Toward Transparent ESG Reporting: Analyzing Promise and Constraint Claims

Anonymous EMNLP submission

Abstract

In the era of increasing regulatory scrutiny and stakeholder expectations, understanding how companies frame their sustainability commitments and limitations is essential for assessing corporate accountability. This study explores promise and constraint claims made by 1,898 companies worldwide from 2020 to 2024. Promises are forward-looking claims that tend to lack specific, measurable actions or mechanisms for accountability, while constraints - are sentences that mention some impediments, restrictions, or obstacles at a company, society, or governmental level that may restrict the company from fulfilling its promises. Our study provides a rigorous and well-defined definition of promise and constraint claims based on the sustainability reports, the terms that have not been explored before, and offers a comprehensive dataset of 5,299 annotated sentences from 2,221 reports[1]. The research presents a lightweight alternative to resource-intensive models by employing ClimateBERT and fine-tuning it as a ClimateBERT-Promise-Constraint model on the collected data. The analysis identifies the distribution of constraint claims across four primary sectors: natural resources, manufacturing, retail, and information. This work contributes a comprehensive dataset and modeling framework, supporting future research on corporate accountability and transparency within environmental, social, and governance (ESG) disclosures, aligned with emerging regulations such as the EU Corporate Sustainability Reporting Directive (CSRD).

1 Introduction

Corporate accountability has emerged as a cornerstone of sustainable development in an era marked by increasing environmental challenges and societal expectations. Companies are now expected to generate profit while still aligning their operations with environmental, social, and governance (ESG) principles (Yudoko, 2024). Sustainability reporting, a growing practice among organizations worldwide, is critical for documenting and communicating these efforts.

Our study looks at sustainability reports provided by companies within 2020-2024. The sustainability reports are public documents organizations use to disclose their environmental, social, and governance (ESG) performance. These reports provide transparency about a company's impact on society and the environment, detailing efforts in areas such as reducing carbon emissions, improving labor practices, and enhancing corporate governance. By sharing this information, companies aim to build trust with stakeholders, including investors, customers, and employees, and demonstrate their commitment to sustainable development (Gunawan, 2023).

The practice of providing sustainability reports has gained significant recognition globally. For instance, as of 2020, 96% of the world's largest 250 companies (the G250) reported on their sustainability performance (Global Reporting Initiative and Sustainability Accounting Standards Board, 2021). This widespread adoption reflects the growing acclaim for the importance of transparency in corporate social responsibility initiatives.

By leveraging the information from the reports, we get a closer look at *promises* a company pledges to fulfill within a specific time or *constraints* - claims that identify impediments that obstruct a company's realization of a particular objective (see Table 1 with provided examples). The promises and constraints go hand-in-hand as the company intends to justify some of its failures or set lower stakeholder expectations.

The main resolution of this study is to identify

[1]The main GitHub repository of the project with the data and code (the link will be added after the review to comply with the anonymity policy). The fine-tuned model and dataset are also provided via the HuggingFace (the links will be added after the review to comply with the anonymity policy)

Sentence	Label
We aim to be a net zero company in the entire value chain.	Promise
We will promote the introduction of a solar power generation system to realize the shift to renewable energy.	Promise
However, as the world shifts to a lower carbon economy, various federal, state, and/or provincial legislative mechanisms could cause our operational costs to increase significantly.	Constraint
Various factors may result in substantially different outcomes.	Constraint

Table 1: The most representative examples of promise and constraint claims from the gathered dataset.

such claims and to foster research connected with promise and constraint identification in ESG reportings. We believe our contribution will be fruitful in the light of the European Union Corporate Sustainability Reporting Directive (CSRD), which came into force on January 1, 2025, and mandates that companies meeting specific criteria disclose risks related to climate change and their societal impacts (Carey, 2024).

The study contributes to the established paradigm of environmental claims detection by providing a curated, comprehensive, and extensive dataset with 5,299 annotated text samples from the 2,221 gathered sustainability reports. We juxtapose our model to the existing ClimateBERT sentiment analysis model trained on the climate sentiment dataset (Webersinke et al., 2022) and highlight the inability of the ClimateBERT sentiment analysis model to perform well on the Promise-Constraint dataset[2], indicating the need for a better model.

Further, we map the distribution of constraint claims over the 2020–2024 period and detail their frequency across four primary sectors: natural resources, manufacturing, retail, and information. The distribution of constraint and promise claims across sectors reveals how companies frame ESG narratives differently depending on their operational domain. These distinctions shed light on recurring justifications and commitments embedded in corporate ESG disclosures.

Our study is significant for research and application purposes. Analyzing the constraint statements facilitates further insights into the company's greenwashing strategies. It provides a method to keep the companies accountable when the identified constraints are insufficient for the company to drop their reduction targets. Overall, the findings directly affect policymakers, investors, and organizations. Policymakers can utilize insights from this study to develop more stringent disclosure requirements and assess corporate compliance with ESG standards. Investors, in turn, can leverage the findings to make informed decisions by identifying discrepancies or overstatements in sustainability claims.

[2]The Promise-Constraint dataset via HuggingFace (the link will be added after the review).

2 Background

This study builds upon a growing body of research that applies language models to environmental text analysis, including claim detection (Bulian et al., 2020; Coan et al., 2021a; Rolnick et al., 2022), fact-checking (Luo et al., 2020; Webersinke et al., 2022), and sentiment analysis (Webersinke et al., 2022).

