# Exploring Pollution Levels and Media Trends Between States: A Study of New York Times News Reports

## Data Science I Final Project

TIAN TONG, WENDY SHI, IRENE CHEN

## 1. Introduction

The role of news media in shaping public opinion has been a well-discussed topic by scholars. Murukutla (2019) categorized media as the "watchdog" that guides the audience to weigh the severity of health and environmental problems and encourages means for risk prevention. Furthermore, numerous scholarly evidence exists of the positive effect of news in pushing for political and social changes (Van Zoonen, 1992). Murukutla (2019) has found the topic of pollution is commonly discussed in the economy, national defense, and politics. Therefore, the media coverage of pollution has underrepresented the negative health impact in shaping public discourse and forming perceptions.

In the digital information era, the general public turns to online news platforms on mobile electronic devices seeking news reports. The easily accessible data online makes research on how digital media shapes public perception possible. With the New York Times API and US news ranking on pollution, our research found no significant relationship between higher pollutant levels and higher media coverage in terms of the absolute number of news reports they receive and the mean sentiment level of each article. This means states with higher pollutants deserve more media attention in raising awareness of the potential health risks.

Pollution Level | Media Coverage | New York Times | Textual Analysis | Sentiment Analysis | Similarity Test | Bigram Analysis

## 2. Methodology

### *2.1. Data collection and index selection*

In this paper, we plan to focus on one of the biggest and most reputable online media platforms in the United States, the New York Times. We want to explore whether media coverage accurately reflects pollution levels in each state. Using textual analysis techniques, we also want to investigate whether the focus on pollution varies among states and evaluate the potential factors that attract higher media coverage.

We used the New York Times Article search API as our based API to obtain all news articles under the search query of pollution in a set time range. Its response field is the same as the Archive API, which returns NYT articles for a given month, with data traced back to 1851. Each article represents a pollution incident in a specific state within the selected time range. Due to the limited article that we found when setting the time parameters to be below five years, we expanded our time limit and collected all news available by the state in the time frame from January 1, 2000, to December 1, 2023. The time restriction and geolocation are set based on a filtering command of `begin_date`, `end_date`, and `'fq': f'glocations:({i})'`."

Using data wrangling techniques in Python, we created a function that yielded the content of the first article collected in a single API pulling, looping through all 50 states. We located 31 states with data available for this research by observing if we pulled empty data. Next, we manually pulled all news available for each state that contains information such as published date, abstract, titles, URL, etc for each article. The final working data set is the combination of all data

from 31 states.

We picked the US news ranking on pollution to measure pollution levels. The US news pollution level index is calculated based on the mean performance on reducing chemical pollutants and lowering risks for long-term chronic diseases. The US News ranks all 50 from 1 to 50, with 1 representing the best states in handling pollutants and health risks to the worst.

### *2.2. Data summary with visualization*

For better visualization, we first created a geo plot (figure 1) that represents the US news ranking, the index we used to measure the pollution level. The color is scaled from 1 to 50, with one representing a lighter color, which matches directly to the index of pollution level from the best to the worst. Therefore, states ranked lower in the US news are associated with a darker color. As we can see from the graph, states in the middle tend to produce more chemical toxins and have a higher risk for chronic diseases compared to states in the coastal area.

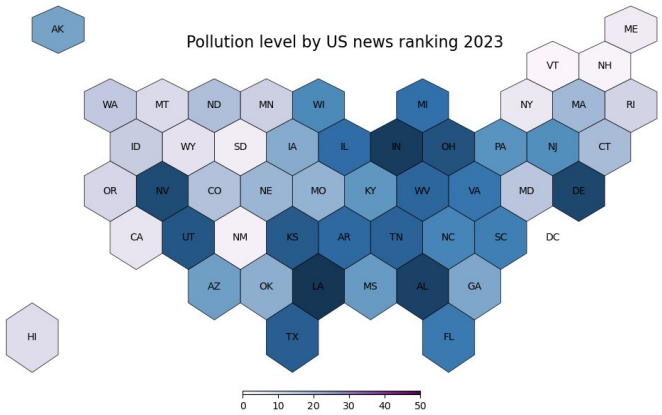


Fig. 1. Geo plot of real pollution level by states

For comparison, we also created a geo plot (Figure 2) with color representing the total number of articles found. The color is scaled from 0 to 100 to match the total number of articles found. Therefore, states that do not have news articles available and states with limited news articles will be in lighter colors, states with incidents that attract more media coverage will be in darker colors.

We found that most states have a total number of articles between 0 and 100. However, there are a total of five states that contain more than 100 articles in the past 20 years. They are California with 599 articles, New Jersey with 342 articles, Connecticut with 149 articles, Texas with 113 articles, and Alaska with 105 articles.[1]

Comparing Figure 1 with Figure 2, we found evidence of a discrepancy between pollution levels and the number of reports on pollution available for each state. For example, Alaska and California do relatively well in emission control, but the number of pollution incidents in these states is significantly higher compared to other states. On the other hand, in states such as Delaware, Indiana, and Louisiana, where pollution level is high, the number of news reports available is limited to non-existing. Pollution levels and media coverage do not always correlate for many states, meaning higher pollutants and health risks for each state do not necessarily lead to higher media coverage.

