A Dataset for Detecting Real-World Environmental Claims

Center for Law & Economics Working Paper Series

Number 07/2022

A Dataset for Detecting Real-World Environmental Claims

Dominik Stammbach
Nicolas Webersinke
Julia Anna Bingler
Mathias Kraus
Markus Leippold

September 2022

A Dataset for Detecting Real-World Environmental Claims

Dominik Stammbach
ETH Zurich
dominsta@ethz.ch

Nicolas Webersinke
FAU Erlangen-Nuremberg
nicolas.webersinke@fau.de

Julia Anna Bingler
Council on Economic Policies
jb@cepweb.org

Mathias Kraus
FAU Erlangen-Nuremberg
mathias.kraus@fau.de

Markus Leippold
University of Zurich
markus.leippold@bf.uzh.ch

Abstract

In this paper, we introduce an expert-annotated dataset for detecting real-world environmental claims made by listed companies. We train and release baseline models for detecting environmental claims using this new dataset. We further preview potential applications of our dataset: We use our fine-tuned model to detect environmental claims made in answer sections of quarterly earning calls between 2012 and 2020 – and we find that the amount of environmental claims steadily increased since the Paris Agreement in 2015.

Keywords Environmental Claims · Climate Change Dataset · Natural Language Processing

1 Introduction

The potential for companies to be held liable for their impact on the environment and society is increasing. This trend is due to changing public and investor attitudes to global warming, as well as proposed reporting requirements on Environmental, Social, and Governance (ESG) issues in Europe and the US. ESG is likely to become a significant source of liability exposure in the future and will continue to grow, particularly around disclosures related to climate change and the environment. Hence, this paper introduces a dataset with 3K human-annotated environmental claims made by listed companies. In Figure 1, we show 4 examples from the dataset. For constructing the dataset, we were inspired by the European Commission (EC), which defines such claims as follows: *Environmental claims refer to the practice of suggesting or otherwise creating the impression (in the context of a commercial communication, marketing or advertising) that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services.*[1] While such claims can be truthful and made in good faith, boasting about environmental credentials can also be monetized [1]. For example, consumers are willing to spend more money on environmentally friendly products [2], and fairly produced products in general [3]. The Commission states if environmental claims are too vague, unclear, or altogether misleading, we are confronted with an instance of "greenwashing".

Our dataset contains environmental claims by firms, often in the financial domain. We collect text from sustainability reports, earning calls and annual reports. Our particular focus on the financial domain is motivated by the fact that financial markets play a crucial role in combating climate change. These markets can allocate capital efficiently to finance the energy transition. Investment flows into sustainable strategies have accelerated since the COVID-19 pandemic, and global assets managed under a sustainability label are on track to exceed $53 trillion by 2025, more than a third of total assets under management. However, unfortunately, the boom has been accompanied by rampant greenwashing, with companies boasting about their environmental credentials.[2]

[1]From the Commission Staff Working Document, Guidance on the implementation/application of Directive 2005/29/EC on Unfair Commercial practices, Brussels, 3 December 2009 SEC(2009) 1666. See section 2.5 on misleading environmental claims.

[2]See, e.g., The Economist, May 22nd, 2021.

Environmental claim: A total population of 6148 is getting the benefit of safe potable drinking water due to this initiative.
Environmental claim: Hydro has also started working on several initiatives to reduce direct CO2 emission in primary aluminium production.
Negative example: Generally, first of all our Transmission department is very busy, both gas and electric transmission, I should say, meeting the needs of our on-network customers.
Negative example: Teams are thus focused on a shared objective in terms of growth and value creation.

Figure 1: Environmental Claims and Negative Examples from the dataset.

We situate the task of detecting environmental claims at the intersection of claim detection, for example [4] and pledge detection [5, 6]. An environmental claim is produced to increase the environmental reputation of a firm or a product. We show that models trained on current claim and pledege detection datasets perform poorly at detecting environmental claims, hence the need of this new dataset.

The dataset, baseline models and associated code to train baseline models is available on github.[3] We will also release the dataset[4] and models[5] via huggingface Transformers [7]. In the reminder of this paper, we describe the dataset creation process, investigate models trained on the dataset and present a small case study displaying the share of environmental claims uttered in quarterly earning calls between 2012 and 2020.

2 Related Work

This work is part of an ongoing effort at the intersection of environmental topics, climate change related topics and NLP. Resulting datasets and methods assist the research community to investigate such topics at scale using computer assistance. Methods include for example ClimateBERT, a language model pre-trained on climate-related text [8]. NLP tasks and datasets include climate change topic detection [9], claim verification of climate change related claims [10], detecting media stance on global warming [11], and the analysis of regulatory disclosures [12].

