Environmental Risk Discourse Evolution in SEC 10-K Filings

Deep Learning for the Social Sciences

Robert Seidel, Ludwig Kunz, Timothy Gunson, Maximilian Wallhöfer

Supervisor: Giordano De Marzo

August 31, 2025

Contents

1 Introduction

Environmental risks have become an increasingly central concern in corporate governance, financial markets, and public policy. Climate change, regulatory pressure, and social expectations require companies to recognize and communicate environmental threats to their operations. One of the most important sources for studying corporate risk perception is the mandatory disclosure of risk factors in annual SEC 10-K filings. These documents provide a standardized and comparable view of how firms articulate threats to their business, including environmental risks, over time.

The aim of this project is to analyze the evolution of environmental risk discourse in the SEC 10-K Risk Factors sections between 2005 and 2024. By examining how companies describe environmental challenges, we seek to uncover patterns of growing awareness, shifts in terminology, and strategic framing of risks. Previous research has shown that textual analysis of financial disclosures can reveal important insights into corporate priorities and communication strategies. However, the dynamics of environmental risk discourse in financial filings remain underexplored, especially in relation to major climate events and regulatory milestones.

To address this gap, the project combines large-scale data collection with modern natural language processing. We extract the risk factor sections from SEC filings, preprocess and align them across firms and years, and apply both dictionary-based methods and deep learning models for classifying parts regarding environmental risk. We analyze the sentiment and strategic positioning of the so-classified parts with respect to environmental risk. This enables us to track temporal trends, differences between industries, and the relationship between disclosure strategies or sentiment and external events. Ultimately, the findings contribute to a better understanding of how corporations integrate environmental concerns into financial reporting and how these narratives evolve in response to external pressures.

2 Data Collection and Cleaning

The data set consists of SEC 10-K filings for firms in the S&P 500 index over the period 2005–2024. We focus on Section 1A (*Risk Factors*), which contains standardized corporate disclosures on business risks.

To ensure temporal consistency, we first identified 215 permanent constituents of the S&P 500 between 2005 and 2024. Historical index membership data was obtained from Farrell Aultman's `sp500` GitHub repository, which records changes in index composition. By intersecting membership lists across years and matching company identifiers with SEC Central Index Keys (CIK), we obtained a list of permanent constituents.

For each identified company, the filings were accessed via the SEC EDGAR API. Two sets of requests were required: (i) metadata retrieval through `data.sec.gov` to collect accession numbers and filing dates, and (ii) document retrieval from `www.sec.gov` to download the actual 10-K reports. The extraction process targeted all available 10-K filings from 2005 onward.

Due to the very heterogeneous and sometimes convoluted structures of the extracted filings, the Section 1A text was extracted using a two-step strategy. First, anchor links pointing to "Item 1A. Risk Factors" and "Item 1B. Unresolved Staff Comments" were identified in the HTML structure. If this method failed – for example, due to missing or malfunctioning anchor links – a fallback search over bold-styled headings was applied to locate the section boundaries. The extracted texts were cleaned by removing table of contents artifacts, isolated page numbers, and duplicate fragments. For each firm, the

extracted sections from all available years were stored in a pickle DataFrame.

To ensure reproducibility, extraction errors were logged. This allowed tracking of cases where Section 1A could not be located or where the parsing returned empty text. In total, 3,896 out of 4,246 filings were successfully collected, yielding a comprehensive longitudinal dataset of corporate risk disclosures.

3 Preliminary Data Analysis and Visualization

3.1 Preprocessing

After retrieving the Section 1A files, text segmentation was applied to prepare the dataset. The filings underwent comprehensive preprocessing using NLTK and custom regex patterns. Texts were Unicode (NFKD) normalized, cleaned of various PDF extraction artifacts, some of which were document-specific, and segmented into sentences using NLTK's sentence tokenizer. We extracted and standardized all Risk Factors sections and applied a basic dictionary-based filtering approach to identify environmental keywords. To ensure temporal comparability, filings were aligned on the basis of fiscal years. Climate-related content was identified using a dictionary of 65 climate-specific terms (e.g., "emissions", "carbon dioxide"), derived from prior research Kim, C. Wang, and Wu (2023) and additional words of interest. To identify the keywords a two-stage matching system was implemented for optimal performance:

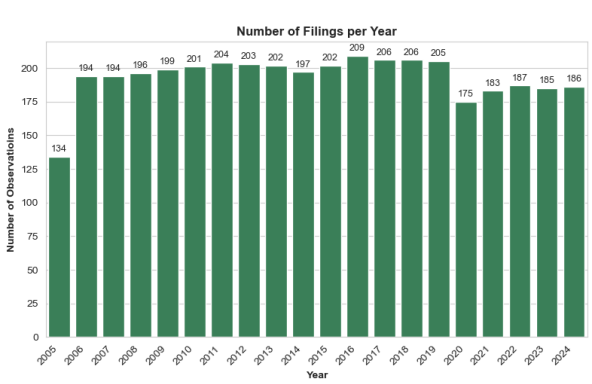
Stage 1: Fast regex-based filtering for initial candidate identification.

Stage 2: Linguistically precise matching through a complete NLP pipeline combining POS-tagging (Parts of Speech), lemmatization, and n-gram analysis.