Central to our methodology is ClimateBERT, a language model pre-trained on climate-related corpora (Webersinke et al., 2022). Among general-purpose language models, ClimateBERT demonstrates superior performance on downstream tasks like sentiment analysis, fact verification, and classification. Its use on the Climate Sentiment dataset is particularly relevant, where paragraphs are labeled as expressing a positive opportunity, negative risk, or neutral stance. However, our focus diverges from sentiment framing and instead centers on identifying constraints and promises explicitly stated by companies in sustainability disclosures.

Unlike ClimateBERT-Climate-Sentiment, which was trained on paragraph-level sentiment categories, our ClimateBERT-Promise-Constraint model is fine-tuned on sentence-level annotations. This granularity allows the model to capture precise linguistic signals linked to specific ESG commitments or limitations. As a result, our model is better suited for detecting narrowly defined ESG claims, offering a more accurate alternative to sentiment-driven classifiers that are ill-equipped for this task.

Prior works have also addressed environmental claim detection at the sentence level. For example, Stammbach et al. (2023) trained a binary classifier using ClimateBERT to detect environmental benefit statements or pledges. Schimanski et al. (2023) introduced ClimateBERT-NetZero, a model trained to classify sentences as referencing a net-zero tar-

get, a reduction goal, or neither. These efforts align with ours in leveraging sentence-level supervision to uncover climate-specific assertions.

Our study introduces two targeted categories of ESG discourse: promise claims and constraint claims. Promise claims are typically forward-looking and characterized by vague or non-measurable language. They often lack enforceable mechanisms or timelines, relying instead on optimistic phrasing, such as the use of future simple or present continuous tenses. These statements indicate actions in progress or planned but seldom provide sufficient operational detail.

This framing echoes the "cheap talk" concept in the climate finance literature (Bingler et al., 2024), where firms issue vague declarations that appear substantive but are not verifiable or enforceable. While not all promise claims are devoid of operational value, statements such as “net-zero Scope 1 and 2 emissions company-wide by 2035” can be concrete when well-scoped, many however lack clarity on implementation, financing, or governance. Our model is designed to surface this rhetorical ambiguity by identifying commitments that lack accountability mechanisms.

Constraint claims, by contrast, are concrete statements identifying barriers such as technological, financial, regulatory, or societal that hinder the achievement of sustainability targets. For instance, a company may note that the absence of viable technology limits its ability to reduce emissions. This differs fundamentally from the "negative risk" sentiment label in Webersinke et al. (2022), which captures generalized pessimistic framing (e.g., climate threats or costs); and contrarian claims (Coan et al., 2021b) that refer to statements that challenge or reject mainstream climate science or policy solutions, often associated with climate change misinformation and delay rhetoric. In contrast, constraint claims focus not on sentiment or climate change denial but on specifying a factual barrier tied to an ESG objective.

It is crucial to underscore that our model and dataset are not designed for sentiment classification and do not serve as proxies for it. The Climate Sentiment dataset (Webersinke et al., 2022) aims to map tone or risk framing, whereas we propose a distinct classification framework that isolates two types of actionable corporate statements: promises and constraints.

This methodological distinction introduces a new lens for analyzing ESG disclosures grounded not in the overall tone of a paragraph but in the presence or absence of concrete commitments and justifications. Our approach enables sharper insights into the ways companies articulate accountability, negotiate expectations, and frame their sustainability efforts.

Finally, this work fits within broader explorations into automated claim verification Leippold et al. (2024) and ESG-related fact-checking (Diggelmann et al., 2020; Gencheva et al., 2018; Subramanian et al., 2019). However, constraint detection remains underexplored in this space. By focusing on this overlooked category, we contribute a critical piece to the larger puzzle of tracking and evaluating ESG narratives—supporting more informed decision-making by both public and private stakeholders (Schimanski et al., 2024).

3 Dataset

Sentence frequencies		
Sector	**Initial Set**	**Final Labeled Set**
Manufacturing	47027	2065
Natural Resources	39802	1430
Retail	13904	748
Information	87957	1143

Table 2: Statistics of sentences before they are filtered and manually checked and after the labels are assigned.

3.1 Data Gathering Process

We obtain 2,221 sustainability reports from 1,898 companies via an open source platform[3], which allows to download specific records gathered from enterprises around the world. Thus, our data is not geographically restrained and provides rich metadata for further exploration. While the corpus includes reports from companies worldwide, the majority are authored in English. Consequently, our dataset’s applicability may be limited when analyzing non-English corporate disclosures, potentially affecting the generalizability of our findings to non-English-speaking regions.

The sustainability reports cover four main sectors: natural resources, manufacturing, retail, and information. The selection of the four industry sectors was guided by their substantial contributions to global greenhouse gas (GHG) emissions and their critical roles in the global economy, making them

[3] Sustainability Reports’ Repository

particularly pertinent for Environmental, Social, and Governance (ESG) disclosures.

Natural resources encompasses agriculture, forestry, and land use, is responsible for approximately 18% of global GHG emissions, primarily due to activities like deforestation and agricultural practices (Ritchie, 2020). Based on the report by Ritchie (2020) the manufacturing category includes food & tobacco, paper & pulp, machinery, iron & still, non-ferrous metals, chemical & petrochemical industries which together account for approximately 24.2% of global GHG emissions. The manufacturing sector's emissions stem from energy-intensive processes and the use of fossil fuels in production activities (Ritchie, 2020). While direct emissions from retail operations are relatively lower, the sector's Scope 3 emissions (those that come from its supply chain) are significantly higher. For instance, in the retail sector, Scope 3 emissions can be up to 92 times higher than direct operational emissions, highlighting the sector's extensive indirect environmental impact [4].The Information and Communication Technology (ICT) sector contributes to approximately 2% to 4% of global GHG emissions, a figure that is expected to rise with the increasing demand for digital services and data centers [5].