Some work focuses on detecting environmental commitments and sustainable sentences. For example, [13] analyze websites of firms in the U.S. metal industry and investigate whether textual environmental commitments of firms correlate with SO_2 concentrations from satellite data. They also employ text as data methods to automatically predict these commitments. In one of the subtasks of FinSim4[6], participants were asked to classify sentences into sustainable or unsustainable sentences [14]. While both are related to this work, they are not strictly concerned with detecting environmental claims as we do in this paper.

We situate environmental claim detection at the intersection of claim spotting and pledge detection, but in the domain of text produced by listed companies with the goal of boosting their environmental credentials. Claim spotting is the task of finding fact-check worthy claims [4, 15, 16]. Pledge detection aims to detect pledges made in, for example, political campaigns [5, 6]. Environmental claims trivially are a sort of claim, thus the connection to claim detection. Also, they convey an intention for a material impact (some sort of environmental improvement) which would benefit the audience of the pledge (the consumer). In this work, we are interested in detecting environmental claims made by listed companies, primarily in the financial sector. In this setting, an environmental claim is usually made to increase a company's reputation and gain a more environmental-friendly reputation.

Our data is annotated by 16 domain experts.[7] The authors drafted annotation guidelines in an iterative process and added examples of clear and borderline environmental claims and negative examples to the guidelines. Employing domain experts, a moderate inner annotator agreement, and carefully drafted annotation guidelines lead us to believe that our dataset is of high quality.

[3] `https://github.com/dominiksinsaarland/environmental_claims`
[4] `https://huggingface.co/datasets/climatebert/environmental_claims`
[5] `https://huggingface.co/climatebert/environmental-claims`
[6] `https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022/shared-task-finsim4-esg`
[7] All annotators passed a core course on sustainable investing with a high grade. This course is part of the executive education program for the Master of Advanced Studies in Sustainable Finance, offered by the University of Zurich. Most of the annotators have prior work experience in the financial sector.

split	# examples	mean length	claims (%)
train	2400	24.4	0.23
dev	300	24.4	0.21
test	300	24.8	0.21
all	3000	24.5	0.22

Table 1: Dataset Statistics

We provide dataset statistics in Table 1. While it is not a large-scale dataset, this is the result of a conscious decision to prioritize quality over quantity by employing experts to annotate the data. Because such annotations are costly, our final dataset consists of 3K examples in total. In Section 3, we show that performance of models slowly converges after being trained on more than 60% of the training set, hence our decision to stop annotation here and release it with 3'000 annotated examples in total.

In pilot studies, we decided to only keep sentences having more than 10 and less than 40 words[8], leading to a mean length of 24.5 words per sentence. To extract the sentences annotated in our dataset, we use a preliminary model to sample candidate sentences from various text sources produced by firms. Furthermore, we randomly sample sentences from different clusters obtained with k-means to increase coverage of the domain. We describe the sampling process of the dataset in detail in Appendix B.

We annotate claims on sentence-level. We acknowledge that sometimes there is not enough context to make a determination whether a sentence is an environmental claim. However, such claims usually try to increase the environmental reputation of a firm in a clear and concise fashion, so we believe sentence-level annotations to be sufficient for the majority of such claims. Thus, we note in the annotation guidelines that sentences requiring more context should not be annotated as claims, see Appendix Table 4.

Each sentence is assigned to four annotators. The annotations are aggregated by majority vote. 60% of the 3K samples were decided unanimously by the annotators, and 88% of the annotations made were part of a majority decision. Less than 12% of the samples could not be resolved this way. The overall inter-annotator agreement measured in Krippendorff's alpha is 0.47, indicating moderate agreement. The authors took an effort to fix the 12% of the samples with tied annotations. We show the complete annotation workflow in Appendix Figure 4.[9]

3 Experiments

We conduct two types of experiments: (1) We train baseline models for detecting environmental claims, using standard transformer models [17, 18, 19, 8], followed by a small error analysis and a study on dataset size. And (2), we apply our models on text produced by listed companies, which leads to a small case study demonstrating the intended use case of the dataset.

3.1 Environmental Claim Detection Models

We report micro- and macro-F1 performance for a 5-fold cross-validation split of the whole dataset, and the development and test set in Table 2. We present results for five baselines and four transformer models. We find that both a random and a majority baseline perform poorly. Next, we fine-tune RoBERTa$_{\text{base}}$ model on ClaimBuster data [4], and use this model to detect environmental claims in the dataset[10]. However, we find that the model transfers poorly to predicting environmental claims, hence the need for a dedicated dataset and specialized models for this task. We also train a RoBERTa$_{\text{base}}$ model on a Pledge Detection dataset [5]. While this model seems to adapt better to detecting environmental claims than the models discussed until here, performance is still only slightly better than random guessing.

