, overall sentiment distributions, and sentiment by docket were produced.

4. Results

Shown here are five key graphics produced throughout the project. These visuals tell the story of the trends extracted from the data and different ways to visualize the data. Firstly, the chart below compares and contrasts the different dimensionality-reducing visualization techniques with 500 samples of comments. The algorithm applied moving from left to right gets increasingly complex, but subsequently increasingly more useful at showing patterns in the data. The PHATE visualization produced a geometrically distinct visualization of the embeddings compared to the other techniques displayed. This is because PHATE did more than simply clustering the data as shown by PCA, t-SNE, and UMAP, but also preserved the global and local structure of the data.

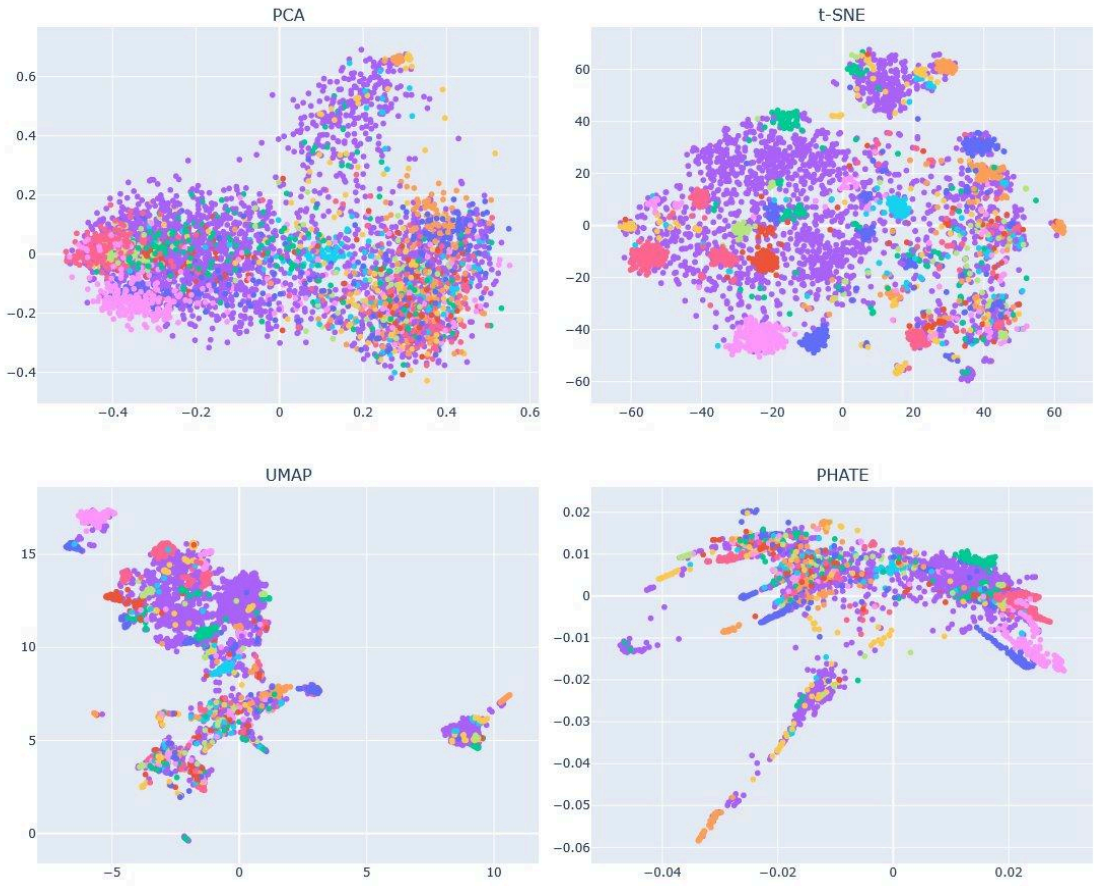

Figure 1: Comparison of different dimensionality-reducing visualization techniques: PCA, t-SNE, UMAP, and PHATE with 5000 samples

Once determined that PHATE was the superior method of data visualization, this figure was produced to visualize how clusters of textual data interact with each other. This contains the top 50 most occurring topics in the data and the visualization produces branching, streak-like

clusters. The distance between clusters may yield valuable information for the observer as it shows which topics are related to each other according to the model embeddings. This chart is able to reveal relationships between topics not necessarily obvious to someone reading through the raw comments of each federal document.