An important distinction from other datasets is that our collection is restricted to sustainability reports published between 2020 and 2024. This specific time frame was chosen due to the significant impact of global events on corporate sustainability strategies and reporting practices during this period. The onset of the COVID-19 pandemic in early 2020 disrupted global supply chains and heightened stakeholder awareness of corporate resilience and social responsibility. Empirical studies have shown that companies increased their sustainability reporting during the pandemic, with average ESG scores rising from 53% pre-COVID-19 to 62.3% during the pandemic period, indicating an increase emphasis on sustainability disclosures (Nakpodia et al., 2024).

Additionally, geopolitical events such as the Russo-Ukrainian war, escalating U.S.-China tensions, and shifts in international trade policies have influenced corporate sustainability priorities. These developments have led to increased regulatory scrutiny and a reevaluation of sustainability commitments, prompting companies to adjust their ESG strategies accordingly [6].

Focusing on 2020–2024 period, our dataset captures a transformative era in sustainability reporting characterized by amplified transparency and expressing commitments and constraints. This focus allows a nuanced analysis of how external crises and geopolitical shifts have shaped corporate sustainability narratives.

Despite the deliberate sector selection and time period, the resulting dataset exhibits a notable class imbalance, with promise claims significantly outnumbering explicit neutral or constraint statements. To address potential issues arising from this imbalance, our evaluation metrics explicitly incorporate macro-averaging techniques, ensuring that model performance adequately represents minority classes and accurately reflects the nuanced reality of ESG-related corporate communication.

3.2 Data Processing

To extract relevant textual content for annotation, we begin by segmenting each sustainability report using the 'SpacyTextSplitter'[7] with a maximum chunk size of 2,000 characters. Each resulting paragraph is then evaluated for climate relevance using the ClimateBERT Environmental Claims model by Stammbach et al. (2023), which produces domain-specific embeddings optimized for environmental discourse.

We compute the cosine similarity between each paragraph's embedding $\vec{e}_p$ and a manually curated centroid embedding $\vec{e}_c$ representing key climate-related themes (constructed from ClimateBERT Environmental Claims dataset (Stammbach et al., 2023)). A paragraph is retained if:

$$\cos(\vec{e}_p, \vec{e}_c) = \frac{\vec{e}_p \cdot \vec{e}_c}{\|\vec{e}_p\| \|\vec{e}_c\|} \geq \tau$$

where $\tau = 0.78$ is the empirically chosen similarity threshold maximizing recall for relevant content based on a development subset. This initial filtering step yielded 140,690 candidate paragraphs.

The selected paragraphs are tokenized into individual sentences using the spaCy transformer pipeline[8], which is optimized for high-accuracy sentence boundary detection. Sentences containing fewer than three words, non-ASCII characters, or

[4] Scope 3 Inventory Guidance

[5] World Bank. Measuring the Emissions & Energy Footprint of the ICT Sector

[6] Top Geopolitical Risks.

[7] SpacyTextSplitter by LangChain library.

[8] spaCy Sentence BERT.

formatting artifacts are discarded. At this stage, 108,520 sentences remain.

Given the substantial volume of remaining data, we implemented a pre-annotation filtering procedure based on large language models. We employ the GPT-3.5-turbo model (Brown et al., 2020) to perform few-shot filtering, grounded in annotation guidelines derived from Stammbach et al. (2023) and Bingler et al. (2024). We construct a domain-specific prompt (see Figure E) and filter out sentences classified with high confidence as "None" (i.e., not promise, constraint, or related neutral claims) where model confidence exceeds 70%. The prompt includes examples for each class and a forced-choice format, minimizing ambiguity. This step retains 27,334 sentences.

We emphasize that the use of GPT-3.5-turbo is not due to an assumed bias in lightweight Transformer models per se, but due to domain coverage limitations of off-the-shelf models. While ClimateBERT Environmental Claims is trained on climate-related corpora, it is optimized for general ESG disclosure binary classification tasks, and lacks sentence-level supervision for constraint-specific language. A study by Garrido-Merchán et al. (2023) shows that Transformer-based models, when fine-tuned on paragraphs or non-specialized claim tasks, tend to oversample policy pledges and under-identify subtle constraint indicators (e.g., regulatory hurdles or supply chain risks). The GPT-based pre-filtering thus acts as a semantic gatekeeper to ensure only well-formed, contextually relevant sentences proceed to the human labeling phase.

Once filtered, the dataset consists of 27,334 sentences presented to three annotators: two professional linguists and one with an environmental policy background. Annotators receive sector-specific sentence batches along with paragraph-level context for anaphora resolution. The annotation task involves assigning each sentence a label: Promise, Constraint, or Neutral. The complete set of annotation instructions is detailed in Appendix B. After this stage 5,299 sentences are kept as relevant for the task and manually annotated.

Labels are finalized using a majority vote mechanism, with ties adjudicated by the lead author. Krippendorff's alpha for inter-annotator agreement is 0.84, demonstrating strong reliability. The majority vote accounts for 83% of the final labels, while 17% required adjudication. Importantly, all GPT-generated labels are discarded prior to human labeling to prevent priming effects (Richmond and Burnett, 2022). The distribution of annotated labels across sectors is shown in Figure 2.