Support vector machines [20] already achieve an acceptable macro-F1 score of over 70% on all data splits using tf-idf features. This indicates that the choice of environment related keywords is indicative of whether a sentence is an environmental claim. However, all transformer models explored in this study outperform the SVM by a large margin. We take this as evidence that the presence of environmental keywords alone is not sufficient to determine whether a sentence is an environmental claim or not.

[8]Shorter sentences are often section titles. We found that longer sentences usually are the result of a failure in preprocessing.

[9]In Appendix C, we list the annotation guidelines along with examples and rationales which the authors discussed in pilot annotation rounds, and which were given to the annotators. We release the dataset with a train, development, and test split which allows to replicate our experiments and other researchers and practitioners to contribute to the field of environmental claim detection.

[10]We reduce the task to a binary classification task and train the model to distinguish fact-checkworthy claims vs. all other claims. The model works exceptionally well on the ClaimBuster testset with a micro-F1 of 97.9% and a macro-F1 of 97.0%.

model	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$
	CV		dev		test	
majority	77.7	43.7	80.0	44.4	79.0	44.1
random	50.7	46.6	48.3	43.3	47.0	41.4
ClaimBuster RoBERTa	49.4	46.3	48.0	43.8	55.3	50.9
Pledge Detection RoBERTa	65.1	52.4	66.0	49.4	59.3	46.3
tf-idf SVM	85.5	77.4	86.3	77.3	81.7	72.5
DistilBERT	90.2	86.6	90.3	86.3	84.7	80.1
ClimateBERT	90.5	87.2	**92.0**	**88.6**	86.7	81.9
RoBERTa$_{\text{base}}$	90.6	87.4	91.7	**88.6**	85.7	80.8
RoBERTa$_{\text{large}}$	**90.9**	**87.6**	91.3	87.9	**90.7**	**86.8**

Table 2: Main results: We report micro- and macro-F$_1$ (μ-F$_1$ and m-F$_1$) on a cross-validation split (CV), the development set split (dev) and the test set split of the environemntal claims dataset. All numbers are reported as % and best performance per split is indicated in bold.

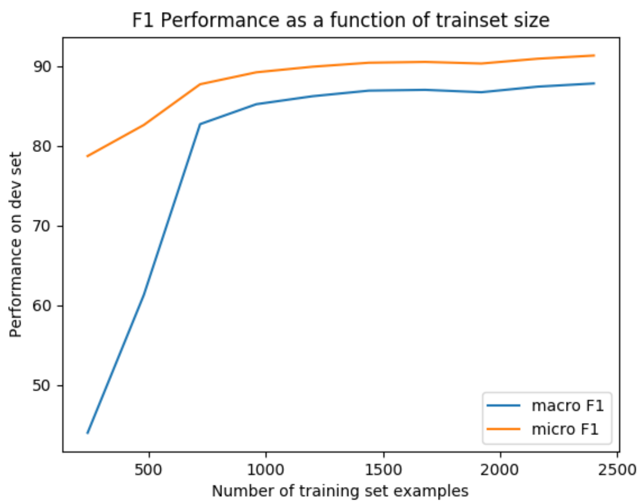

Figure 2: Development Performance of ClimateBERT fine-tuned on subsets of the training dataset

We find that DistilBERT [19] achieves a macro-F1 score of more than 80% on all data splits. We can see that the bigger models, that is RoBERTa$_{\text{base}}$ and RoBERTa$_{\text{large}}$ [18], unsuprisingly perform better than DistilBERT. We also fine-tune ClimateBERT [8], a language model pre-trained on over 1.6 million climate-related paragraphs on detecting environmental claims. This model performs slightly better than the DistilBERT which uses the same amount of parameters. Hence, further pre-training on climate-related text seems beneficial to detect environmental claims. We speculate that further pre-training bigger models on such text would also increase the performance. Overall, we observe that the different transformer models used in our experiments all achieve better results than the non-neural baselines. Specifically, all transformer models achieve a macro-F1 above 80% on each split of the dataset.

For training our models, we use Hugging Face [7] and standard RoBERTa hyper-parameters. We use the Adam optimizer with a learning rate of 2e-5, a batch size of 16 and train models for 3 epochs. The authors were surprised by the performance of their initial models on the development set of the dataset, thus there was no additional exploration of hyper-parameters.

