Text as Data Final Project

Tereza Petrovicova

December 2024

1 Research Question

1. What is your research question? Explain its significance. Have others looked at this question in the past? If so, what have they found? [We expect 4 paragraphs – two on the significance of the research question and two that review other's relevant work in this area with citations. The last few sentences should clearly state the contribution you hope to bring to the current state of the literature.]

My research question is: How can we measure firm exposure to climate and energy risk stemming from regulation? This question is important because understanding how firms are exposed to regulatory risks related to climate change and energy can provide valuable insights for investors, policymakers, and firms themselves. Climate and energy regulations are becoming more prominent as governments around the world implement stricter rules to combat climate change. Understanding the exposure of firms to these regulations is critical in assessing the financial and operational impacts they may face. Measuring this exposure effectively can inform investment decisions, risk assessments, and policy design, particularly as regulatory environments evolve to address climate challenges.

There is growing interest in using textual data from earnings calls and 10-K reports to analyze political and climate risk. A seminal paper by Hassan et al. (2019) examines political risk in earnings calls and validates their measure using 10-K reports. They find that political risk is better explained at the firm level rather than at the industry level, with over 90% of the variation in political risk attributed to firm-specific factors. This finding suggests that political and regulatory risks have a significant impact on firm-level dynamics, which is critical for assessing how climate and energy policies affect individual companies. Their work sets the foundation for further research into climate-related regulatory risk by suggesting that firm-level communication and disclosures are key to understanding a firm's exposure to political and regulatory changes.

Building on this foundation, Hassan et al. (2019) focus specifically on climate risk in earnings calls, with one of their key measures being climate regulatory risk. They find that corporate communication in earnings calls provides valuable insights into firms' climate risk exposures, emphasizing the importance of qualitative data in understanding the impact of environmental and regulatory changes on firm operations. This research lays the groundwork for using textual data to understand regulatory risk and highlights the role of earnings calls in analyzing climate exposure. However, 10-K filings offer a more formal and regulated source of information, and thus, for my research, I am shifting the focus from earnings calls to 10-K reports, as companies are legally required to disclose risks related to regulation and climate change in these filings.

My measure is based on a recent working paper by Baz et al. (2023) examines climate regulatory exposure using 10-K filings, investigating the relationship between climate regulatory risk and stock returns. They use dictionaries for climate change and government policy terms to measure the proximity of these terms in the text, creating a measure of climate regulatory exposure. However, the study uses a dictionary-based approach to quantify exposure, so my contribution is to improve upon it by incorporating more advanced text analysis methods. By leveraging cutting-edge text-as-data techniques, I hope to provide a more nuanced and accurate measure of firms' exposure to climate and energy regulation, enabling better insights into how these risks influence firm behavior and market outcomes.

2 Data

2. What data did you choose to address this research question? In what ways does this data shed light on the research question? What might its limitations be? [One paragraph description of the data, several plots or tables of relevant descriptive statistics about your data (e.g. how many documents? what are the length of the dos? tables of metadata? One paragraph discussion of value and limitations relevant to your question.]

Since I am analyzing climate regulatory discussions in 10-K filings, I intentionally selected firms within sectors likely to be rich in environmental regulation language. This study includes a sample of 149 companies, drawn from two NAICS sectors: 93 companies in Sector 21 (Mining, Quarrying, and Oil and Gas Extraction) and 56 companies in Sector 22 (Utilities). Although I gathered data from the past 10 years, I focused exclusively on 2024 10-K filings to optimize the processing efficiency. Environmental risk disclosure has become increasingly prominent in recent years, so I think that 2024 will be the richest year in environmental regulatory language.

The format of 10-K reports is standardized across firms by the SEC, which allowed me to narrow the scope of the analysis to *Part 1A: Risk Factors*. Each firm's filing includes a *Risk Factors* section, which, although typically spanning at least two pages, was the richest source of environmental regulatory content. Given that the section averages around 13,179 tokens per firm (see table 1), I decided to further subset this into 10-sentence chunks in order to facilitate hand-coding.

NAICS Sector	Number of Companies	Number of Chunks
21 (Mining, Quarrying, and Oil and Gas)	93	3,762
22 (Utilities)	56	1,651

Table 1: Summary of Data Sample

In my final dataset, I have 5,437 chunks, each approximately 10 sentences long, selected from the Risk Factors sections. Of these, I hand-labeled 270 annotated chunks as training data. After training the model, I applied it to 5,227 unlabeled chunks to classify them based on the presence of environmental or energy regulation. Initially, I reviewed several risk sections and developed a codebook, refining it after obtaining annotations for 50 chunks from a second coder.