[1]See Figure 11

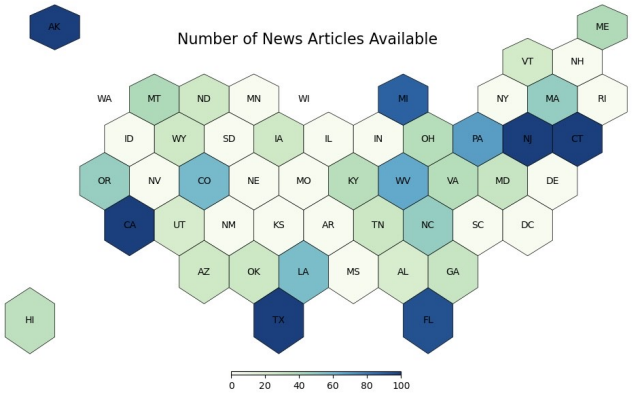


Fig. 2. Geo plot for the total number of news articles published

## 3. Analysis

### *3.1. Semantic analysis based on sentiment score*

For semantic-based sentiment analysis, we used the transformers model, the pre-trained language model that serves as the core of many AI language models such as Chat-GPT. For every sentence, the language model will return a numerical number ranging from -1 to 1, representing negative and positive sentiment. To evaluate the sentiment of each article, we apply the transformer model to every abstract of all news articles available. Next, we calculated the mean sentiment score by state. We saved the mean sentiment score in a separate data sheet with the state ranking on pollution level and the total number of articles found for comparison and data visualization.

Figure 3 is the geo plot of the mean sentiment score by states. The mean sentiment score in our dataset ranges from -1 to 0, meaning the news report on pollution for all states appears to be negative. Therefore, we constructed the color scale from -1 to 1 to match the mean sentiment score of each state. The darker the color, the more negative NYT portrait pollution will be.

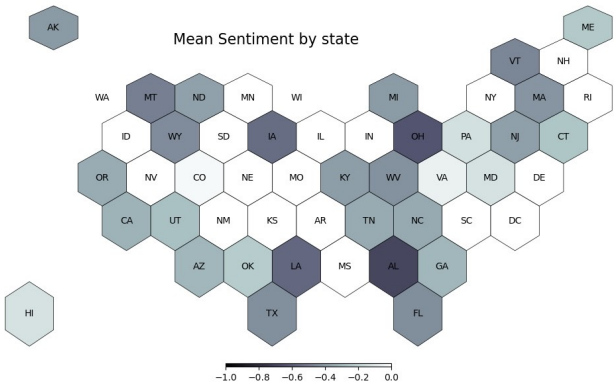


Fig. 3. Geo plot for mean sentiment score by state

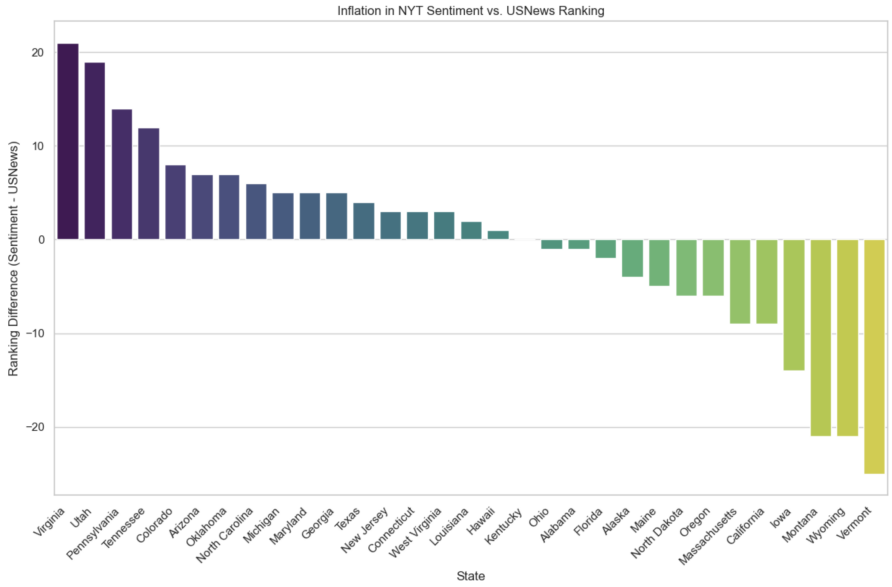


Fig. 4. Inflation in NYT Sentiment for each state

Comparing the sentiment score in Figure 3 with the pollution level in Figure 3, we did not observe a correlation between higher pollution levels and a more negative sentiment level. New York Times media coverage does not precisely match pollution levels for each state. On the contrary, states that do relatively better in emission control receive a more negative sentiment score, with the example of Oregon and California on the West Coast, Maine, Vermont, and Massachusetts on the East Coast. All states mentioned seem to have a democratic political orientation. These patterns also raised the question of whether sentiment score is associated with party interest. In the future, it may be possible to explore how party interests directly or indirectly result in insufficient news coverage in states with higher pollution levels.