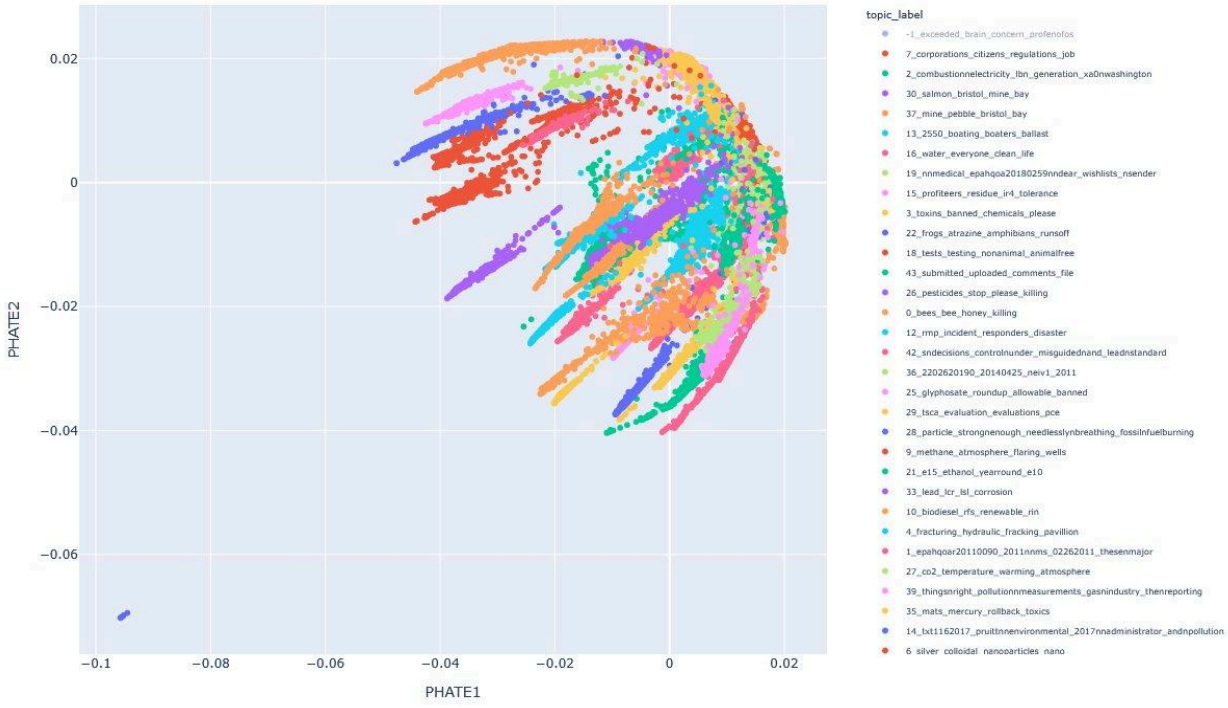

Figure 2: PHATE Visualization of top 50 topics

Figure 3 shows how the frequency of comments for each topic changes over time, highlighting evolving trends in public discussion. The most frequently discussed topic overall is related to bees, though it remains relatively consistent without major spikes. In contrast, distinct surges are observed in 2009, 2011, and 2019, corresponding to increased discussion around nanosilver, petroleum refineries, and asbestos, respectively—indicating heightened public attention during those periods.