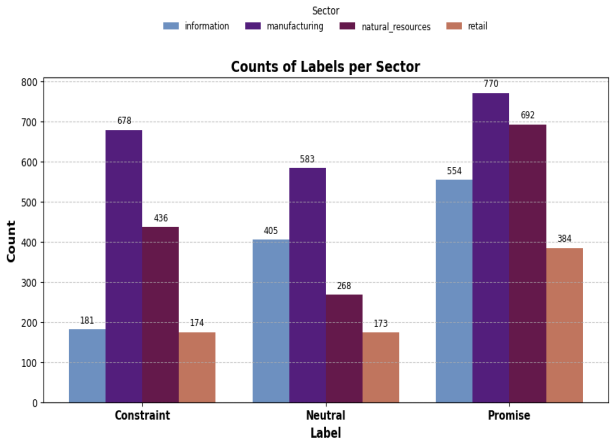

Figure 1: The statistics of label counts per sector.

4 ClimateBERT-Promise-Constraint Classifier

We evaluate three types of models: traditional machine learning classifiers trained on manually labeled data, transformer models fine-tuned on the climate sentiment dataset, and transformer models fine-tuned directly on our Promise-Constraint dataset. The goal is to assess whether domain-adaptive pretraining or task-specific fine-tuning yields better performance for identifying discrete corporate ESG claims.

Among the traditional models, logistic regression (Lever et al., 2016), support vector machines (SVM) (Hearst et al., 1998), and XGBoost (Chen and Guestrin, 2016) are trained on feature vectors generated using CountVectorizer and sparse TF-IDF representations from the Promise-Constraint data. These classifiers show strong baseline performance (macro-F1 scores around 0.74), demonstrating that lexical patterns alone can reasonably separate the three label categories. Their simplicity and competitive performance make them suitable for lightweight deployment, though they may struggle to capture nuanced ESG phrasing.

We then compare these baselines to several BERT-family models (RoBERTa_base (Liu et al., 2019), RoBERTa_large (Liu et al., 2019), DistilRoBERTa (Sanh et al., 2019)) fine-tuned on the climate sentiment dataset. These models perform

Model name	F1	Accuracy	Precision	Recall
Machine Learning Models Trained on Promise-Constraint Data				
Logistic Regression	**0.7486 ± 0.0098**	**0.7643 ± 0.0567**	**0.7621 ± 0.0098**	**0.7488 ± 0.0907**
SVM	0.7376 ± 0.0108	0.7458 ± 0.0457	0.7427 ± 0.0167	0.7362 ± 0.0493
XGBoost	0.7457 ± 0.0045	0.7615 ± 0.0476	0.7585 ± 0.0154	0.7463 ± 0.0703
BERT Models Fine-Tuned on Climate-Sentiment Data				
RoBERTa_large	**0.5610 ± 0.0206**	**0.5946 ± 0.0267**	**0.6298 ± 0.0343**	**0.5940 ± 0.0156**
RoBERTa_base	0.5209 ± 0.0070	0.5466 ± 0.0565	0.5979 ± 0.0701	0.5463 ± 0.5442
DistilRoBERTa	0.4575 ± 0.0094	0.5266 ± 0.0489	0.5945 ± 0.0927	0.5272 ± 0.0310
ClimateBERT-Climate-Sentiment	0.1485 ± 0.0074	0.2013 ± 0.0313	0.1377 ± 0.1377	0.2004 ± 0.0214
BERT Models Fine-Tuned on Promise-Constraint Data				
RoBERTa_large	**0.8837 ± 0.0360**	**0.8860 ± 0.0358**	0.8939 ± 0.0373	**0.8780 ± 0.0348**
RoBERTa_base	0.8803 ± 0.0284	0.8841 ± 0.0277	**0.8954 ± 0.0294**	0.8731 ± 0.0273
DistilRoBERTa	0.8358 ± 0.0132	0.8375 ± 0.0133	0.8360 ± 0.0151	0.8366 ± 0.0117
ClimateBERT-Climate-Sentiment	0.8677 ± 0.0337	0.8728 ± 0.0320	0.8728 ± 0.0342	0.8610 ± 0.0314

Table 3: The table presents the macro-averaged precision, recall, F1 score, and accuracy expressed as fractions. Standard deviations are calculated from the 5-fold cross-validation results on the held-out test sample.

markedly worse on our task (macro-F1 ranging from 0.45 to 0.56), reinforcing the misalignment between sentiment classification and the identification of explicit promise or constraint claims. As our task is not about calculating tone or risk framing but about identifying concrete commitments or limitations, sentiment-optimized embeddings fail to generalize (Bingler et al., 2024; Webersinke et al., 2022).

To address this, we fine-tune the same BERT-family architectures on our Promise-Constraint dataset. RoBERTa_large achieves the highest performance (macro-F1: 0.8837), followed closely by RoBERTa_base and DistilRoBERTa. A fine-tuned version of ClimateBERT, initially developed for environmental sentiment classification, also performs competitively (F1: 0.8677). Despite slightly lower performance compared to RoBERTa_large, we log the fine-tuned ClimateBERT variant as our ClimateBERT-Promise-Constraint model. This choice reflects both practical and theoretical priorities: ClimateBERT's domain-specific embeddings offer better generalization to unseen ESG text, and using an openly available climate-focused model promotes reproducibility for future work (Webersinke et al., 2022).

Ultimately, our findings underscore the need for task-specific fine-tuning in ESG contexts. General sentiment classification objectives are poorly suited for pinpointing structurally and semantically precise claims such as promises and constraints. The Promise-Constraint model addresses this by learning sentence-level linguistic patterns that general sentiment models overlook.