We then perform a small error analysis which we detail in Appendix D. While we introduce a rather small dataset, Figure 2 shows that model performance as a function of dataset size converges quickly.[11] Hence, we believe that our dataset is sufficient in size and we do not expect model performance to increase drastically anymore if we were to annotate more data points.

[11]We fine-tune a ClimateBERT model on different subsets of the training data, e.g. on 10%, on 20% etc. In Figure 2, we find diminishing marginal utility after having fine-tuned a model on more 60% of the dataset.

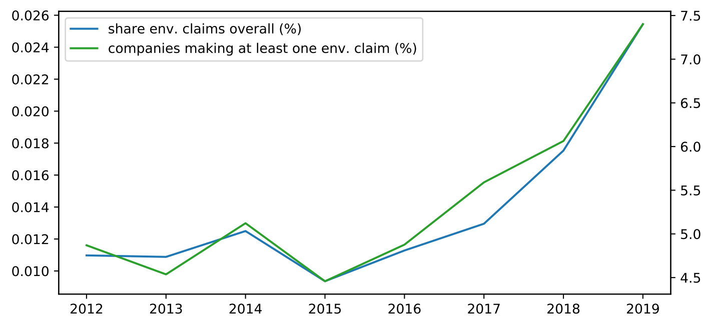

Figure 3: Amount of environmental claims (in %) made in earning calls answer sections. The blue line (y-axis on the right) shows the share of environmental claims made each year. The green line shows the share of companies making at least one environmental claim in a given year.

3.2 Earning Calls

We use our trained model to detect environmental claims in corporate earning calls between 2012 and 2020. These are conference calls between the management of a publicly traded company, analysts, investors and the media to discuss the company's financial results and other topics for a given reporting period (mainly quarterly). The conference calls consist of different segments, of which the segment with questions and answers is the most interesting for our purposes. Therefore, we focus on the management responses, which consist of 12 million sentences from 3,361 unique companies. All earnings conference call transcripts are obtained from Refinitiv Company Events Coverage. Due to the size of the data and computational constraints, we use our ClimateBERT model fine-tuned on detecting environmental claims instead of the RoBERTa$_{\text{large}}$ model.

We would expect that the amount of environmental claims made by corporations and business leaders has steadily increased since the Paris Agreement in 2015. In Figure 3, we find that this is indeed the case. The amount of environmental claims is not only increasing, the increase is also accelerating. In 2019, the amount of such claims is almost three times as high as in 2015.[12] This small case study illustrates one of the intended use cases of the dataset and its associated models: We have access to a tool that allows us to analyze environmental statements made by listed companies at scale.

4 Conclusion

Identifying corporate environmental claims is a first step in uncovering potential greenwashing activities. Given the vast and ever-growing volume of corporate disclosures, regulatory filings, and statements in the news, an algorithmic approach to detecting these claims is essential. Therefore, we introduce a dataset concerned with detecting environmental claims in text. It is a small but high quality dataset annotated by domain experts. We described the resulting dataset and the construction process. We release baseline models and show one of the intended use cases of such models in a case study involving earning calls.

We envision several directions of future work of this line of research. First, we plan to investigate "greenwashing", the practice of making a false, vague, unclear or misleading environmental claim. To make progress on this front, it is mandatory that we can detect environmental claims in the first place. Hence, our effort to publish the dataset. Second, models trained on detecting environmental claims have merits of their own, as previewed in our case study. We plan to explore more such applications in detail, e.g. analyzing annual reports and TCFD[13] reports at scale. For example, it would be interesting to see in which sections of TCFD reports firms make environmental claims. Lastly, we expect similar efforts at the intersection of environment, climate change, and NLP to increase in the near future and hope this work contributes to these efforts.

[12]In Table 5 of the Appendix, we give the 5 highest and lowest scoring sentences based on our model, and Figure 6 shows the word clouds of non-claims and environmental claims.

[13]Task Force on Climate Related Financial Disclosures

References

[1] Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, 32(1):19, Feb 2020.

[2] Nielsen Media Research. `https://nielseniq.com/global/en/insights/analysis/2015/the-sustainability-imperative-2/`, 2015. Accessed: 2022-07-04.

[3] Raluca Dragusanu, Daniele Giovannucci, and Nathan Nunn. The economics of fair trade. *Journal of Economic Perspectives*, 28(3):217–36, September 2014.

[4] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. A benchmark dataset of check-worthy factual claims. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):821–829, May 2020.