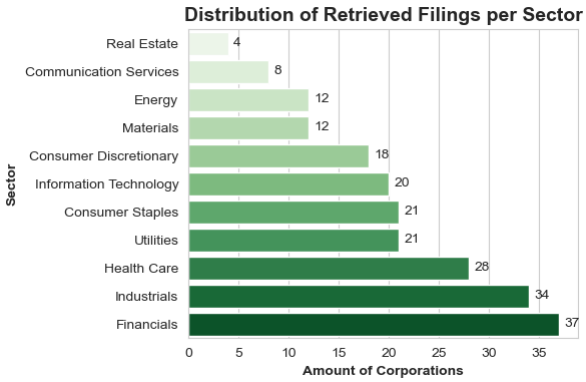
This approach normalizes different morphological variants (e.g., "emissions" $\rightarrow$ "emission") and handles multi-word expressions such as "Kyoto protocol" through sliding window n-gram matching up to the maximum keyword length. After the keyword processing, climate related sentences, words, and word counts were calculated and stored in a CSV DataFrame. Additionally sector data of corporations were retrieved from Wikipedia (2025) and matched for each corporation. This resulted in 10 different sector categories. Duplicates were removed, resulting in 3846 observations. Notably, the weak dictionary approach has limits due to its context insensitivity and potential failures of regexes due to the heterogenous structure of the filing reports. 47 (72.3%) different climate-related words of the climate dictionary were retrieved successfully within the filings.

3.2 Exploratory Data Analysis

A closer inspection of the retrieved data shows that filings for all years are almost equally distributed ensuring requirements for later trend analysis. In Figure 1a the annual number of filings in the dataset is visualized, revealing that in 2005 the number of filings deviates noticeably.



(a) Amount of retrieved filings per sector



(b) Sector Information

Figure 1: Filing and Sector Information

Furthermore, in Figure 1b it is obvious, that financial corporations together with the industrial sector already account for approximately 33% of the dataset, whereas the two least frequently occurring sectors, Real Estate and Communication Services represent only about 5.5%, resulting in a marked imbalance. Therefore, sector analyses should be carefully interpreted. To showcase outlier detection, sentence count analysis was performed with 1.5 * IQR (Interquartile Range) criteria. In total 130 (3.4%) filing outliers by length were found, all of them exceeding the upper bound criteria (581 sentences per report). Some of the outliers represent unusually comprehensive filings (very long), possibly reflecting firm size or litigation-heavy industries. The lower bound did not detect any outliers, due to the heavily right-tailed distribution of sentence counts. 269 (7%) filings contained fewer than 50 sentences, which might indicate extraction errors.

Aggregates by sector in Table 1 indicate that the Energy sector exhibits the highest mean climate-word density (0.22%), followed by Utilities and Materials. In contrast, the mean sentence share is substantially larger (e.g., 7.34% for Energy), which implies that a non-negligible share of sentences in those filings is climate-related even though the fraction of climate words is comparatively small. This difference arises from the distinct interpretations of the two normalization metrics: word density captures the intensity or verbosity of climate-related language within text, whereas sentence normalization measures the prevalence of climate related statements across the document.

Sector	mean density (%)	median density (%)	mean sentences (%)	median sentences (%)	avg words / filing	avg sents / filing	filings
Energy	0.22	0.19	7.34	6.09	7.94	11.92	203
Utilities	0.13	0.12	5.61	5.38	10.42	22.40	368
Materials	0.10	0.07	3.80	3.70	4.83	8.58	218
Real Estate	0.09	0.09	3.29	3.23	6.55	8.41	80
Industrials	0.08	0.05	2.52	1.92	3.69	5.42	619

Table 1: Sector-level descriptive aggregates for selected industries: mean and median climate word densities, sentence densities, average words and sentences per filing, and number of filings.

Comparing mean and median densities also reveals right-skewed distributions in several sectors: Particularly in Energy and Industrials a few very large filings drive up the mean. In the Utilities sector the right skewness is substantially weaker, though it remains detectable.

Sector	Filing count	Total Sentences	Total Climate Words	Total Climate Sentences
Energy	203	32,251	1,611	2,420
Financials	689	274,821	2,767	6,225
Utilities	368	136,998	3,834	8,243
Health Care	490	120,417	1,428	2,201

Table 2: Sector-level statistics for selected industries (updated with current data).

Table 2 reveals that Financials have a high absolute climate-related words or sentence count but considering normalization the proportion drops significantly. This is due to the fact, that their total amount of report sentences outnumber other sectors by far. Additionally, the Energy and Utilities sector highlight a high median sentence proportion of their filings in comparison to other sectors. Plot 2b showcases that especially these two sectors contribute more climate-related content in their risk section.

This might be due to the vulnerability of their core business operations. Presumably, Energy and Utilities are highly affected by regulatory policies as well as the increased number of climate related events, risking loss of revenue. On the other hand one could argue that Communication Services and Health Care sectors are substantially less affected by climate regulations/events and thus do not report the same intensity in their risk sections. To further analyze if sector differences are meaningful, we conducted pairwise Mann–Whitney U tests on the 5% significance level with FDR (False Discovery Rate) correction, after Kruskal-Wallis was statistically significant for all sectors (H = 954.805, p = 1.01e-198). The test is robust to skewed data and thus suitable for comparing the distributions of normalized climate-related sentence frequencies across sectors. Although, it is important to note that Mann Whitney U has certain limitations. Specifically, its reliability across subgroups has been questioned, and some items may not fully capture the construct of interest. These shortcomings should be considered when interpreting the results. Further analysis for effect sizes would result in more stable interpretations but could not be performed further due to text length limitations.