5 Conclusion

The study introduces a comprehensive dataset encompassing both the promises companies make toward achieving sustainability goals and the constraints they identify as hindrances to reaching net-zero targets. We analyze sustainability records from four key industries: manufacturing, natural resources, retail, and information. Although our dataset covers a limited period (2020–2024), it effectively exposes the major risks and barriers related to emission reduction objectives. We present a fine-tuned version of ClimateBERT that performs well in detecting these constraints and promises, alongside a sector-based analysis of constraint and promise prevalence. Future work may explore finer-grained discourse strategies or the rhetorical framing companies use to construct accountability in sustainability reporting.

Limitations

Despite the listed contributions, the study has several limitations.

1. The dataset is limited to 2020 to 2024 and does not account for earlier texts that may provide additional context regarding the constraints.

2. The fine-tuned ClimateBERT model, while showing high performance on our dataset, might not scale well to other similar datasets, such as sentiment analysis. Moreover, the

model is fine-tuned on the sentence level and may not be as robust on the paragraphs.

3. The GPT-based filtering of the sentences may have excluded some samples pertinent to our studies. Hence, the dataset may be limited to only the instances that contain explicit constraint and promise claims, discarding other potentially relevant data points.

4. The reliance on English-language reports excludes valuable insights from non-English corporate disclosures, potentially limiting the generalizability of the findings.

Ethics Statement

This research adheres to the ethical standards in data collection, annotation, and analysis. The dataset comprises publicly available sustainability reports, ensuring no breach of confidentiality or proprietary information. Annotators were provided with detailed guidelines to minimize subjective biases, and all annotations underwent rigorous quality checks. We explicitly specified the purpose of data annotation and realized that some bias may still be present in the data. However, the participants were positive towards the main idea of the study and dedicated much time and effort to data labeling. The findings aim to enhance corporate accountability and transparency without targeting or discrediting specific organizations. Our commitment to ethical AI extends to using models that prioritize energy efficiency and align with the sustainability objectives supporting this research.

References

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Jannis Bulian, Jordan Boyd-Graber, Markus Leippold, Massimiliano Ciaramita, and Thomas Diggelmann. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Brian Carey. 2024. Brussels sprouts more red tape on reporting. *The Times*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021a. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11(1):22320.

Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021b. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *ArXiv*, abs/2012.00614.

Eduardo C. Garrido-Merchán, Cristina González-Barthe, and María Coronado Vaca. 2023. Fine-tuning climatebert transformer with climatext for the disclosure analysis of climate-related financial risks. *Preprint*, arXiv:2303.13373.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2018. A benchmark dataset for check-worthy fact-checking claims. In Leif Azzopardi, Allan Hanbury, Gabriella Pasi, and Benjamin Piwowarski, editors, *Advances in Information Retrieval*, volume 10772 of *Lecture Notes in Computer Science*, pages 178–193. Springer.

Global Reporting Initiative and Sustainability Accounting Standards Board. 2021. A practical guide to sustainability reporting using gri and sasb standards.

Juniati Gunawan. 2023. *Sustainability Report and Sustainability Reporting*, pages 3366–3371. Springer International Publishing, Cham.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2024. Automated fact-checking of climate change claims with large language models. *Preprint*, arXiv:2401.12566.

Jake Lever, Martin Krzywinski, and Naomi Altman. 2016. Logistic regression. *Nature Methods*, 13(7):541–542.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Franklin Nakpodia, Rilwan Sakariyahu, Temitope Fagbemi, Rasheed Adigun, and Oluwatoyin Dosumu. 2024. Sustainable development goals, accounting practices and public financial management: A pre and post covid-19 assessment. *The British Accounting Review*, page 101466.

Lauren L. Richmond and Lois K. Burnett. 2022. Chapter six - characterizing older adults' real world memory function using ecologically valid approaches. In Kara D. Federmeier and Brennan R. Payne, editors, *Cognitive Aging*, volume 77 of *Psychology of Learning and Motivation*, pages 193–232. Academic Press.

Hannah Ritchie. 2020. Sector by sector: where do global greenhouse gas emissions come from? *Our World in Data*. Https://ourworldindata.org/ghg-emissions-by-sector.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2022. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. pages 15745–15756, Singapore.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. pages 1051–1066, Toronto, Canada.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019. Deep ordinal regression for pledge specificity prediction. pages 1729–1740, Hong Kong, China.

Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text. Available at SSRN: https://ssrn.com/abstract=4229146 or http://dx.doi.org/10.2139/ssrn.4229146.

Gatot Yudoko. 2024. Toward sustainable operations strategy: A qualitative approach to theory building and testing using a single case study in an emerging country. *Sustainability*, 16(21).

A Data Collection

Figure 2 illustrates the distribution of label counts derived from the collected data. Notably, 2022 has the highest number of labels, indicating the shift in the general rhetoric regarding climate change, goals, and restrictions. On the other hand, we see from Figure 3 that the counts of sentences within each sector remain consistent through the years, with a peak in 2022 and a slight reduction in 2023. This pattern may be explained by the COVID-19 pandemic in 2020–2021, during which some companies either chose not to issue sustainability reports or concentrated on highlighting how lockdowns temporarily alleviated environmental pressures. As anticipated, the United States of America dominates the dataset with the largest share of sustainability reports per sector, followed by Japan, the United Kingdom, Canada, China, and South Korea. Within the European Union, Germany and France emerged as the most proactive nations concerning ESG disclosures.