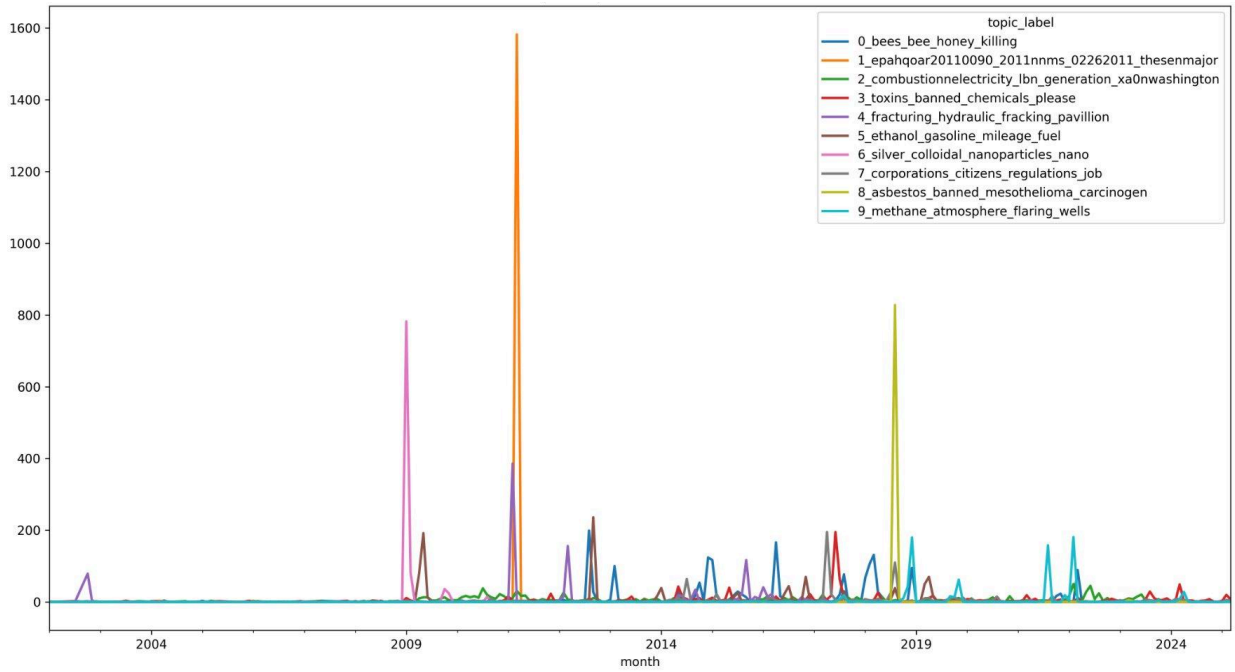

Figure 3: Frequency of Top 10 Topics over Time

Looking at the dataset as a whole, the histogram below displays the sentiment distribution of the comments, revealing a strong skew toward negative sentiment**.** This aligns with the well-documented negativity bias in psychology, where individuals are more likely to express dissatisfaction or concern than to offer praise. As shown, comments with positive sentiment are significantly less frequent, suggesting that people are more compelled to speak up when they are critical or disapproving.