There are several limitations to my dataset, primarily stemming from its focus on a subset of firms and the splitting of text into 10-sentence chunks. First, by only examining firms in two sectors, I risk creating a measure that may not generalize well to other industries, as the language used to describe climate regulation may vary depending on the specific environmental concerns faced by different sectors. Additionally, my analysis is limited to the Risk Factors sections of 10-K reports. While this section is often the most comprehensive source for environmental risk discussions, it is not the only place where references to environmental regulation may occur, potentially biasing the measure. Additionally, dividing the text into 10-sentence chunks can introduce issues by fragmenting context or splitting discussions in ways that make the data less coherent, which could impact the accuracy of my analysis, as well as raises concerns how to aggregate my measure up to the document level. Finally, focusing only on 2024 could raise some temporal issues if environmental regulatory discussion has evolved over time.

3 Creating a Measure

How did you go about using the data to create a measure to answer your research question? Why did you choose this approach? One-page (or however long necessary) step-by-step description of how you went about creating a measure with justification. Why does your particular test answer your research question and make you vulnerable to be proven wrong?

To answer my research question, I developed measures based on textual data from the Risk Factors sections of 10-K reports. The process began with a careful review of several 2023 10-K filings that were not included in my sample (in my sample I only looked at 2024). This allowed me to familiarize myself with the language and patterns used in the discussion of risks, particularly those related to environmental and energy regulation. I designed two key measures: `regulation_present` and `regulation_type`. These measures aim to identify and classify references to environmental or energy regulation in the text.

I first created a binary measure, `regulation_present`, to determine whether a discussion of environmental or energy regulation is present in a chunk of text. The second measure, `regulation_type`, categorizes the type of regulation discussed into environmental, energy, both, or neither. These variables allow for a nuanced analysis of how companies disclose regulatory risks in their filings.

This approach directly addresses my research question by leveraging the fact that firms are legally required to disclose material risks in their 10-K filings. If a company discusses environmental or energy regulation extensively, it suggests that they perceive these regulatory areas as significant risks. This method provides a clear way to quantify the extent of a firm's exposure to climate and energy regulation. However, there are limitations that make my findings vulnerable to potential criticism. For example, some firms may over-report these risks due to factors like preemptive compliance strategies, efforts to signal commitment to sustainability, or even just the strategic desire to highlight certain risks to stakeholders. Other firms may also under-report environmental or energy regulatory risks to avoid negative perceptions, reduce legal and financial liabilities, maintain flexibility, or due to uncertainty in the regulatory landscape. Furthermore, the level of discussion across firms may vary—some may provide detailed accounts while others offer only cursory mentions, potentially skewing the results. Lastly, my approach may miss nuanced or implicit discussions of regulatory risks, as it relies on explicit language and key terms. Therefore, while the measure provides a valuable first step in assessing regulatory exposure, it may not capture the full complexity of the risks discussed in these filings.

I improved the definitions after doing some inter-coder reliability tests, and the refined definitions are as follows:

1. **`regulation_present`**

- This binary variable captures the presence of either environmental or energy regulation specifically within the text. For an entry to be coded as "1," there must be an explicit reference to regulatory terms that are clearly related to environmental or energy concerns. General mentions of "governmental regulation" or similar non-specific terms are not included here. Examples of indicators include references to specific agencies or terms like "EPA regulations" or "energy compliance."

After conducting intercoder-reliability tests for `regulation_present`, I examined the discrepancies and updated my codebook. I found that some texts included statements like, "the actual or perceived health risks of handling hydraulic fracture sand." While these references may be relevant to energy firms, they are more closely related to health and safety concerns rather than environmental or energy regulation. Therefore, I excluded these instances from the analysis unless the regulation specifically addressed changes in environmental or energy policy. Similarly, health and safety regulations were not considered unless they directly impacted environmental or energy issues.

2. `regulation_type`

To distinguish between the types of regulation discussed, I created the `regulation_type` variable. This is a categorical variable that assigns a value based on the nature of the regulation described in the text:

- `1` for Environmental Regulation: Includes mentions of regulations or policies aimed at environmental protection, pollution control, emissions standards, clean water standards, etc.
- `2` for Energy Regulation: Includes policies governing the production, transmission, and distribution of energy, such as FERC rules or electricity rate standards.
- `3` for both Environmental and Energy Regulation: Assigned when both types of regulation are mentioned together in the text.
- `0` if neither type is mentioned or if the regulation is unclear in the context.