B Annotation Guidelines

We adopt the annotators' guidelines described in Stammbach et al. (2023) paper for environment claim detection. Each annotator was presented with the subsets of data per sector after the GPT filtering of the sentences. The timing between each sector annotation encompassed 4 days. Overall, it took 3 to 5 days to annotate one subset of data. This way, we reduce annotators' fatigue and give them time to focus on other topics. Unlike other studies, we did not directly involve people with prior experience in sustainability finance or climate change analysis. Therefore, we supplied the annotators with two

Hyperparameters Tuning				
Hyperparameters	**RoBERTa_large**	**RoBERTa_base**	**DistilRoBERTa**	**ClimateBERT-Climate-Sentiment**
learning_rate: 1e-05, num_train_epochs: 3, batch_size: 16	0.8036 ± 0.0112	0.7902 ± 0.0125	0.8046 ± 0.0131	0.7984 ± 0.0099
learning_rate: 2e-05, num_train_epochs: 3, batch_size: 16	**0.8188 ± 0.0081**	0.7935 ± 0.0112	0.8114 ± 0.0159	0.8085 ± 0.0101
learning_rate: 2e-5, num_train_epochs: 3, batch_size: 8	0.8078 ± 0.0068	**0.7948 ± 0.0107**	0.8098 ± 0.0158	0.8086 ± 0.0062
learning_rate: 3e-5, num_train_epochs: 3, batch_size: 16	0.8058 ± 0.0114	0.7926 ± 0.0086	**0.8117 ± 0.0093**	**0.8213 ± 0.0070**

Table 4: The table describes the hyperparameters grid along with the corresponding F1-Macro scores and their standard deviations. Each set of parameters was trained using an 80/20% split, and the results are reported for the test subset based on 5-fold cross-validation. The best score per model is highlighted in bold.

Evaluation of ML Models on the Sectors					
Model name	**Full dataset**	**Information**	**Manufacturing**	**Retail**	**Natural Resources**
Logistic Regression	**0.748 ± 0.0098**	0.684 ± 0.0113	**0.729 ± 0.0067**	**0.747 ± 0.0089**	**0.729 ± 0.0116**
SVM	0.737 ± 0.0108	**0.75 ± 0.0057**	0.722 ± 0.0112	0.734 ± 0.0054	0.702 ± 0.0101
XGBoost	0.745 ± 0.0045	0.731 ± 0.0116	0.676 ± 0.0108	0.691 ± 0.0076	0.644 ± 0.0042

Table 5: Exploratory analysis of data separability across subsets and the full dataset, leveraging standard machine learning techniques. The best score per model is highlighted in bold. The results are given from the F1-Macro metric.

primary papers by Stammbach et al. (2023) and Schimanski et al. (2023) to get acquainted with the topic and to look through the existing data. Moreover, the participants have either linguistic or environmental studies background, which makes them well-trained candidates for the annotation. We provided the annotators with two types of data: sentences and paragraphs to perform anaphora resolution when needed and better understand if the sentence is ambiguous. We decided not to use paragraphs as final data points because more concrete restrictions can be better inferred on the sentence level rather than paragraph.

Annotation Guidelines

Your task is to label sentences connected with some environmental topics. You must choose between three labels: Promise, Constraint, and Neutral. You need to be careful and rely exclusively on explicit ideas and stances in the sentences.

Promises - are forward-looking claims that tend to lack specific, measurable actions or mechanisms for accountability. These claims frequently rely on optimistic language, future simple and present continuous tenses, stating that some actions are either in progress or planned without defining specific criteria for their execution or time periods.

Examples of promise claims:

- We aim to be a net zero company in the entire value chain.
- We aim to reach net zero first in our own and then in our whole value chain.
- We will promote the introduction of a solar power generation system to realize the shift to renewable energy.

Constraints - sentences that mention some impediments, restrictions, or obstacles at a company, society, or governmental level that may restrict the company from fulfilling its promises. You do not need to consider descriptions of risks connected with climate change as constraints unless it is explicitly stated that some climate issues now influence a company's performance or have a visible impact on its operational abilities.

Examples of constraints:

- Accordingly, please be advised that the actual results may differ from such statements due to various changes.
- However, as the world shifts to a lower carbon economy, various federal, state, and/or provincial legislative mechanisms could cause our operational costs to increase significantly, given the industry's current reliance on natural gas.
- Various factors may result in substantially different outcomes.

The sentences that do not fall under these three categories are considered neutral.

The sample from the final dataset with the assigned labels by annotators is depicted in Table 6.

C Models Fine-Tuning

The hyperparameters grid is presented in the Table 4. We follow the most commonly adopted hyperparameter configurations for BERT-family models with the Adam optimizer (Kingma and Ba, 2014). Each configuration is evaluated through a five-fold cross-validation process to ensure robustness and minimize overfitting. The final model is selected based on the configuration that yielded the highest performance on the held-out set. We subsequently download the same fine-tuned model from the HuggingFace to confirm that the results remain consistent while re-applying the five-fold cross-validation check.This approach ensures reliability and transparency in the evaluation process for the future research.

D Environmental Impact

We understand that the task of fine-tuning language models like BERT is not new, but with the rise and omnipresence of the large language models, we consider the necessity of fostering the application of the smaller models, such as BERT, which can achieve high results while being less harmful to the environment. We use the Compute Cluster, which provides V100 GPUs with 64GB VRAM for fine-tuning and inference. Running the fine-tuning jobs and inference on the trained models took approximately 20 hours. The lightweight machine learning models were trained locally on the discrete NVIDIA GeForce RTX 4080 (Laptop, 140W) GPU. Although the production of such a laptop is also carbon-demanding, it is still less damaging to the environment than using a proprietary GPU cluster for a large language model inference. The GPT-3.5-turbo took 1 hour 30 min for the sentence distillation task. In total, we spent about 26 hours in computation time.