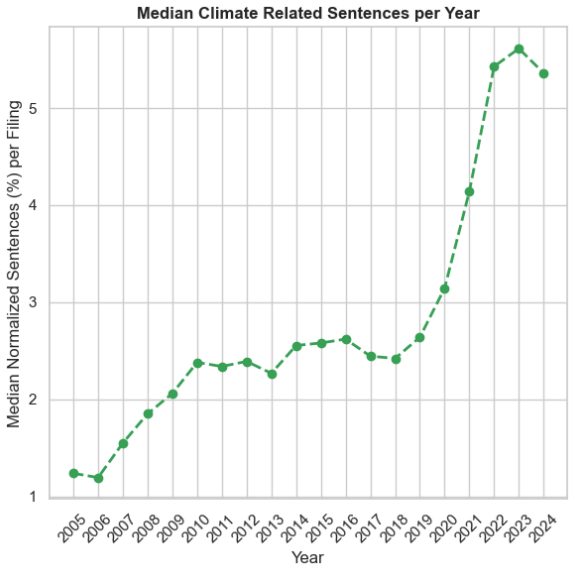
Sector comparison	U-statistic	p-value	p-adjusted	Significant
Energy vs. Health Care	90401	1.40e-64	1.92e-63	Yes
Financials vs. Utilities	32880	8.44e-88	2.32e-86	Yes
Health Care vs. Utilities	13441	2.99e-101	1.65e-99	Yes
Consumer Discretionary vs. Info Tech	61151	6.02e-01	6.13e-01	No
Materials vs. Real Estate	9822	9.45e-02	1.11e-01	No

Table 3: Selected pairwise Mann–Whitney-U results (FDR-corrected p-values). Three strongly significant comparisons and two non-significant comparisons to illustrate key EDA findings.

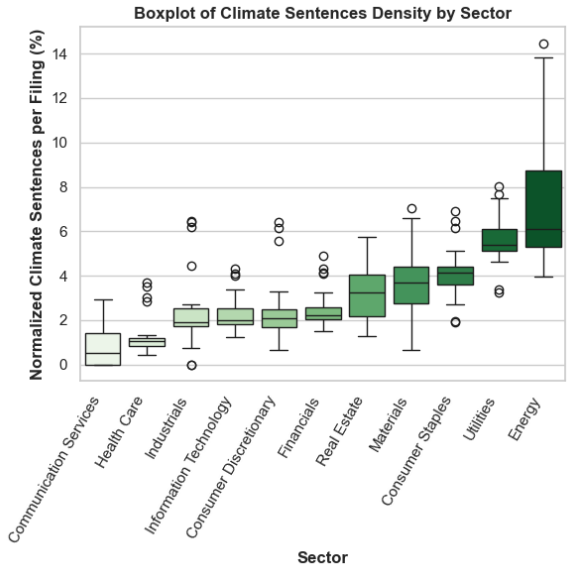
The pairwise Mann–Whitney U tests with FDR correction confirm that several sectors differ significantly in the normalized frequency of climate-related sentences. Notably, Utilities differs strongly from both Health Care and Financials (extremely small adjusted p-values), supporting the earlier observation that Utilities contains a comparatively large share of climate-related sentences. The comparison Energy vs Financials also shows a very strong difference, consistent with the higher proportion of climate-related sentences observed in Energy. By contrast, Materials vs Real Estate and Financials vs Industrials are not statistically different after correction, which suggests that these sector pairs exhibit similar patterns of normalized climate reporting.

Although risk related reportings seem to differ within the sector structures, Figure 2a reveals a global trend towards climate

risk disclosure. Visually, a first distinct change in reports began in the late 2000s due to the fact that climate related risks were incorporated as a mandatory part of the risk section report. Secondly, after 2019 a substantial rise in reportings occurred, signaling increased public awareness to climate-related topics.



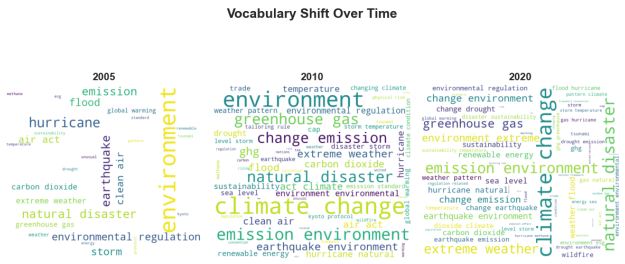
(a) Lineplot with Median Climate Related Sentences per Year



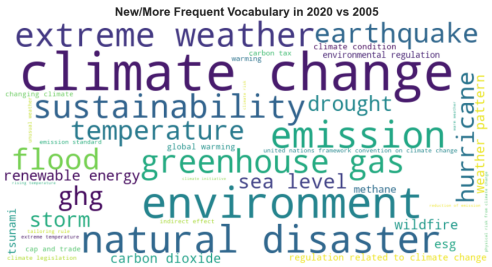
(b) Normalized Sentences per Sector

Figure 2: Annual climate related trends and distribution of climate related sentences.

To gain insights on how climate disclosure is evolving over time, we introduced word clouds to visualize the trend of climate-related disclosures. This made it possible to illustrate both gradual changes in terminology (Figure 3a) as well as the prominence of specific keywords in 2020 relative to 2005 (Figure 3b). Not only has the number of distinct climate-related words increased, but the overall tone of risk section filings has also shifted toward more climate-oriented language. Especially highlighting natural disaster phenomena, which are widely regarded as globally more frequently occurring and potentially be of great jeopardy for supply chains, energy related corporations or industrial sites.