After hand-coding and intercoder reliability tests, I realized that it is quite difficult to disentangle energy regulation from environmental regulation, as energy regulation is often driven by environmental concerns. This overlap complicates the analysis, as regulatory discussions related to energy frequently include significant environmental considerations. Given this challenge, I decided to drop the `regulation_type` variable for further analysis, as it did not provide additional clarity and may have introduced unnecessary complexity in differentiating the two types of regulation. By removing this variable, I aim to focus more on the core measure of regulatory exposure without the complications of categorizing regulations into environmental or energy-specific types.

In my hand-coded dataset, I also included variables to capture the level of government regulation (`federal_regulation, state_regulation, local_regulation`), with the aim of understanding whether climate and energy regulations are being discussed at different levels of government. However, I found that in the case of large companies, there was very little mention of specific local regulations in the 10-K filings, which limited the usefulness of the local level variable. When it came to state and federal regulations, it was sometimes possible to discern whether the text referred to specific federal or state agencies. However, distinguishing between these two levels proved challenging because either there was no reference to a specific regulatory body, or the coder had to look up and figure out at which level the regulatory body operates, which is not always straightforwards. Additionally, my intercoder reliability was not high for these categories, indicating that the classification of regulation at the state and federal levels was subjective and prone to error.

4 Results

How well did you do at capturing your measure? What were the results of your analysis? [Some metrics and plots that show validation of your measure. Explain your results.].

4.1 Classification of Unlabeled Data

To classify my unlabeled data, I followed a series of preprocessing steps that included stemming the text, removing non-alphabetic characters, and eliminating stopwords. The classification process was carried out using a range of models, each with its own strengths and weaknesses. Figure 1 illustrates the performance of the models, showcasing key metrics such as accuracy, precision, and recall for each. In particular, precision and recall were evaluated with a focus on the "1" category, representing segments that discuss environmental or energy-related topics. This emphasis on the "1" category is crucial, as the primary goal of the analysis is to capture content that addresses critical areas related to environmental and energy regulations. By prioritizing recall, I aimed to maximize the identification of these important discussions, even at the cost of occasional false positives.

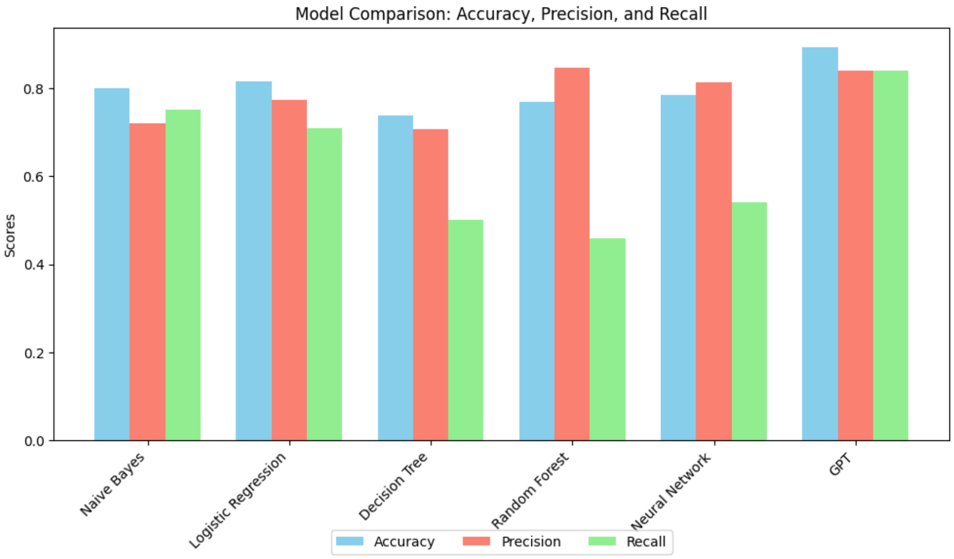

Figure 1: Model Comparison

Among the models tested, ChatGPT emerged as the most robust performer, excelling across all evaluation metrics, including precision, recall, and accuracy. The Naive Bayes model followed closely behind, demonstrating a solid performance, particularly in terms of recall, though it was slightly outpaced by ChatGPT in overall accuracy. Logistic Regression exhibited higher accuracy and significantly higher precision than Naive Bayes but lagged behind in recall, making it less suitable for this task where capturing all relevant environmental and energy-related content is essential. Given the importance of recall in this context—ensuring that no significant environmental or energy topics are missed— I used Naive Bayes model to classify my unlabeled data