### *3.2. Similarity test within grouped states*

To further explore the possible reasons behind the stark discrepancy between the U.S. News Ranking a state received and its sentiment score based on New York Times articles, we ran separate cosine similarity tests for states in groups to see whether there exists any patterns or keywords that would systematically bias the New York Times reports.

We rearranged the U.S. News Ranking to obtain a relative rank within these 31 states. Then we subtracted the updated 31-state U.S. News Ranking from their sentiment rankings to build a new index, which we will temporarily refer to as “Inflation in NYT Sentiment”. This number represents the difference between the sentiment rankings a state achieved and its U.S. News Ranking. A positive number indicates that compared to a relatively lower rank in the index built off from scientific measures in U.S.News, the NYT’s reports about this state are more neutral and mild in tone. A negative number would imply that a state receives unfairly harsh comments and heavy criticism that are disproportionate to their real-life status. A greater magnitude in the absolute value of this index points to a larger mismatch between the state’s true affairs and the NYT reports.

Figure 4 displays each state’s “Inflation in NYT Sentiment” index in descending order. The states on the left side of the spectrum receive inflated sentiment ranking in NYT reports, while

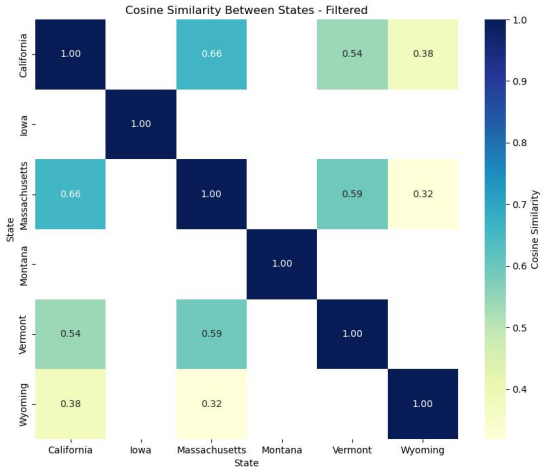


Fig. 5. Cosine Similarity between states with the lowest Inflation in NYT

those on the right suffer unduly critical comments compared to their ranking in U.S.News. For example, Vermont ranks 1st in U.S. News rating for their environmental protection and pollution control, topping the list. But a -27 in this index suggests that based on the sentiment rating of NYT articles, Vermont would in fact be among the states that receive the most negative comments regarding pollution.

No clear connection was found when comparing the states at far ends of the spectrum in terms of geography, economy, or politics. Therefore, we conducted distinct similarity tests for the states that positioned on the leftist side and rightest side of this graph, in an attempt to explore possible similarities in their NYT coverage. What specific language or themes trigger excessively negative coverage? In other words, we seek to understand the underlying similarities among states at the far ends that could systematically bias the NYT articles, making them overly harsh on some issues while inadequately tolerant towards others.

Based on the top 20 bi-grams we extracted for each state,[2] we first ran a cosine similarity test for the six states at the right-hand side of the spectrum, and we found a strong connection within this group that received the heaviest criticism in comparison to their relatively well-managed pollution control. (See Figure 5) We set a threshold of 0.35 to filter out pairs of states that share more than 35 percent similarity in their top 20 bi-grams for clearer comparison. A more detailed performance of bigram analysis would be conducted in the following section.

We then created a combined barplot to show the bigrams that these states have in common. (See Figure 12) We limited the number of bigrams to the most frequent ten for better visualization. Here, we can see that the word-pairs that are shared the most are “global warming”, “carbon dioxide”, “climate change”, “bush administration”, “power plants” and “electric cars”, four of which point to the theme of global warming and use of new energy.[3]

However, when running a similar test for states on the leftist side (who might have received news coverage that is too mild in tone in contrast to their low rating in U.S.News), we did not

[2]We opted for a test based on bigrams instead of single words because, in the context of pollution, high-frequency words mostly come in pairs (such as Carbon Dioxide, Climate Change, and Global Warming). If we used the single highest-frequency words, the results would be diluted by words that only appear together.

[3]See Figure 12

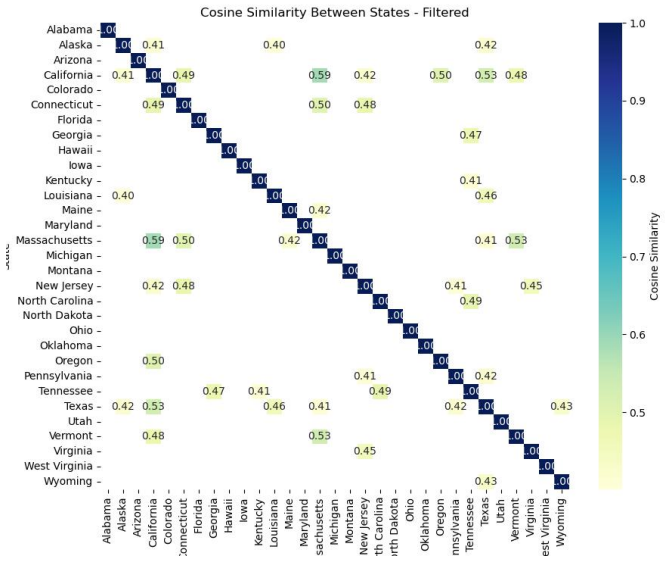


Fig. 6. Cosine Similarity between all 31 states (filtered by >0.4)

detect many similarities for the top 10 bigrams in these states. [4] "Fresh kills" and "kills closed" are likely referring to a certain one-time event or incident; and even though "natural gas" and "power plant" are highly prominent in two of the states, they're not commonly shared by others.[5]