(a) Vocabulary Shift over Time



(b) Change in Tone: New/More Frequent Climate Related Words (2020 vs. 2005)

Figure 3: Word Cloud Analysis

It is important to notice that climate events could overestimate individual words which can appear with more frequency in those years and limit interpretability. It is yet unknown if risk assessments of corporations are part of acknowledging global warming trends and thus realizing that regulations and natural disasters cause a significant loss in revenue and how this is stated in their sentiment tone in the risk sections. Therefore further analysis with FinBERT are applied in the following

sections.

4 Environmental Risk Classification and Sentiment Analysis

4.1 Preprocessing

In order to classify environmental risk disclosures within Section 1A filings, we created a dataset at sentence level from the preprocessed 10-K corpus preserving document structure. We obtained a cleaned dataset of 381,186 candidate disclosure sentences. We restricted our analysis to individual sentences rather than full texts or surrounding context, both to reduce complexity and because the base model of our choice (FinBERT-ESG) is also trained on single sentences.

As a first step, we implemented two baseline labeling strategies. The first is a dictionary-based approach, whereby a sentence is labeled as environmentally relevant if it contains specific environmental risk-related words (see 3.1). This provided a binary indicator that served as a weak reference for model development. The second baseline uses FinBERT-ESG, a transformer model pre-trained for financial, environmental, social, and governance (ESG) classification (yiyanghkust/finbert-esg; Huang, H. Wang, and Yang 2023). Applying this model to the entire dataset produced a baseline label for environmental relevance, whereby FinBERT's original four-class prediction was reduced to environmental versus non-environmental categories.

To prepare training data for supervised learning, we constructed a manually annotated sample. Using the FinBERT-ESG baseline labels, we performed stratified sampling across filings while capping the number of allowed sentences per document. This ensured that the sample was balanced between positive and negative classes and that no single filing dominated the dataset. In total, 1,184 sentences (514 positive; 670 negative) were annotated. Each entry was manually reviewed and labeled as environmentally relevant or not, resulting in a high-quality dataset for fine-tuning.

4.2 Methods

For supervised classification, we fine-tuned the pre-trained FinBERT-ESG model on the manually labeled dataset. The model was initialized with a binary classification head to distinguish environmentally relevant from non-relevant sentences. The dataset was split into training, validation, and test sets in a 70/15/15 ratio while preserving class balance.

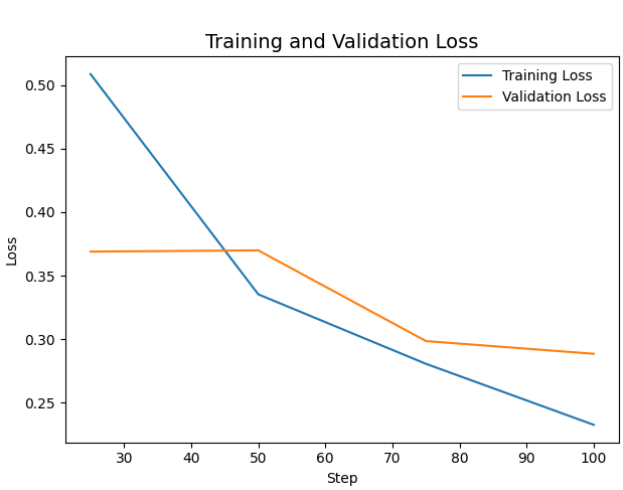
Text inputs were tokenized with the FinBERT-ESG tokenizer, applying padding and truncation to a maximum sequence length of 216 tokens. To enable efficient adaptation, we employed Low-Rank Adaptation (LoRA) modules on the main BERT layers (`query`, `key`, `value`, `dense`). We used a dropout rate of 0.1 and weight decay to regularize updates. Hyperparameter tuning was conducted over three key dimensions: the LoRA rank $r \in \{4, 8, 16\}$, the training batch size $bs \in \{8, 16\}$, and the learning rate $lr \in \{5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\}$. All models were trained using the standard cross-entropy loss for binary classification.

Model selection was guided by validation performance, using recall as the primary metric, since capturing as many environmentally relevant sentences as possible is critical for the downstream analysis. Training was run for up to 5 epochs with early stopping with a patience of 3 evaluations to prevent overfitting. After each run the checkpoint with the highest recall was retained for comparison.

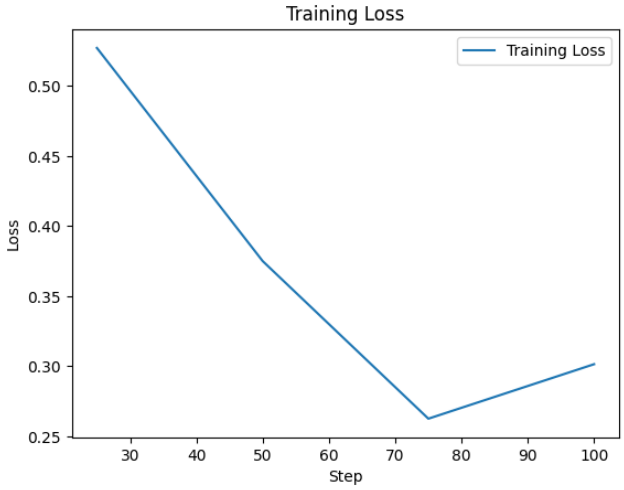
Out of the hyperparameter candidates, the final model was selected based on validation performance, prioritizing recall as the main metric and using F1 score as a secondary criterion. Among models with comparable recall, the configuration with lower training complexity was preferred to improve generalization. Threshold tuning was not found to improve the model's performance and was therefore excluded.