Category	Labeled Data	Unlabeled Data
Proportion of 0	0.68	0.59
Proportion of 1	0.32	0.41
Actual Count of 0	3112	184
Actual Count of 1	2115	86

Table 2: Proportion and Actual Counts of Categories in Labeled vs. Unlabeled Data

Looking at the distribution of my labeled and unlabeled data in table 2, I observed that approximately 40% of the chunks in my dataset mention environmental or energy regulation. This aligns reasonably well with my training data, where 32% of the chunks were coded as mentioning regulation. The similarity between these percentages suggests consistency in the data and coding process. This distribution is also logical given the context of the dataset, which comprises chunks extracted from the Risk Factors sections of 10-K reports filed by companies in the mining and utilities sectors. These industries are inherently sensitive to environmental and energy regulations, as compliance with such policies often represents both a critical operational challenge and a source of significant financial risk.

If the goal is to measure exposure to environmental regulation, these companies, which are likely to have high exposure, do indeed discuss environmental and energy regulation frequently in their filings. This provides a strong signal that my methodology is capturing relevant content. Companies in these sectors, by necessity, must highlight regulatory risks as a key concern in their disclosures, validating the utility of focusing on these chunks to gauge regulatory exposure.

Rank	Top Tokens	Rank	Top Tokens
1	epa	11	ghg
2	emiss	12	dioxid
3	fractur	13	land
4	greenhous	14	biden
5	hydraul	15	mandatori
6	air	16	presid
7	climaterel	17	climat
8	pari	18	fossil
9	gase	19	injunct
10	methan	20	ban

Table 3: Top 20 Tokens Associated with `regulation_present`

Based on the Naive Bayes model, the top 20 tokens associated with `regulation_present` are depicted in Table 3. These tokens are consistent with expectations, as they include terms such as emissions, greenhouse, and relevant regulatory entities like President Biden, EPA, and climate-related keywords like "cap-and-trade." The presence of these terms aligns with the model's focus on environmental and energy regulation, and the inclusion of key political figures reflects the importance of policy in driving regulatory discussions in these areas.

4.2 Validation

To validate my measure, I reconstructed the dictionary-based measure proposed by Baz et al. (2023), which uses a dictionary approach to assess climate regulatory exposure (CRE). Baz et al. built on two existing dictionaries: (a) the climate change vocabulary (CCV) from Chou and Kimbrough (2019), which captures climate-related terms, and (b) the government policy vocabulary (GPV) from Koijen et al. (2016), which focuses on government-related regulatory terms. Baz et al. (2023) created a continuous variable that identifies instances where words from both dictionaries appear in close proximity within the text. These occurrences are then summed across the entire document to compute a measure for CRE.

However, for my hand-coded analysis, I simplified their approach by focusing on smaller text units as well as creating a binary measure. Specifically, instead of examining the full document, I restricted my analysis to 10-sentence chunks of text. Instead of calculating a continuous measure, I simplified the CRE measure by checking whether at least one term from each dictionary appeared within a given chunk. So, while Baz et al. used a continuous measure, I employed a binary approach, which is more suitable for the validation of my measure.

I used the regulatory terms identified by Koijen et al. (2016) for the GPV, as well as the climate-related terms from Chou and Kimbrough (2019). The dictionaries are as follows:

- `ccv_terms` = { 'climate change', 'climate', 'global warming', 'clean air act', 'ghg', 'rps', 'weather', 'rain', 'emission', 'carbon', 'carbon dioxide', 'carbon tax', 'cap-and-trade', 'sustainable', 'renewable', 'green' }

- `gpv_terms` = { 'congress', 'government regulation', 'political risk', 'congressional', 'government approval', 'politics', 'debt ceiling', 'government debt', 'price constraint', 'federal', 'government deficit', 'price control', 'federal funds', 'government intervention', 'price restriction', 'fiscal imbalance', 'law', 'regulation', 'government-approved', 'legislation', 'regulatory compliance', 'government-sponsored', 'legislative', 'regulatory delay', 'governmental', 'legislatory', 'reimbursement', 'government program', 'political', 'subsidy', 'government program', 'political', 'subsidies' }

Upon comparing the two dictionaries, I observed that the combination of these dictionaries does not adequately capture the top 20 most frequent words identified by the Naive Bayes model in table 3. This discrepancy arises because the government policy vocabulary used here is not well-suited to capture terms directly associated with environmental regulation. For example, terms like "EPA," "Paris Agreement," "Biden," and other related actors would not be included in this government vocabulary. These terms are critical for identifying climate regulatory exposure in the context of environmental risk, but they are missing from the dictionary-based approach.