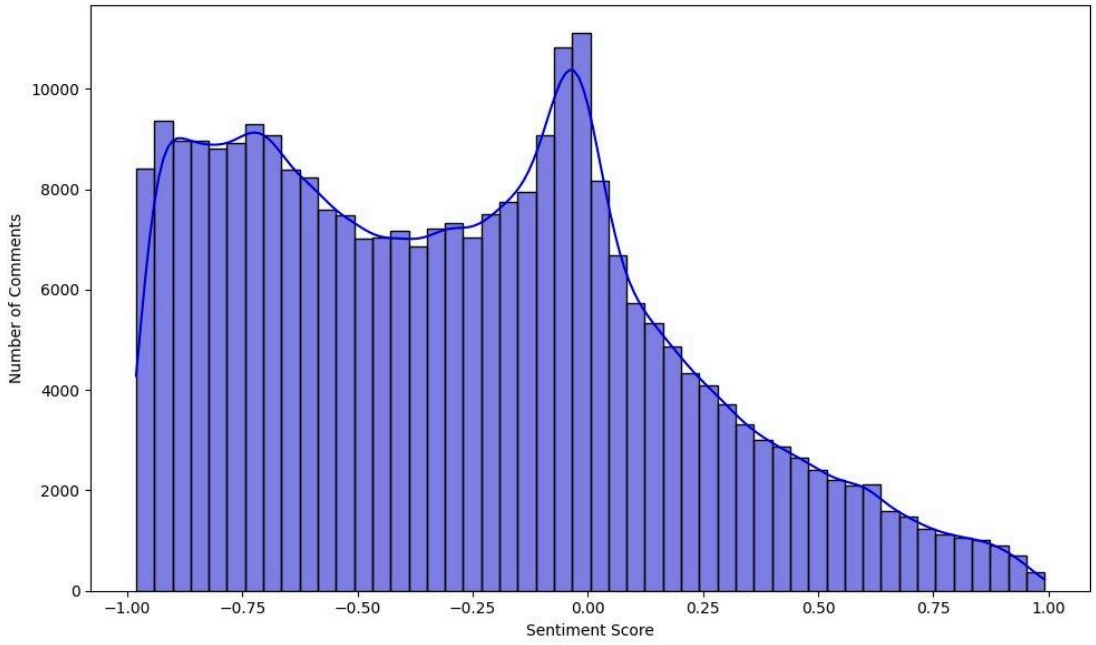

Figure 4: Histogram of overall comment sentiment from negative (-1) to positive (+1)

Zooming into the top ten most frequent Environmental Protection Agency dockets, the bar chart in Figure 5 shows the relative sentiment distribution of these comments. It should be noted that the group of three dockets with great negative sentiment all belong to dockets covering public health. The second docket from the right covers regulatory reform evaluations, and all of the other dockets cover fossils and emissions.

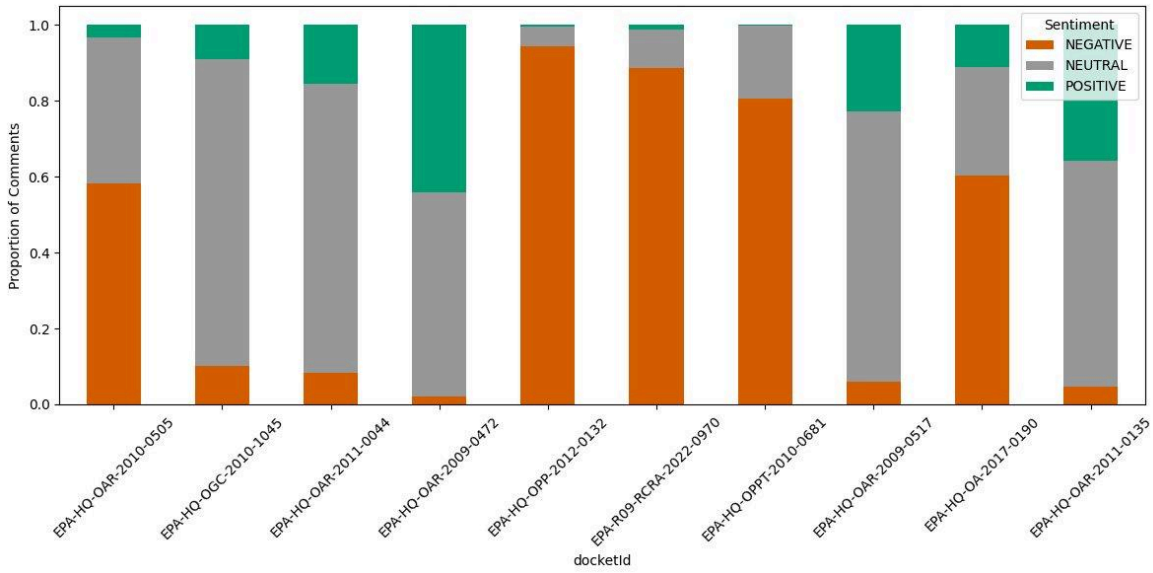

Figure 5: Sentiment proportions on top 10 EPA dockets

5. Discussion

This research displays that the integration of natural language processing techniques with PHATE and other visualizations provides an effective framework for analyzing large-scale public feedback on environmental policies. By preserving both local and global structures within the data, PHATE enables the identification of nuanced thematic relationships that traditional methods may overlook. Our analysis also reveals a strong tendency toward negative sentiment in public comments, highlighting the critical importance of addressing public concerns in the regulatory process.