The optimal configuration was obtained with a LoRA rank of $r = 8$, a batch size of $bs = 8$, and a learning rate of $lr = 5 \times 10^{-4}$. Training was run for 100 steps (1.5 epochs), corresponding to the early stopping point with the highest recall. We observed that models trained on fractional epochs generalized better on the validation set. In this setting, 1.207% of the model parameters were updated during training, ensuring efficiency while retaining model capacity. Since we wanted to make the most of our small, manually annotated dataset, we trained our best model using both the training and validation sets.

During hyperparameter tuning, the validation loss plateaued while the training loss continued to decrease, indicating stable learning without substantial overfitting. In the final training run on the combined training and validation set, the loss declined steadily up to around 75 steps before showing a slight increase near step 100. Since our primary objective is generalization, this fluctuation does not undermine the overall performance assessment of the model.



(a) Training and validation loss curves of the best candidate model during hyperparameter tuning.



(b) Training loss curve of the selected best model during final training on the combined training and validation set.

Figure 4: Loss curves for the best FinBERT-ESG fine-tuning setup.

4.3 Performance

The dictionary-based baseline provides only limited coverage of environmentally relevant sentences. Both accuracy and consistency improve notably when using the FinBERT-ESG baseline, which benefits from domain-specific pretraining. The fine-tuned FinBERT-ESG model achieves the strongest balance of recall and F1 score, making it the most reliable choice for identifying environmental risk disclosures in our dataset.

The confusion matrix shows a well-balanced performance across both classes, with the model accurately identifying relevant and non-relevant sentences at similar rates.

Model	Split	Accuracy	F1	Recall	Precision
Dictionary Baseline	Validation	0.68	0.63	0.68	0.76
Dictionary Baseline	Test	0.71	0.68	0.71	0.80
FinBERT-ESG Baseline	Validation	0.85	0.85	0.85	0.85
FinBERT-ESG Baseline	Test	0.84	0.84	0.84	0.87
Fine-Tuned FinBERT-ESG	Validation (during tuning)	0.87	0.86	0.91	0.81
Fine-Tuned FinBERT-ESG	Test	0.88	0.88	0.88	0.88

Table 4: Performance overview of baseline and fine-tuned models for environmental risk classification.

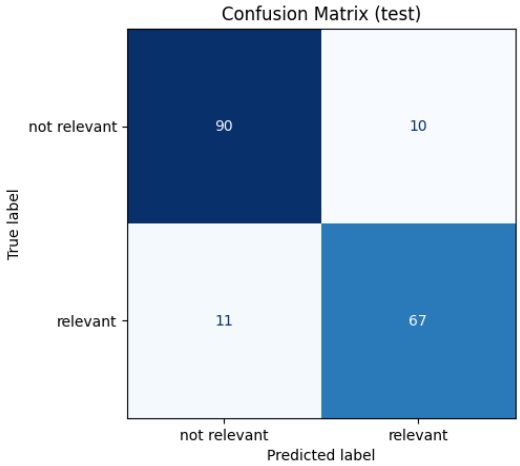


Figure 5: Confusion matrix of the best fine-tuned FinBERT-ESG model on the test set.

An inspection of misclassified examples helps to illustrate the model's limitations. For instance, the sentence *"Operation of fueling stations, landfill gas collection and control systems and waste to energy plants involves additional risks of fire and explosion."* was incorrectly labeled as environmentally relevant. While it mentions energy infrastructure, the disclosure relates primarily to operational and safety risks rather than environmental risks, highlighting how the model can be misled by overlapping terminology.

Looking at all misclassified test set examples (10 false positives, 11 false negatives) reveals recurring error patterns. False positives mainly occur when the model confuses energy infrastructure or regulatory mentions with environmental risks, such as disclosures about nuclear fuel, fossil fuels, or generic EPA regulations (e.g., "air quality under the Clean Air Act"), where the emphasis lies on operational, safety, or legal compliance rather than environmental impacts. Similarly, references to political or budgetary constraints around renewable energy policies are often misclassified as environmental risks, indicating an over-reliance on surface-level keywords like "energy," "clean," or "renewable."

False negatives, by contrast, arise when genuine environmental exposures are expressed indirectly or in technical or strategic language. Examples include forward-looking regulatory initiatives (e.g., California's executive order on zero-emission vehicles), energy transition strategies, or sector-specific risks such as fiber sourcing and nuclear decommissioning. These cases suggest that the model underestimates implicit or contextual signals of environmental relevance, particularly when embedded in broader financial or strategic framing.

Overall, the fine-tuned model tends to (i) overpredict environmental relevance when strong keywords appear without substantive context, and (ii) underpredict when environmental content is implicit or technically framed. This highlights a reliance on lexical cues at the expense of deeper contextual understanding and points to the need for further fine-tuning or more context-aware architectures.

4.4 Sentiment Analysis

After classifying environmentally relevant risk sentences with the fine-tuned FinBERT-ESG model, we constructed a dataset consisting only of these disclosures (2,587). For sentiment analysis we did not have the resources to manually annotate data for fine-tuning, and therefore employed the pre-trained FinBERT-tone model, a transformer-based architecture trained for financial sentiment classification with three output categories: negative, neutral, and positive (yiyanghkust/finbert-tone).

To capture the overall sentiment at the filing level, we grouped all environmentally relevant sentences by company identifier and fiscal year, preserved their original order, and concatenated them into a single text per filing. Each aggregated text was then classified with FinBERT-tone, using a sufficiently large sequence length to ensure that no content was truncated. This approach allowed us to assign a single sentiment label to every filing, reflecting how the firm communicated environmental risks in that year.