E AI Assistance

The AI model GPT-4o was used to make paragraph "2 Background" sound more coherent while preserving the information written by the authors. Moreover, the model was used for exploration to find information regarding greenhouse gas contributions by each sector.

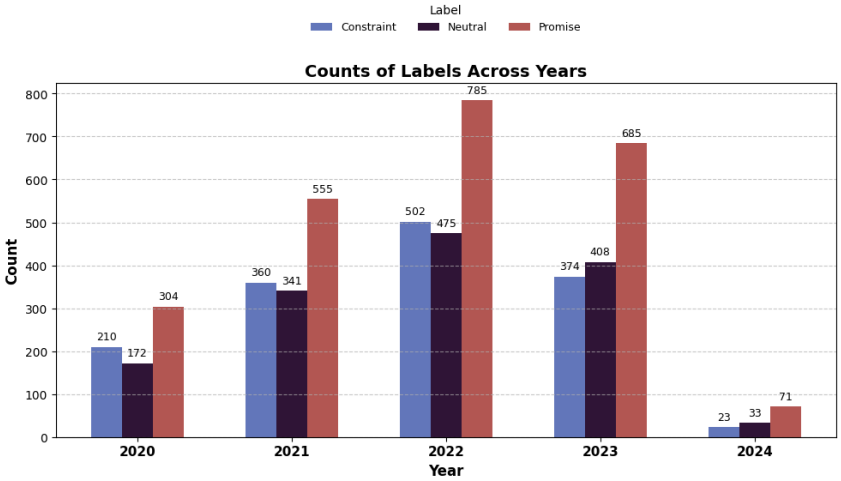

Figure 2: The distribution of labels across years.

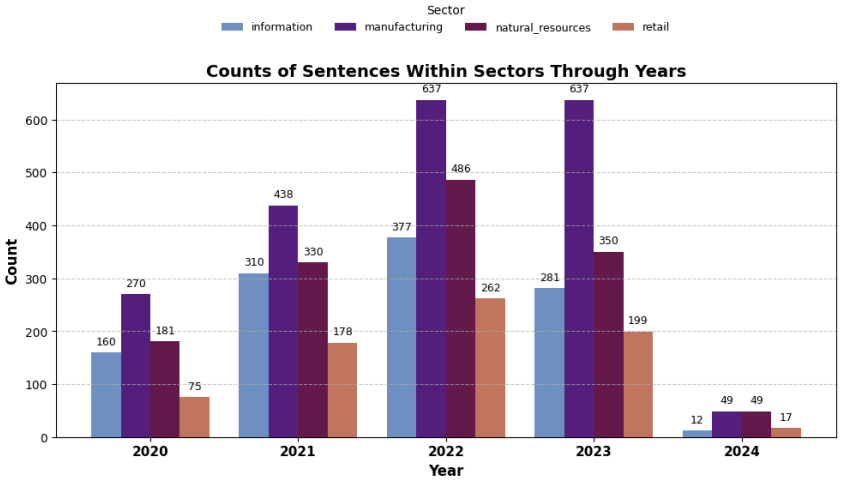

Figure 3: The statistics of sentence counts collected for each sector in a period starting 2020 to 2024.

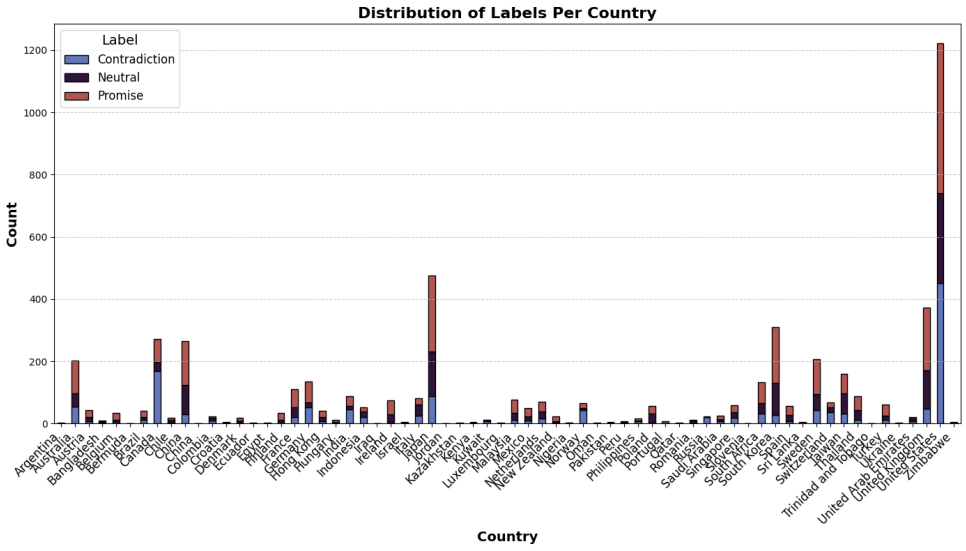

Figure 4: The distributions of promise, constraint and neutral sentences across countries.