[5] Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. Deep ordinal regression for pledge specificity prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1729–1740, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Tommaso Fornaciari, Dirk Hovy, Elin Naurin, Julia Runeson, Robert Thomson, and Pankaj Adhikari. "we will reduce taxes" - identifying election pledges with language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3406–3419, Online, August 2021. Association for Computational Linguistics.

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[8] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text, 2021.

[9] Francesco Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. Climatext: A dataset for climate change topic detection. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*. NeurIPS, December 2020.

[10] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

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

[12] Julian F Kölbel, Markus Leippold, Jordy Rillaerts, and Qian Wang. Ask bert: How regulatory disclosure of transition and physical climate risks affects the cds term structure. *Journal of Financial Econometrics* (forthcoming), 2022.

[13] Sebastian Schmidt, Jan Kinne, Sven Lautenbach, Thomas Blaschke, David Lenz, and Bernd Resch. Greenwashing in the US metal industry? A novel approach combining SO2 concentrations from satellite data, a plant-level firm database and web text mining. *Science of The Total Environment*, 835:155512, 2022.

[14] Elvys Linhares Pontes, Mohamed Benjannet, Jose G. Moreno, and Antoine Doucet. Using contextual sentence analysis models to recognize esg concepts, 2022.

[15] Pepa Atanasova, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. 2018.

[16] Alberto Barron-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. 2020.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

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

[20] Bernhard Schölkopf and Alexander Johannes Smola. *Learning with Kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning series. MIT Press, 2002.

[21] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[22] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021.

[23] Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Towards climate awareness in nlp research, 2022.

A Ethical Considerations

Intended Use: This dataset will benefit researchers analyzing environmental claims made by listed companies at scale. Also, we see this as a first step towards algorithmic greenwashing using NLP methods. It might also be useful to regulators in both the financial sector and the legal domain. Next, we hope companies are inspired by our work to produce more careful environmental claims. To conclude, we envision that the dataset and related models bring large positive impact by encouraging truly environmentally friendly actions and less verbose boasting about environmental credentials.

Misuse Potential: Although we believe the intended use of this research is largely positive, there exists potential for misuse. For example, it is possible that for-profit corporations will exploit AI models trained on this dataset while drafting environmental claims.

Bias: Although the performance of NLP models can usually achieve higher than 80% F1, it is commonly known that ML models suffer from picking up spurious correlations from data. Furthermore, it has been shown that large pre-trained language models such as ClimateBERT suffer from inherent biases present in the pre-training data leading to biased models.

Data Privacy: The data used in this study are mostly public textual data provided by companies and public databases. There is no user-related data or private data involved.

B Sample Selection

The basis for selecting samples are documents from four domains in text produced by companies. We consider TCFD reports, voluntary self-disclosing by firms about their environmental impact, but not legally binding. Furthermore, we consider annual reports, comprehensive reports about activities conducted by a firm in a given year. We also consider corporate earnings calls (only the answer sections), which are conference calls between the management of a public company, analysts, investors, and the media to discuss the company's financial results and other business relevant topics during a given reporting period. Earnings conference call transcripts are obtained from Refinitiv Company Events Coverage (formerly Thomson Reuters StreetEvents). Lastly, we include the language data on environmental risks, targets and performance from the CDP disclosure questionnaire responses from 2021. We denote the universe of these documents by D_{large}.

A random selection of sentences from these documents would lead to a high number of sentences not related to environment, thus, is impracticable. We also decided against using a keyword search to pre-filter D_{large} for two reasons. If we use a keyword set which is too narrow, we might have dataset artifacts. On the other hand, if we use a set which is too loose, we might again end up with too many non-climate related sentences, which again is impracticable.

As a remedy, we start with a handpicked selection of 250 environmental claims used in a recent marketing study about greenwashing in French investment funds by 2DII, an independent, non-profit think tank working to align financial markets and regulations with the Paris Agreement goals. We also consider 200 non-environmental claims as negative samples, randomly sampled from company websites. The authors translated them to English (if necessary) and losely annotated these sentences to double-check their quality and to help come up with annotation guidelines. However, these 450 sentences do not appear in the final version of the dataset. Next, we train a preliminary $\text{RoBERTa}_{\text{base}}$ model on this dataset and use this trained model to compute the likelihood of each sentence from D_{large} being an environmental claim. Using this likelihood, we use the following strategy to select both samples with a high chance of being environmental claims, samples with a low chance of being environmental claims, and samples which are semantically similar but lead to very different results compared to our base transformer model:

1. First 300 samples were sampled, which are adjacent to our starting selection of 250 environmental claims in SBERT embedding space [21], but for which the base transformer model assigned a small score of being an environmental claim.
2. Then, 1500 samples with a score greater than 0.7 from our preliminary transformer model are selected.
3. Next, 500 samples with a score between 0.2 and 0.5 from our preliminary transformer model are selected.
4. Then, we selected 200 samples with a score lower than 0.2 from our preliminary transformer model.
5. Finally, all encoded samples from SBERT are clustered into 2000 clusters using k-means. The largest clusters, from which no sample was selected in steps 1-4, are then represented by a random sample from the cluster.