Out of all filings, 2,026 were classified as negative, 539 as neutral, and only 12 as positive. This distribution confirms that Section 1A predominantly emphasizes risks and potential adverse developments, while neutral statements are less common and positive framings of environmental risks remain exceptional.

5 Risk Communication Classification

5.1 Methods

To classify companies' strategic self-positioning in their environmental risk communication, we used the dataset consisting of all environmentally relevant text per filing, including the sentiment labels from the FinBERT-tone model. The baseline model is Qwen3-0.6B , a small pre-trained transformer LLM and part of the Qwen3 family (Qwen/Qwen3-0.6B; Qwen 2025). It contains roughly 600 million parameters, spread across 28 layers. We initialized the model with a binary classification head to classify whether a filing was merely acknowledging the environmental risk at hand, or whether it emphasized concrete actions to manage it.

For this classification, we used the following prompt:

> *You are an expert classifier for corporate risk communications.*
>
> *Classify the following environmental risk sentences for a 10-K filing.*
>
> *Labels: 0=acknowleding (states the risk descriptively without stressing or minimizing); 1=actively managing (emphasizes concrete actions, commitments, or leadership in addressing the risk).*
>
> *Output only one number. Do not output anything else.*
>
> *Sentiment:* `sentiment` *(context only)*
>
> *Sentences:* `filing text`
>
> *Final Answer: [0|1]*

This prompt was used with the base model to create baseline labels[1]. A randomly sampled dataset with 1000 observations was then created for manual annotations. Stratification was not employed here, due to the difficulty of the classification task making the validity of the baseline labels uncertain. 500 filings were evaluated, 376 of which were labeled as *acknowledging*

[1]Due to GPU memory constraints, the full prompt including the filing text was designed to be no longer than 2,048 tokens. If a text was too long, it was truncated at the end, in order to still include the "Final Answer" section.

and 124 as *actively managing*.The resulting dataset was split into a training (70%), a validation and a test dataset (15%, respectively) while preserving the class ratio.

LoRA was employed for fine-tuning, to ensure memory-efficient optimization. Our configuration contains a LoRA adaption rank of 4, a scaling factor of 32 and a dropout rate of 0.1 to regularize updates. Adaptation was applied to a broad set of target modules, including `query`, `key`, `value` and `output projections`, as well as `feedforward components`. Training was conducted for two epochs, with a per-device batch size of 4 and gradient accumulation over 4 steps, effectively simulating a larger batch size under GPU memory constraints. The learning rate was set to 2×10^{-5} with a linear scheduler and a warmup ratio of 0.06. To mitigate overfitting, L2 regularization was applied via a weight decay of 0.01, and label smoothing was used to avoid overconfident predictions. The loss function was a weighted cross-entropy loss, where class weights were computed to address label imbalance. In addition, weighted random sampling was applied during training to ensure that minority classes were adequately represented in each batch. An early stopping procedure was implemented with a patience of 3 evaluations (conducted every 5 steps over 44 training steps) to prevent overfitting. Threshold tuning was not found to improve the model's performance and was therefore excluded.

Model selection was based on the F1-score, which served as the primary evaluation metric due to its ability to balance precision and recall in imbalanced classification tasks. Training logs included both training and validation loss curves, as well as F1 trajectories, to monitor convergence.

In Figure 6, the training loss shows a steady decrease over the first steps but fluctuates thereafter, suggesting some instability during optimization. In contrast, the validation loss stabilizes around 0.6 relatively early, indicating that the model generalizes reasonably well without severe overfitting. The F1-score improves markedly up to step 15 and then plateaus with minor oscillations, showing that the model reaches a strong performance level early in training and maintains it consistently.

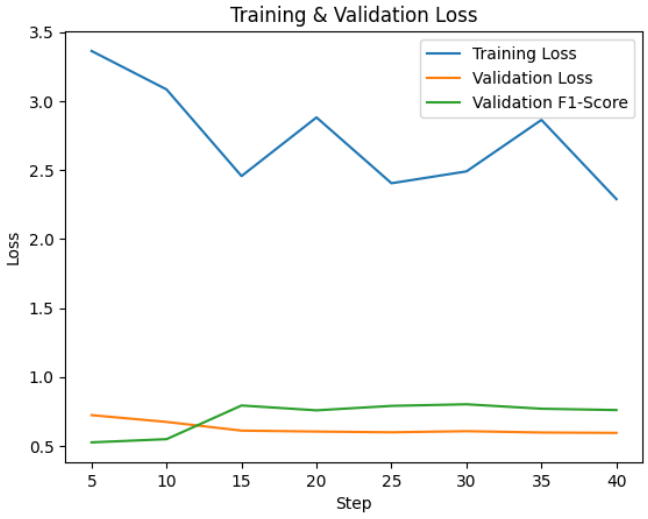


Figure 6: Training, validation loss and F1-score curves of the fine-tuned Qwen 3 model for strategic positioning classification.