An interesting finding not previously discussed was the presence of climate denialism throughout the collected data. While negative opinions on environmental policies were expected, it was surprising to observe numerous comments outright rejecting the existence of climate change, often without engaging with the specific policy in question. These comments were typically treated as noise, as they did not contribute meaningful insights into public sentiment toward individual dockets. Although this project did not specifically analyze which topics attracted the most climate denialism, this could be a valuable direction for future research. Investigating whether climate denialism has been a consistent phenomenon or if it emerged at a particular point in time would offer historical context into this finding.

For future work, there are many avenues to pursue. The most important is to expand the existing dataset to include all comments available in the Federal Register. Since our current data only contains EPA-related comments, expanding to all comments would greatly enrich the sentiment analysis. In addition, incorporating older comments would be valuable. Our current database only includes comments after 2000, as those were the easiest to collect, but adding earlier

comments could reveal how sentiments and public discourse have evolved. A challenge with this would be dealing with handwritten comments and accurately extracting text from PDFs. Another future direction would be further refining the sentiment analysis models and adopting emerging large language models (LLMs) to enhance the accuracy, depth, and interpretability of insights drawn from this public discourse. There are continually bigger and better models for sentiment analysis and future work could revolve purely around testing different models and fine-tuning with the data collected. Additionally, building and training custom sentiment analysis models tailored specifically to public comment data could be an avenue for further improving the model's clustering performance and extracting more nuanced insights.

6. Conclusion

This study demonstrates the potential of NLP in analyzing public comments on environmental policies. By systematically extracting themes and employing PHATE for exploratory data analysis, our approach enhances the transparency and inclusivity of policymaking.

Future work will focus on expanding the dataset to include the full Federal Register, fine-tuning existing sentiment analysis models, experimenting with newer LLMs for improved sentiment classification, and generating more refined visualizations using PHATE or other dimensionality reduction techniques.

Acknowledgments

This work was conducted in collaboration with Two Six Technologies and the Applied Economics Office of the Engineering Laboratory (EL) at NIST. We gratefully acknowledge Professor Dirk Colbry from the CMSE 495 course at Michigan State University for his guidance and support throughout the project.

References

1. Altaweel, M., Bone, C., & Abrams, J. (2019). Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances. *Ecological Informatics, 51*, 82. https://doi.org/10.1016/j.ecoinf.2019.02.014
2. Corallo, A., Fortunato, L., Matera, M., Alessi, M., Camillò, A., Chetta, V., Giangreco, E., & Storelli, D. (2015). Sentiment analysis for government: An optimized approach. In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 98–112). Springer International Publishing. https://doi.org/10.1007/978-3-319-21024-7_7
3. Hagen, L., Uzuner, Ö., Kotfila, C., Harrison, T. M., & Lamanna, D. (2015). Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In *2015 48th Hawaii International*

Conference on System Sciences (pp. 2134–2143). IEEE. https://doi.org/10.1109/HICSS.2015.257

4. Mirrulations. (n.d.). *Mirrulations/Mirrulations* [Computer software]. GitHub. https://github.com/mirrulations/mirrulations
5. Moon, K. R., Van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., & others. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology, 37*(12), 1482–1492.
6. Planas, J., Firebanks-Quevedo, D., Naydenova, G., Sharma, R., Taylor, C., Buckingham, K., & Fang, R. (2022). Beyond modeling: NLP pipeline for efficient environmental policy analysis. *arXiv*. https://arxiv.org/abs/2201.07105
7. Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2022). Tackling climate change with machine learning. *ACM Computing Surveys (CSUR), 55*(2), Article 42, 1–96. https://doi.org/10.1145/3485128
8. Swarnakar, P., & Modi, A. (2021). NLP for climate policy: Creating a knowledge platform for holistic and effective climate action. *arXiv*. https://arxiv.org/abs/2105.05621
9. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics* (pp. 194-206). Cham: Springer International Publishing.
10. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.