Class	**Precision**	**Recall**	**F1-Score**
0	0.89	0.82	0.86
1	0.59	0.71	0.65
Accuracy	0.80		
Macro avg	0.74	0.77	0.75
Weighted avg	0.81	0.80	0.80

Table 4: Classification Report for Dictionary Approach

To further validate my measure, I assessed the performance of the dictionary-based approach against my hand-coded measure using several performance indicators for the `regulation_present` variable. In this analysis, I treated my hand-coded values as the "true values" and the dictionary-derived values as the "predicted values." Table 4 presents the classification report for the testing data, which includes precision, recall, and F1-score for both classes. The dictionary-based model achieved an overall accuracy of 80%, but it showed a lower recall (71%) and much lower precision (59%).

I believe this comparison provides some validation of my own measure, as the results are not drastically different from the established measure by Baz et al. (2023). However, it also highlights the limitations of using a dictionary-based approach to capture climate regulatory exposure, particularly when key environmental terms are absent from the dictionary. Despite these limitations, the accuracy of the dictionary-based model is still notable, even with the constrained vocabulary used. This suggests that with an expanded dictionary that includes more climate-related terms, the dictionary-based approach could potentially improve in effectiveness. In fact, such an expansion could reduce the need for more advanced text-as-data methods, making dictionary-based approaches a more viable option for measuring regulatory exposure in the future.

5 Findings and Limitations

How does your analysis shed light on your original research question? What are its limitations and what would be your future work? [2-3 paragraphs.]

My research question is: How can we measure the exposure of firms to climate and energy risk stemming from regulation? Based on my analysis, 10-K filings have proven to be a rich source of environmental regulation discussions for firms in sectors like mining, utilities, and oil and gas. These sectors are more likely to engage with environmental regulations, making the 10-K documents a good place to look for relevant regulatory discourse.

My analysis also sheds significant light on the challenge of measuring a firm's exposure to climate and energy regulation, and the difficulties inherent in disentangling these two types of regulation. The creation of a joint category for both environmental and energy regulation provided a useful starting point, but I found that this combined category often masks the nuanced differences between the two. Many instances of energy regulation are deeply rooted in climate concerns, as policies governing energy production and consumption are frequently motivated by climate change mitigation goals. This overlap makes it challenging to distinguish between the two types of regulation, especially when the regulations address similar issues from different angles. Through hand-coding and analyzing the text of 10-K filings, I observed that energy regulation is often presented alongside environmental concerns, further complicating the task of categorizing them separately.

A dictionary-based approach, such as the one employed by Baz et al. (2023), also fares well in answering this question. The approach uses pre-defined dictionaries to identify and quantify the presence of climate-related and government policy-related terms in text. While this method is effective, I believe there is potential to improve upon it by refining the dictionaries used. By expanding and customizing the dictionaries to better capture the nuances of climate and energy regulation in the context of 10-K filings, we can improve the measure's accuracy and relevance.

In terms of future work, a more refined classification system would be needed to better capture the complexities of climate and energy regulation. One possibility could involve developing separate dictionaries for environmental and energy regulation, combined with advanced machine learning techniques like topic modeling, which could help uncover latent themes and trends in the data. Additionally, extending the analysis to consider both the timing and intensity of regulations could provide further insights into how these regulatory risks evolve over time and impact firms in different ways. Overall, while this study provides a useful framework for understanding regulatory exposure, it also highlights the need for further refinement and exploration of these issues.

Another key aspect of future analysis involves the challenge of aggregating regulatory exposure to the document level. In my current work, I focused on 10-sentence chunks, but for a more comprehensive and holistic understanding, it is important to extend the analysis to the entire 10-K document. Aggregating results from smaller text units (e.g., the 10-sentence chunks) to the document level presents challenges, particularly as regulatory terms might appear sparsely across the document. One potential solution is to perform the analysis at a finer granularity, such as at the sentence level. This would allow for better identification of regulatory terms even when they are scattered throughout the document, providing a more accurate measure of regulatory exposure.

References

Baz, S., Cathcart, L., Michaelides, A., and Zhang, Y. (2023). Firm-level climate regulatory exposure. *Available at SSRN 3873886*.

Chou, C. and Kimbrough, S. O. (2019). Talking about climate change: What are enterprises saying in their sec filings? *Available at SSRN 3509765*.

Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Koijen, R. S., Philipson, T. J., and Uhlig, H. (2016). Financial health economics. *Econometrica*, 84(1):195–242.