5.2 Performance

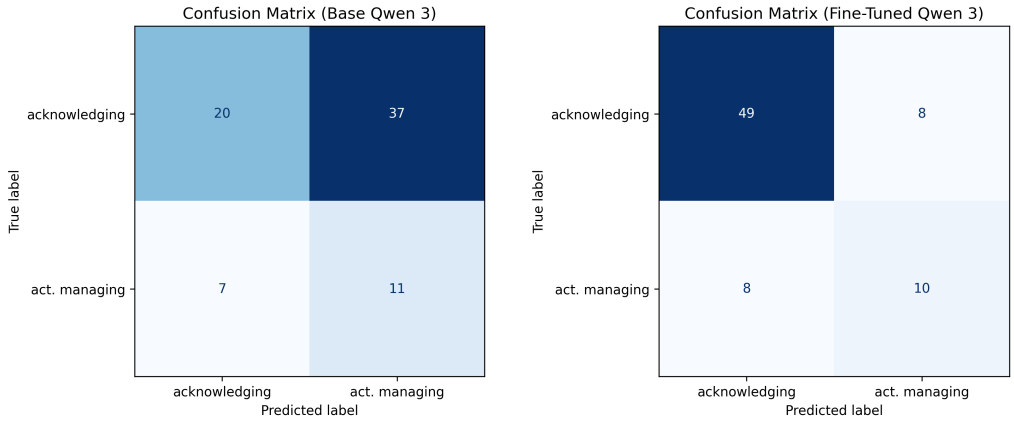
Table 5 shows the performance metrics for the base Qwen 3 model and its fine-tuned version. The base model performs quite poorly, falsely predicting over half of the strategic positioning labels. All recorded performance metrics improve substantially after fine-tuning, with almost 80% of all positioning labels in the test set being predicted correctly. F1, recall

and precision scores are slightly lower than the model's accuracy, but still indicate reliable predictions.

Model	Split	Accuracy	F1	Recall	Precision
Base Qwen 3	Validation	0.31	0.30	0.36	0.38
Base Qwen 3	Test	0.41	0.40	0.48	0.48
Fine-Tuned Qwen 3	Validation	0.80	0.80	0.80	0.80
Fine-Tuned Qwen 3	Test	0.79	0.71	0.71	0.71

Table 5: Performance overview of baseline and fine-tuned Qwen 3 models for strategic positioning classification.

The confusion matrices in Figure 7 corroborate these metrics. The base model significantly overpredicts cases of actively managing communications, while the correct and false predictions on this class are more balanced after fine-tuning. Despite this improvement, the fine-tuned Qwen 3 model still struggles to reliably predict whether a filing contains active commitments. This may be due to the extensive length of some of the filings sections, exceeding the maximum length that can be displayed in the prompt. Due to GPU memory constraints it was, however, impossible to increase the maximum prompt length. The extensive text lengths also prohibit more in-depth analyses of errors and specific examples. Predictions from the fine-tuned model are used for further analysis of strategic risk communication.



(a) Confusion matrix of the base Qwen 3 model

(b) Confusion matrix of the fine-tuned Qwen 3 model

Figure 7: Confusion matrices of the base and fine-tuned Qwen 3 models on the test set.

6 Results

We analyze the evolution of sentiment and risk communication across sectors from 2005 to 2024, fitting linear regressions and analyzing the distribution of sentiment scores and risk communication disclosure. Subsequently, a Seasonal Autoregressive Integrated Moving Average with exogenous regressor (SARIMAX) analysis is conducted to find out whether climate events or regulatory events have an impact on the sentiment or risk disclosure patterns.

6.1 Evolution of Sentiment and Risk Communication

To examine long-term dynamics in disclosure, an ordinary least squares (OLS) regression was applied to annual data from 2005 to 2024. The model revealed a significant trend toward increasingly negative sentiment over time (coeff. = $-0.0012, p < 0.05$). No significant temporal effects were detected for risk communication patterns.

The distribution of sentiment values (0 = negative, 1 = neutral, 2 = positive) is strongly left-skewed, with a mean of 0.22 and

standard deviation of 0.42. Similarly, the binary risk communication indicator (0 = acknowledging, 1 = actively managing) shows a mean of 0.22 and standard deviation of 0.41.

There is a strong difference in how different sectors communicate risk disclosure. Sectors like Utilities and Energy show by far the highest scores for actively managing communication (paralleling the highest ratios of climate sentences in their risk reports, as described in section 3.2), while Information Technology and Communication Services show the lowest scores (see Figure 8).

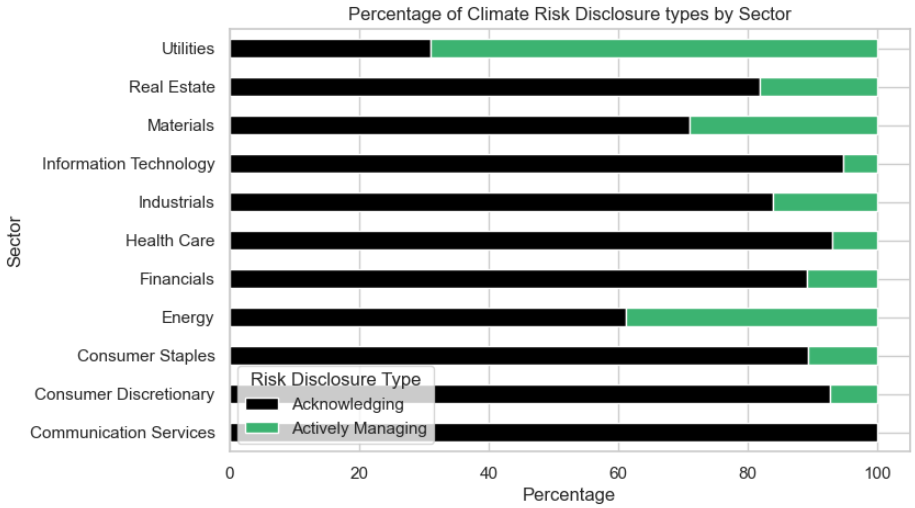


Figure 8: Risk disclosure by sector

6.2 Relationship Between Disclosure Patterns and Regulatory/Climate Events.

After compiling a list of climate events associated with extraordinary damages and casualties, a Seasonal Autoregressive Integrated Moving Average with exogenous regressors (SARIMAX) was implemented to assess the influence of such events on disclosure patterns. Lags of 1–6 months and 1 year were tested.