Furthermore, a closer look at the similarities in top bigrams across all states suggests that much of the similarities could be attributed to common words that would normally appear in a news article about pollution (such as "power plant", and "carbon dioxide"). For example, in Figure 6 we display the cosine similarity between all 31 states in our dataset (setting a threshold of 0.4 to show only strong correlations), and it's not difficult to find that some of the states that are positioned at the leftist side on the spectrum in Figure 4 also shares a lot of commonalities with states that are on the right-hand side.[6] This suggests that possible similarities we detected within the groups at one end of the spectrum could be inflated by these common words.

To see the specific word-pairs that are in common use for all 31 states, we took out a sample of California (the state with the largest sample size, 599 articles in total) and the states that it shared more than 50 percent of similarity with. A larger sample size boosts precision; and a threshold of 0.5 filters out the words that are most heavily used. Figure 7 shows a combined bar plot of the top bigrams in these states, in which we can see that "global warming", "climate change", and "carbon dioxide" continue to loom large in this sample group - a feature that we earlier detected in the six states with the lowest Inflation in NYT. Therefore, we should avoid overstating the influence of these word-pairs (mostly pointing to the theme of climate change) on the disproportionate heavy criticism from New York Times articles.

It's tempting, therefore, to conclude that at this point, no common feature is shared by the states that jointly receive NYT commentaries that don't match their actual progress in pollution

[4] See Figure 13

[5] It is necessary to note that most of the states on the leftist side don't have a large number of news articles as samples, which limits the capacity for a similarity test.

[6] See Figure 10

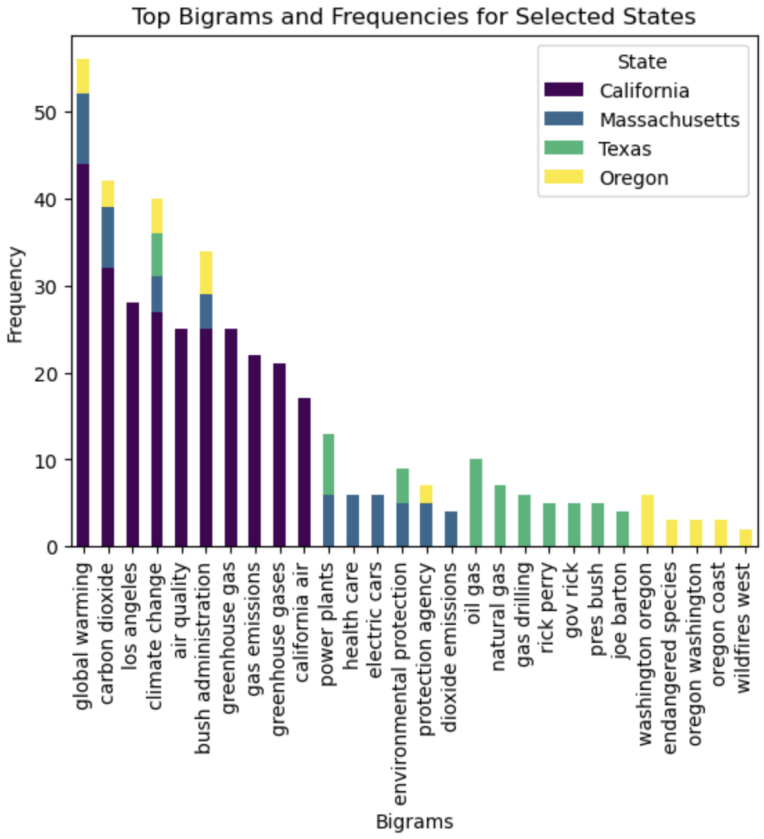


Fig. 7. Combined top bigrams for selected sample states

treatment and environmental protection, because the states that receive inflated sentiment scores don't have a lot in common in their top bigrams, and the states that do share some commonalities seem to inherit much of these shared concerns from a national trend of emphasizing the topic of climate change and new energy.

However, it is necessary to note that key differences still exist between the states at the two ends of the spectrum in Figure 4 . For example, none of the states that receive unduly heavy criticisms are traditionally heavy-industrial states. Moreover, a large proportion of these states are notable for their pioneering status in the science and technology sector. The implications of these commonalities are further discussed in Sector 5: Limitation and Future Research Direction.

### *3.3. Bigram analysis of sample states*

In exploring the potential relationship between higher pollution levels and higher media coverage, our research extended more to the specific factors that capture more media. For this purpose, we conducted a bigram analysis on our dataset sourced from the New York Times, focusing on the articles under the search query 'pollution' within a given time frame. Our analysis prioritizes three sample states to be representative: New Jersey, Connecticut, and Alaska, out of the five states with over 100 articles collected. The approach is designed to draw insights into both the generalized and specific characteristics that influence the media reporting on pollution significantly.