Listing 1: Few-Shot Prompt Template

```
"""<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful assistant who assists human analysts in answering a question regarding the text.
<|eot_id|><|start_header_id|>user<|end_header_id|>

Question: {question}
Sentence: {sentence}
Assign:
- 'Promise' - a forward-looking claim that tends to lack specific, measurable actions or mechanisms for accountability. These claims frequently rely on optimistic language, future simple and present continuous tenses, stating that some actions are either in progress or planned without defining specific criteria for their execution or time periods.
- 'Constraint' - a sentence that mentions some impediments, restrictions, or obstacles at a company, society, or governmental level that may restrict the company from fulfilling its promises. You do not need to consider descriptions of risks connected with climate change as constraints unless it is explicitly stated that some climate issues now influence a company's performance or have a visible impact on its operational abilities.
- 'None' if none of the above is applicable.

## Example 1:
Question: Does the sentence contain any claim that mentions constraints, impediments or goals, targets connected with sustainability, net-zero, environmental or sustainability targets?
Sentence: Disruption of our supply chain, including increased commodity, raw material, packaging, energy, transportation, and other input costs.
[Guess]: Constraint
[Confidence]: 0.8

## Example 2:
Question: Does the sentence contain any claim that mentions constraints, impediments or goals, targets connected with sustainability, net-zero, environmental or sustainability targets?
Sentence: We are committed to achieving net-zero carbon emissions by 2050.
[Guess]: Promise
[Confidence]: 0.9

## Example 3:
Question: Does the sentence contain any claim that mentions constraints, impediments or goals, targets connected with sustainability, net-zero, environmental or sustainability targets?
Sentence: The negative impacts of, and continuing uncertainties associated with the scope, severity, and duration of the global COVID-19 pandemic and the substance and pace of the post-pandemic economic recovery.
[Guess]: Constraint
[Confidence]: 0.7

Reply in the following format:
[Guess]: <Your most likely guess, should be [Promise, Constraint, None].>
[Confidence]: <Give your honest confidence score between 0.0 and 1.0 about the correctness of your guess. 0 means your previous guess is very likely to be wrong, and 1 means you are very confident about the guess.>
Answer: <|eot_id|><|start_header_id|>assistant<|end_header_id|>"""
```

Sentence	Label
If the risks or uncertainties ever materialize or the assumptions prove incorrect, the results of HP Inc. and its consolidated subsidiaries (HP) may differ materially from those expressed or implied by such forward-looking statements and assumptions.	Constraint
Actual results could differ materially due to factors such as the availability of funding for the programs described in this Report.	Constraint
Just as the global pandemic's effects began to ease, the invasion of Ukraine worsened the supply chain disruptions, caused the inflation to surge, heightened political tensions, polarized, and fractured international relations.	Constraint
Suppliers may find it burdensome to install additional equipment, get assessments, or receive certificates to improve their environment.	Constraint
Due to the high growth of the battery industry, the global production capacity of LG Energy Solution is rapidly increasing every year, and the resulting increase in energy use makes it difficult for us to achieve our carbon neutrality.	Constraint
Geopolitical risks also pose a difficulty for collaboration between nation states to progress towards common goals such as climate change adaptation and mitigation, response to pandemic, access to medicines, etc.	Constraint
These targets were not met due to COVID-19 related temporary production shutdowns in FY20 and operation at 75% of production capacity in FY21.	Constraint
Moreover, ongoing conflicts like the Russia-Ukraine war, are causing energy prices to rise and acting as a barrier to the renewable energy transition.	Constraint
Sustainable development is marked by considerable uncertainty because of changing expectations, the complexity of the problem, and the difficulty of its resolution.	Constraint
An analysis of reductions to date shows that meeting the target of the Paris Agreement will be challenging, requiring countries to step up their efforts.	Constraint
Our new commitment is to work toward reducing absolute and by percent by from a base year.	Promise
We will continue to improve different such as product recovery and repair as well as raw material with a focus on to the circular economy.	Promise
We aim to reach net zero by first in our own and then in our entire value chain.	Promise
At Deutsche Telekom, we have set ourselves a vital target of achieving net zero emissions along our entire value chain by 2040.	Promise
The Company set ambitious for Carbon and Water Neutrality and a strategy to achieve them.	Promise
As part of this commitment, T-Mobile aims to reduce absolute emissions by 90% by 2040.	Promise
Regarding the utilization of renewable energy, we plan to start and expand solar power generation.	Promise
Under this policy, we commit to maximize renewable energy use, reduce the carbon footprint in our and work with our business to reduce across the value chain.	Promise
Our goal is to become Climate Positive by removing more carbon than we create, and assuring that future are able to enjoy the world we love.	Promise
We know that, in order to ensure the of the Group and its climate change adaptation and mitigation are absolutely necessary.	Neutral
Even in a year of great change, we continue to drive forward to reach our Net Zero.	Neutral
Flue gas capture is vital to net zero commitment.	Neutral
Meanwhile, we are actively exploring a natural way towards carbon neutrality.	Neutral
Since the competition is to grow to obtain the necessary for carbon neutrality, a stable supply chain will be an urgent task.	Neutral
Our strategy was to to drive our reduction to net zero.	Neutral
The Audit and Risk Committee (ARC) is responsible for the risk management and to mitigate key climate change.	Neutral
As the leader in our industry, our responsibility to reduce the impact of construction on our world.	Neutral
Through our commitment to net zero we demonstrate our leadership in the global climate crisis.	Neutral
The electrification of is a key step toward net zero and air pollution.	Neutral

Table 6: Examples of sentences and their corresponding labels. The full dataset also incorporates metadata with companies names, years, and four sectors.