The first set of models examined the relationship between climate events and sentiment. Results indicated a significant positive effect of climate events on sentiment (coeff $= 0.23, p < 0.01$) at a lag of 1 month, while lags of 2 to 6 months and 1 year were not significant. The positive coefficient suggests that companies present themselves with more favorable sentiment following climate shocks, potentially as a reputational strategy to signal control over climate-related risks.

In a second analysis, major regulatory events were tested as exogenous drivers of sentiment. Eleven regulatory milestones, which were directly related to environmental topics and SEC climate risk disclosure requirements, were included.

Regulatory events had a statistically significant, positive effect on sentiment with lags of 1 month (coeff $= 0.15, p < 0.05$) and 2 months (coeff $= 0.15, p < 0.03$), as well as 5 (coeff $= 0.26, p < 0.05$) and 6 Months (coeff $= 0.23, p < 0.05$). No significant effects were detected at 3-month, 4-month, or 1-year lags. We expected delayed sentiment responses to reflect the time required for adoption and integration into disclosure practices. However, our findings do not support those expectations and suggest no obvious pattern.

Another SARIMAX investigation was conducted on risk communication strategies. Climate events exhibited significant effects at lags of 1, 2, and 3 months (coeff$(1) = 0.11$, coeff$(2) = 0.30$, coeff(3)$= 0.15$, each with $p < 0.05$). Longer lags (4–6 months, 1 year) were not significant. The short-term increase in active risk communication may suggest that firms

respond to climate events by signaling proactive management, either to reassure stakeholders or to protect their reputation.

Figure 9 displays the average monthly sentiment scores and risk communication scores, as well as trend lines, climate events and regulation dates.

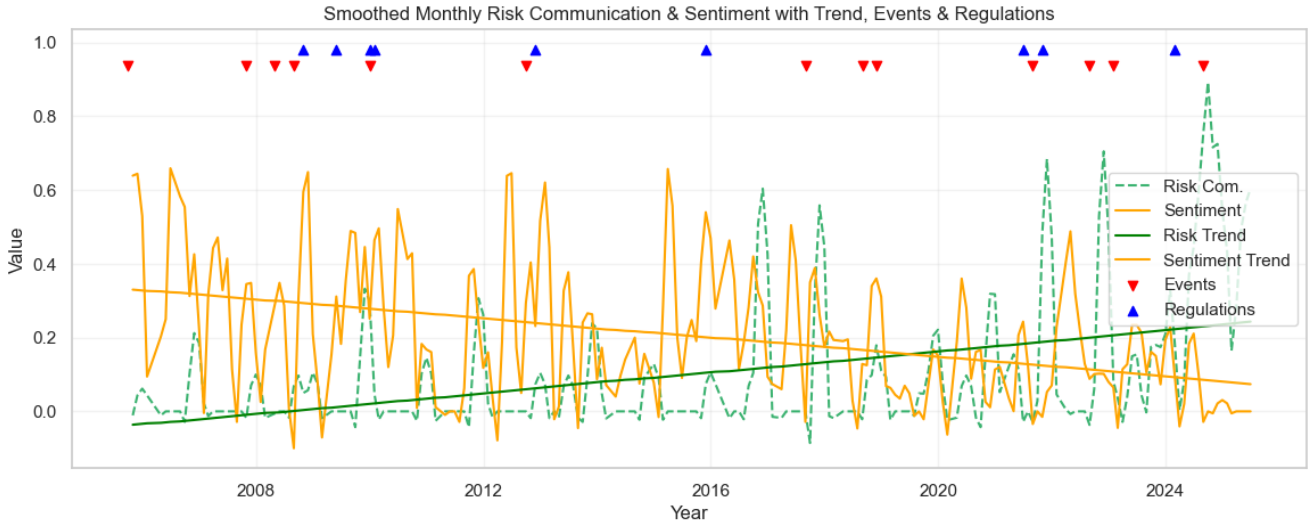


Figure 9: Monthly averages of risk communication and sentiment with event and regulation dates.

The findings provide several implications for environmental finance, corporate sustainability analysis, and policy evaluation. First, regulatory events are associated with shifts in disclosure sentiment but do not consistently translate into more active risk management communication, at least within SEC filings. Second, climate shocks appear to directly alter how firms discuss climate change and its risks, reflecting heightened societal relevance of catastrophic events. However, it must also be acknowledged that these findings merely rely on how the companies portray themselves in financial statements and not on their observable behavior. Greenwashing and strategic downplaying of climate risk remain possible and warrant further investigation.

References

Huang, Allen H, Hui Wang, and Yi Yang (2023). "FinBERT: A large language model for extracting information from financial text". In: *Contemporary Accounting Research* 40.2, pp. 806–841.

Kim, Jeong-Bon, Chen Wang, and Fei Wu (2023). "The real effects of risk disclosures: evidence from climate change reporting in 10-Ks". In: *Review of Accounting Studies* 28.6, pp. 2271–2318. DOI: 10.1007/s11142-022-09687-z.

Qwen, Team (2025). *Qwen3 Technical Report*. arXiv: 2505.09388 [cs.CL]. URL: https://arxiv.org/abs/2505.09388.

Wikipedia (2025). *List of SP 500 companies*. Accessed: 2025-08-31. URL: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.