Under the methodology, the primary focus is to examine the 20 most frequent words for each of the 5 states. The process involves the extraction and visualization of bigrams - pairs of consecutive words - from the 'Abstract' column of the collected articles. We need to understand the patterns of the dominant bigrams in each state to conclude the key factors.

Before the extraction process of bigrams, we performed the common pre-processing steps including tokenization, normalization, stop word removal, and stemming. Except for the general stopwords, we also exclude meaningless words such as 'said', 'says', and 'since' from bigrams like 'official says'. At the same time, we noticed certain bigrams like "York New", "Jersey Connecticut", and "York Connecticut", which originated from the listing of state names such as "New York, New Jersey, Connecticut". To enhance the clarity of plotting, we adjusted our approach by excluding these repeated state names for a more accurate representation. Another interesting observation was that we noticed a high frequency of "New Jersey " within the "Connecticut " bigram plot. This finding lead us to conduct a comparison analysis between the two states, emphasizing our interest in similarities in environmental journalism

### 3.3.1. Alaska

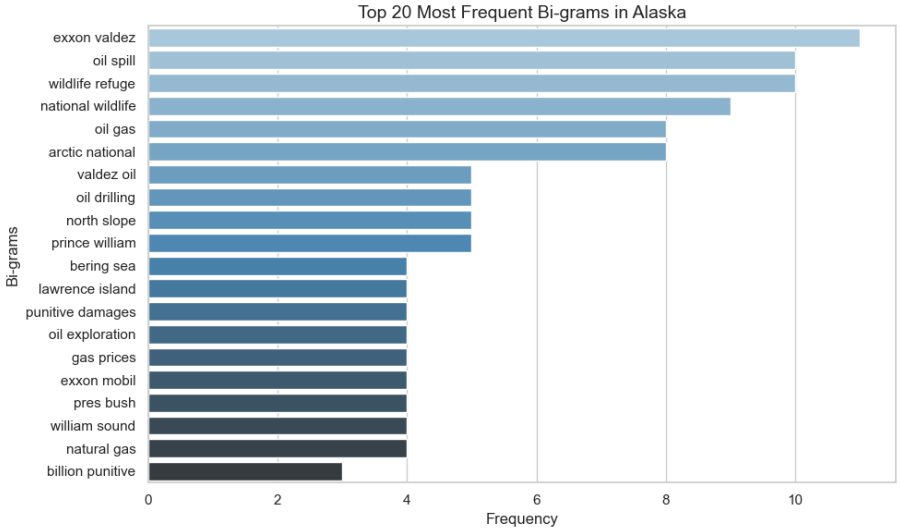


Fig. 8. Top 20 bigrams of Alaska

Our examination of the bigram plot for Alaska provided a distinct perspective of a dominant historical factor compared to other states. The plot revealed a significant emphasis on the 'Exxon Valdez' oil spill incident and its associated words.

On 24 March 1989, the oil tanker 'Exxon Valdez' spilled 260,000 barrels of crude oil in Prince William Sound, Alaska. The incident has maintained its popularity as a heated topic and consistent presence in media reporting over the years with a constant number of reports news each year. This suggests that the consequences are enduring and present rather than merely historical for Alaska and the surrounding regions. The prominence is further reinforced by the correlated words including the geographic references such as 'Prince William', 'Lawrence Island', and 'William sound', alongside 'Exxon Mobil', the responsible entity company, as the reminder of public awareness of a must-taken social responsibility. Moreover, the 'national wildlife' bigram pointed to a broader concern for Alaska's ecosystems. Following the incident, effects were observed across a wide range of habitats and species. Habit damage resulting from oil contamination is underestimated (Peterson, 2001), demanding attention and action.

The media's consistent focus, particularly regarding the oil spill incident, coupled with the economically significant factors, highlighted by 'fossil fuels' and 'natural gas', illustrated Alaska's role as a major energy producer. The continual media spotlight reflected a contemporary societal commitment to addressing long-standing challenges posed by such disasters. It intensified the imperative for more effective policy reforms and more robust environmental protections.

### 3.3.2. New Jersey and Connecticut

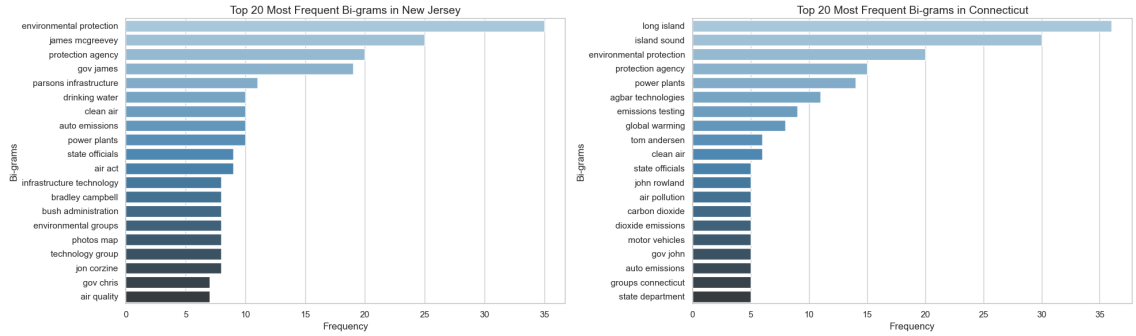


Fig. 9. Combined top 20 bigrams of New Jersey and Connecticut

In our analysis with the 342 articles from New Jersey, a notable predominance of the term 'New Jersey' was observed throughout the frequency distribution over 200 times.It is followed by 'New York', mentioned about 60 times. Such high occurrences of region names in state-focused articles were not surprising. Consequently, these words would overshadow other significant but less frequent bigrams, potentially skewing the interpretative value of the data visualization.