This way we increase the coverage of the whole domain by our selected samples. We selected 500 samples with that strategy.

While we tried to maximize domain coverage using this sampling procedure given the limited annotation budget, it is likely that we missed lots of utterances of environmental claim. Also, the sample is somewhat biased towards our preliminary model, which we used to sample environmental claims from. Moreover, we did not include all domains of text produced by listed companies. For example company websites and advertisements are not included in our universe of documents.

C Annotation Guidelines

Your task is to label sentences. The information we need is whether they are environmental claims (yes or no).

A broad definition for such a claim is given by the European Commission: *Environmental claims refer to the practice of suggesting or otherwise creating the impression [...] that a product or a service is environmentally friendly (i.e. it has a* ***positive impact*** *on the environment) or is* ***less damaging*** *to the environment than competing goods or services [...]*

In our case, claims relate to **products, services OR specific corporate environmental performance.**

General annotation procedure / principles:

- You will be presented a sentence and have to decide whether the sentence contains an **explicit** environmental claim.
- Do not rely on implicit assumptions when you decide on the label. Base your decision on the information that is available within the sentence.
- However, if a sentence contains an abbreviation, you could search online for the meaning of the abbreviation before assigning the label.
- In case a sentence is too technical/complicated and thus not easily understandable, it usually does not suggest to the average consumer that a product or a service is environmentally friendly and thus can be rejected.
- Likewise, if a sentence is not specific about having an environmental impact for a product or service, it can be rejected.
- Final goal: We will train a classifier on these annotations and apply it to massive amounts of financial text to explore which companies/sectors at which time make how many environmental claims. Does the number of environmental claims correlate with sectors/companies reducing their environmental footprint?
- The annotation task is not trivial in most cases. Borderline decisions are often the case. If you are uncertain about your decisions, copy-paste the sentence and add an explanatory note to the sentence. We will then cross-check it in case needed.

In Table 3 and 4, we show examples which were discussed within the author team.

Figure 4 shows our annotation process containing two steps: First, four expert annotations are collected for each of the 3000 selected sentences. In case of a clear majority of the annotators for a sentence (4:0, 3:1, 1:3, or 0:4), the sentence is annotated as such. In case of no majority (2:2), two authors annotate the sentence again. If both authors annotate the sentence as *positive*, the sentence is annotated as positive in our dataset. Else, it is annotated as negative. The rational behind this is, that a sentence annotated as *positive* accuses the writer to claim something. This accusation should be agreed on by the majority of readers (in dubio pro reo - in doubt, rule for the accused).

D Error Analysis

We then perform a small error analysis on all the miss-classified examples by the RoBERTa$_{\text{large}}$ model on the testset. In 40% of the errors, the annotation agreement was 50%, e.g., 2 annotators labeled the example as a claim, two annotators did not. As discussed, the authors re-annotated these examples, but still did not agree whether it is a claim, hence they are treated as non-environmental claims. Given this procedure, it is expected that the model might miss-classify such sentences where annotators were divided as well. To provide additional insights, we discuss 3 miss-classified examples where the annotators agreed unanimously in the following.

1. *Renewables will play a major role in tackling global greenhouse gas emissions and mitigating climate change, the defining issue of our time.* **label:not a claim** – While this sentence contains lots of words related to the environment, it is a pure description of Renewables, and not an environmental claim.

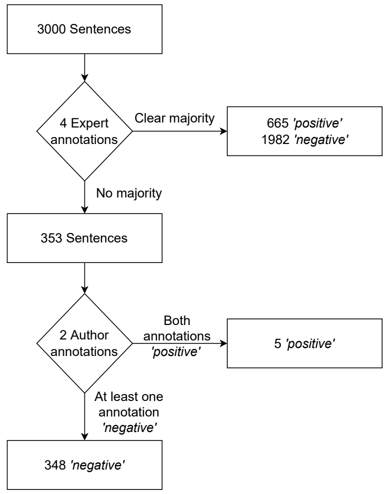

Figure 4: Our two step annotation process. *positive* describes sentences annotated as environmental claims, whereas *negative* describes no environmental claims.