In the analysis conducted for Connecticut, a similar trend was shown. 'New York' appears in over a third of the articles, with about 40 mentions. To discover the most relevant bigrams efficiently, we decided to exclude the state names to highlight other bigrams that could offer deeper insights into the thematic focus of the research question, moving beyond the location references.

Firstly, focusing on the New Jersey text corpus, two environmental concerns of 'air quality' and 'water supply' are clear. Also, we noticed outstanding occurrences of terms related to high-level government officials, such as 'James McGreevey(52$^{nd}$ governor of New Jersey)', ''Bradley Campbell(Commissioner of the New Jersey Department of Environmental Protection), 'Jin Corzine(56$^{th}$ governor)', as well as entities of 'protection agency' and 'environmental groups'. Under the leadership of James E. McGreevey, the long tradition of strong environmental protection has been restored. Significant attention has been given to the restoration of water resources and clean air, as evidenced by the enactment of the Pollution Prevention Act. Especially, the initiatives to preserve the Highlands region, a key contributor to the nationwide drinking water supply, were highlighted in a March-April newsletter from the New Jersey Department of Environmental Protection 2004.

Another rising focus revealed is technology. Residents in Northern New Jersey have been suffering from severe water pollution caused by flooding due to heavy precipitation and extreme weather events. Green infrastructure has emerged as a successful intervention method to mitigate these challenges, as detailed in Wieczerak's 2022 study. The presence of such bigrams indicated ongoing discussions aimed at leveraging technological advancements for environmental management, including infrastructure, working together with legislative efforts.

In the subsequent part of the analysis, we turned to the bigram of Connecticut. The primary

bigram ‘Long Island’ and LIS(Long Island Sound, shown as ‘island sound’ in the plot) captured the ecological and economic significance, recognized as one of the ‘Great Waters’ of the United States and a globally productive water body. The effluents from CT WPCF(Connecticut Water Pollution Control Facility) contributed to the majority of current and historical water pollution of Long Island. Given the reliance on the Sound for commercial, industrial, and recreational uses, the related discussions in media coverage were to be expected.

At the same time, the analysis suggests that air quality was a principal environmental concern in Connecticut. The Connecticut Emissions Testing Program states that about 1.3 million vehicles are tested per year to meet regulations imposed by the EPA (Environmental Protection Agency) and the Clean Air Act. This was reflected in the frequent appearance of bigrams associated with vehicles and carbon dioxide emissions, with the commonly mentioned air quality issue: global warming. Noteworthy within the distribution was the inclusion of ‘Agbar Technology’, a company that was reportedly suspended for failing to meet the CT Emissions Testing Program’s standards. Additionally, the intersection of state governance and environmental regulation was shown by the bigram ‘John Rowland($86^{th}$ Governor)’ in conjunction with ‘protection agency’, indicating the media’s interest in federal actions and their local implications.

To draw the similarities between New Jersey and Connecticut, we would start with their geographical identifiers. Coastal counties in the densely populated Northeast tri-state areas, including New York, New Jersey, and Connecticut, are pivotal not only in terms of population but also serve as a critical component of the world’s economic and financial activities (NOAA. State of the Coast and US Census Bureau 2013, Paterson et al. 2010). A closer examination of the overlapping articles revealed interconnected issues. Here, New York has been perceived as a benchmark for the Northeastern United States, while Connecticut’s location emphasized its vital relationship with Long Island Sound. Moreover, collaborative efforts to mitigate carbon dioxide emissions, facilitated by the partnership with the EPA, have led to a convergence in federal legislation and infrastructure development between the two states.

## 4. Conclusion

In conclusion, our research finds no significant relationship between higher pollutant levels and higher media coverage. Although states geographically located in the middle produced more toxins with higher chronic health risks, it does not receive more media attention measured by the absolute number of reports on pollution incidents and appropriate sentiment level. As a result, media portraits of incidents of pollution, represented by New York Times news, failed to guide the general public understanding of the potential health risks of pollution.

On the contrary, our research shows examples of states with better emission control receiving more media coverage and a more negative sentiment level, with the example of Oregon and California on the West Coast, Maine, Vermont, and Massachusetts on the East Coast. Future research can explore if higher media coverage leads to stronger social awareness or if it results in a stricter standard for emissions. Furthermore, all states mentioned have a democratic political tendency. Exploring the different roles of news media under different political ideologies is also critical in understanding means for advocating for social responsibility in reducing emissions.

Last but not least, our analysis reveals several potential factors that would attract more attention of media generated from the articles of states with larger volume. Prominently, historical incidents, especially whose with enduring impact until now, national disaster, the intersection of federal regulations and local implications, matters of economic and environmental significance. Additionally, geographical identifiers, cultural background also plays a crucial role, alongside the general global environmental issues. These findings indicate the complex interplay of both personal and natural factors in shaping media coverage, offering valuable insights into the dynamics of news prioritization.

## 5. Limitation and Future Research Direction

Under our methodology, the proxies that we picked to evaluate the pollution levels in the US New ranking are the emission level and health risks nationwide which may not capture the whole complexity of pollution's impact. For future research, we could generate our proxy in measuring pollution with actual data sets such as those detailing carbon dioxide emission, and water quality metrics, to provide a more refined and accurate reflection of pollution levels. Additionally, we could conduct a more segmented approach to our analysis by separating the pollution into distinct categories such as air and water, for a more nuanced understanding of each type's specific attributes and effects.

Although we performed our similarity analysis and bigram analysis based on the states with more articles than the remaining, as indicated by over 100 articles here, with only one source to collect, there is still not sufficient data to examine the frequent words for more general conclusions.

Furthermore, we recognized that our primary source, the New York Times, is often considered to be a media platform aligning with the Liberal perspectives and have long demonstrated leftward leaning ethos. Under the discoveries in the similarity tests and visualization of the spectrum, even though we detected no clear patterns for the states that receive milder coverage relative to their real-life pollution, we certainly find that the commonly mentioned "climate change" and "new energy" - environmental issues that are notably more global than local - feature heavily in the right side states. This could partly be explained by and/or further testify to the Liberal editorial stance of the New York Times and the publication's long-recognized tendency towards Liberal viewpoints.

For a more comprehensive and informative examination of media coverage on pollution-related issues, our future research could include media platforms with more diverse political orientations to explore if the conclusions drawn from our current research are consistent across different media platforms with varying political affiliations into a more comparative lens.

## References


1. Murukutla. N, Kumar. N, Mullin. S, "A review of media effects: Implications for media coverage of air pollution and cancer" leukemia **2**, 7–10 (2019)
2. Van Zoonen, E. A. "The women's movement and the media: Constructing a public identity." European Journal Communication **7**(4), 453–476 (1992)
3. Peterson, C. H. "The "Exxon Valdez" oil spill in Alaska: Acute, indirect and chronic effects on the ecosystem" In Advances Marine Biology pp. 1–103, (2001) https://doi.org/10.1016/s0065-2881(01)39008-9
4. Wieczerak, T., Lal, P., Witherell, B. et al. "Public preferences for green infrastructure improvements in Northern New Jersey: a discrete choice experiment approach." SN Soc Sci 2, 15(2022) https://doi.org/10.1007/s43545-022-00315-w

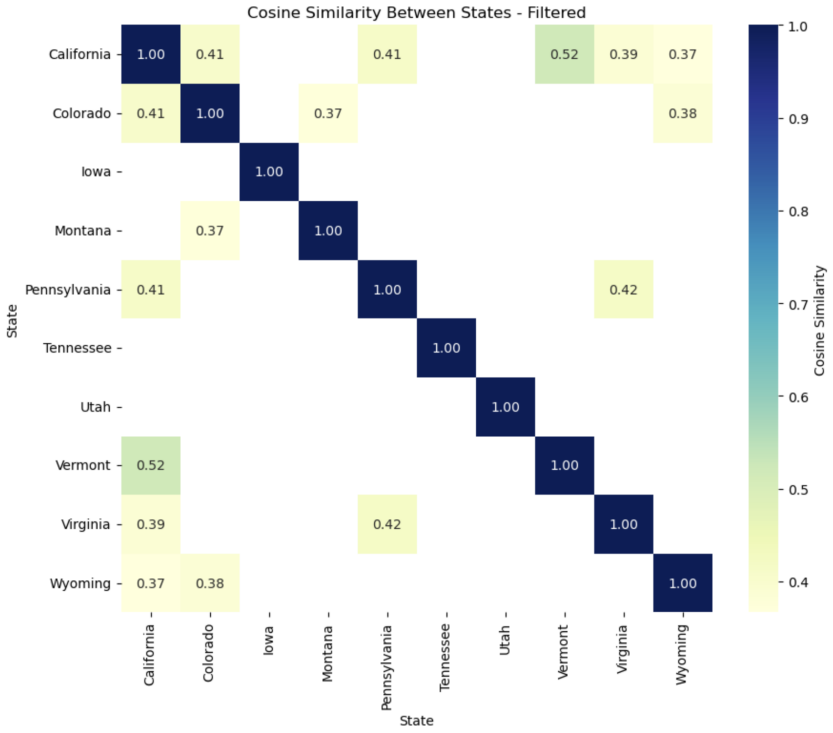


Fig. 10. Cosine Similarity for states at both ends of spectrum in Figure 4.