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	1 (0.98)	0	1.00	#s Ren ew ables will play a major role in tackling global greenhouse gas emissions and mitigating climate change , the defining issue of our time . #/s

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	1 (1.00)	0	1.00	#s Despite the numerous challenges we faced during the year , K V PL and T TE maintained sustainability commitments to preserving and enhancing the quality of our natural capital . #/s

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	0 (1.00)	1	1.00	#s We are listed on the major sustainability market indices , such as the DJ SI , the BM & F BO V ES PA Corporate S ustain ability Index (local acronym IS E) and the Carbon E fficient Index (local acronym ICO 2). #/s

Figure 5: Layer-Wise Relevance Propagation of transformer models for miss-classified (unanimously annotated) examples in the test set. We find that models focus on environmental concepts if they predict a sentence to be an environmental claim.

2. *Despite the numerous challenges we faced during the year, KVPL and TTE maintained sustainability commitments to preserving and enhancing the quality of our natural capital.* **label: not a claim** – Again, we find words related to environment in this claim, e.g. "sustainability", "commitments", "preserving", "natural", further highlighting the difficulty of the task.

3. *We are listed on the major sustainability market indices, such as the DJSI, the BM&FBOVESPA Corporate Sustainability Index (local acronym ISE) and the Carbon Efficient Index (local acronym ICO2).* **label: environmental claim** – Our interpretation is that the model does not handle the proper nouns and acronyms well in this example.

In Figure 5, we show the word importance leading to the model's decision in these sentences. To visualize these importances, we use Layer-Wise Relevance Propagation for Transformers by [22].

E Environmental Impact

In this section, following [23] we describe the environmental impact of our dataset construction and experiments. All experiments were conducted on a carbon-neutral computing cluster in Switzerland, using a single Nvidia GeForce GTX

Figure 6: Word clouds of non-claims (on the left) and environmental claims (on the right) in earnings call transcripts.

1080 Ti GPU with a TDP of 250 W.[14] While the computing cluster we performed the experiments on is superficially carbon-neutral, there are still emissions for the production and shipping of the hardware used. Also, the energy used for our experiments could replace power produced by fossil fuel somewhere else. Therefore, we calculate emissions based on the country's energy mix.

Running the main experiments took less than 1 hour combined. Detecting environmental claims in the quarterly earning calls took an additional 3 hours. For preliminary experiments, we trained a battery of transformer models on loosely annotated data (we used scores assigned by our "best" model to sample the sentences in the dataset). This took roughly 48 hours. Also, we embedded all sentences with SBERT for two additional hours. In total, we spent about 60 hours of computation time.

[14]See `https://www.hpc-ch.org/lake-water-to-cool-supercomputers-at-cscs`

Label	Sentence	Explanation
yes (unanimously)	Farmers who operate under this scheme are required to dedicate 10% of their land to wildlife preservation.	Environmental scheme with details on implementation
yes (borderline)	We prove our commitment to a sustainable world every day—by being a force for change where we work and live and holding ourselves and our suppliers to high standards in three vital aspects of doing business: people, product, and planet.	Very generic sustainability or responsibility wording without clear reference to environmental aspects. Yet the term "sustainability" and "responsibility" includes environmental aspects.
yes (borderline)	Our places, which are designed to meet high sustainability standards, become part of local communities, provide opportunities for skills development and employment and promote wellbeing.	No would be: "Our places, which are designed to become part of local communities, provide opportunities for skills development and employment and promote wellbeing."
yes (borderline)	Fast Retailing has adopted "Unlocking the Power of Clothing" for its Sustainability Statement, and through the apparel business seeks to contribute to the sustainable development of society.	Very generic sustainability or responsibility wording without clear reference to environmental aspects. Yet the term "sustainability" and "responsibility" includes environmental aspects.
yes (borderline)	Hermès, which is currently managed by the sixth generation of family shareholders, is aware of its social responsibility and strives to give back to the world a part of what it gives to the Company.	Very generic sustainability or responsibility wording without clear reference to environmental aspects. Yet the term "sustainability" and "responsibility" includes environmental aspects.
yes (borderline)	In 2016, UTC was placed on the CDP climate change and supplier A List, and in 2017 and 2018 received an A- and Leadership designation.	Environmetal initiatives and leadership.
yes (borderline)	Change internal behavior; Drive low-carbon investment; Identify and seize low-carbon opportunities; Stakeholder expectations.	Intangible but environmentally friendly/ier processes.
yes (borderline)	We are looking into the Insurance Underwriting element, and have taken part in the CRO Forum's Sustainability Carbon Footprinting paper of Underwriting.	Intangible but environmentally friendly/ier processes.
yes (borderline)	In a further demonstration of the importance we place on helping customers to live sustainably, we became signatories of the Task Force on Climate related Financial Disclosures, to provide consistent information to our stakeholders.	Intangible but environmentally friendly/ier processes.
yes (borderline)	As for asset, DBJ Green Building certification for 18 properties, BELS certification for 33 properties, and CASBEE certification for one property have been received.	Official environmental Labels
yes (borderline)	Our clean, safe and high-tech products and solutions enable everything from food production to space travel, improving the everyday life of people everywhere.	Environmentally friendly/ier products and solutions
yes (borderline)	FreshPoint, our specialty produce company, addresses customers' needs for fresh, unique, organic, and local produce items.	Environmentally friendly/ier products and solutions
yes (borderline)	WilLDAR consists of detecting methane leaks with an optical gas imaging camera and repairing those leaks within 30 days.	Environmentally friendly/ier products and solutions
yes (borderline)	These products include climate metrics, Climate Value-at-Risk (VAR), carbon portfolio reporting, low carbon, and climate change indexes as well as tools to identify clean-tech and environmentally oriented companies.	Environmentally friendly/ier products and solutions