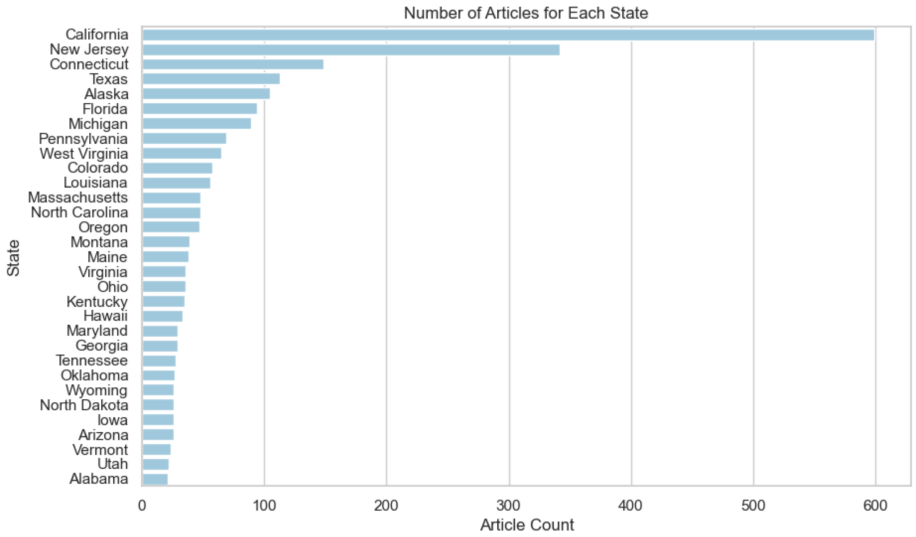


Fig. 11. The bar plot of the total number of articles available for each state.

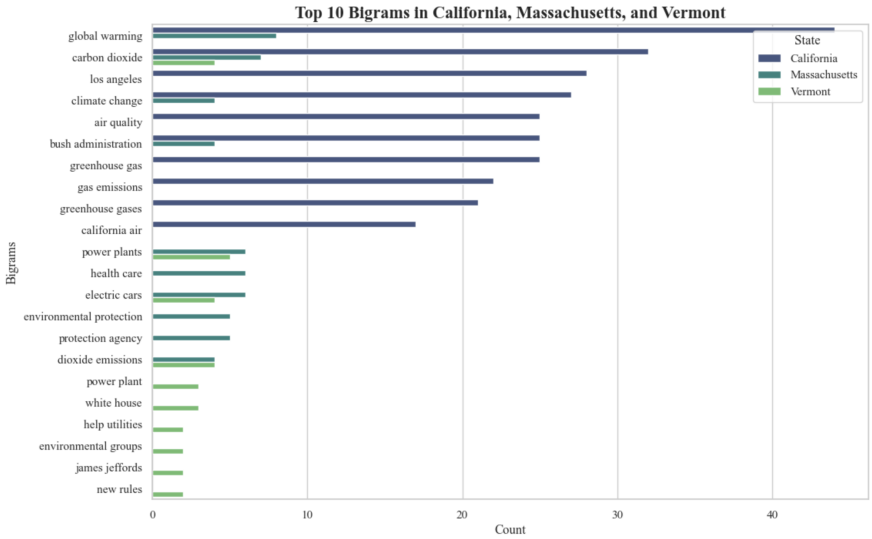


Fig. 12. Top 10 Bigrams in states with the lowest Inflation in NYT

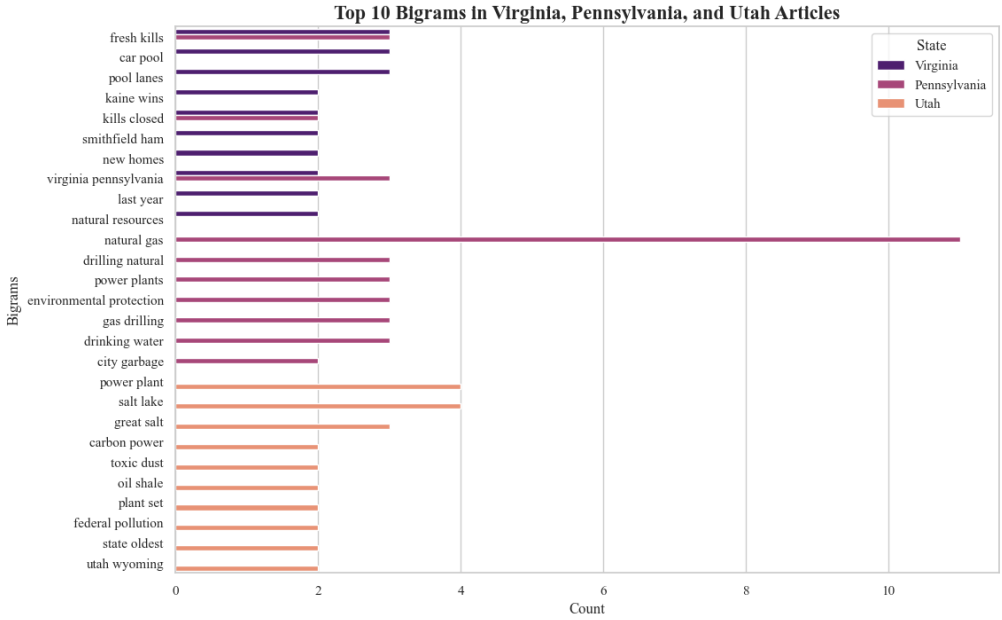


Fig. 13. Top 10 Bigrams in states with the highest Inflation in NYT