Table 3: Environmental Claims with Rationale in Annotation Guidelines

Label	Sentence	Explanation
no (borderline)	We do this for 15 sustainable and impact strategies (equities, bonds and green bonds).	No positive impact or no link to better environmental performance
no (borderline)	We use the EcoAct ClimFIT (Climate Financial Institutions Tool) tool to measure the carbon emissions associated with the household and personal products sector.	No positive impact or no link to better environmental performance
no (borderline)	AUSEA is a miniaturized sensor, fitted onto a commercial drone, that can detect methane and carbon dioxide.	Product with potentially positive environmental impact, but impact is not stated hence no claim
no (borderline)	This will further accelerate Croda's positive impact by creating and delivering solutions to tackle some of the biggest challenges the world is facing.	Unclear whether this relates to environmental positive impacts, only implicit assumptions would make it an environmental claim.
no (unanimously)	Hence, the Scope 2 emission is included in the Scope 1 emission which has been reported in accordance with the ISO 14064-1 requirements as verified by qualified independent assessor.	Technical details, descriptions, and explanations
no (unanimously)	Emissions associated with processing activities are associated with the supply of these ingredients and are included in our Scope 3 supply chain emissions.	Technical details, descriptions, and explanations
no (unanimously)	Emissions are modelled based on sector averages including linear regression and country carbon emissions intensities for GDP.	Technical details, descriptions, and explanations
no (unanimously)	Wood products facilities also operate lumber drying kilns and other processes that can either use the steam from the boilers or, if direct fired, will commonly use natural gas.	Technical details, descriptions, and explanations
no (unanimously)	We use the EcoAct ClimFIT (Climate Financial Institutions Tool) tool to measure the carbon emissions associated with utilities.	Technical details, descriptions, and explanations
no (unanimously)	In the past we have conducted analysis of our portfolio impact on the climate, using scope 3 as a metric.	Technical details, descriptions, and explanations
no (unanimously)	For that, Danone needs organic fresh milk.	Sentence context would be required to understand whether it is a claim
no (unanimously)	UPM Biofuels is developing a new feedstock concept by growing Brassica Carinata as a sequential crop in South America.	Sentence context would be required to understand whether it is a claim
no (unanimously)	Our key sources of emissions are the running of our operations (electricity, business travel, etc), purchased goods and services (consultants, maintenance work, IT services, etc), and land leased to sheep and beef farming (to keep the grass low under our wind farms).	environmental risk exposure description but no commitment / claim to act on reducing the risk or improving impact
no (unanimously)	Extreme weather events and the impacts of transitioning to a low-carbon economy have the potential to disrupt business activities, damage property, and otherwise affect the value of assets, and affect our customers' ability to repay loans.	environmental risk exposure description but no commitment / claim to act on reducing the risk or improving impact
no (unanimously)	At the date of this report, the Group owns 34 mills (29 of which produce containerboard), 245 converting plants (most of which convert containerboard into corrugated boxes), 40 recovered fibre facilities and two wood procurement operations (which together provide raw material for our mills) and 34 other production facilities carrying on other related activities.	environmental risk exposure description but no commitment / claim to act on reducing the risk or improving impact

Table 4: Negative Examples with Rationale in Annotation Guidelines

Environmental Claims	Negative Examples
In support of Apple's commitment to reduce its carbon footprint by transitioning its entire supply chain to 100% renewable energy, we've transitioned our facilities in China to be powered through a series of renewable power purchase agreements.	So there's an annual cycle that, to some degree, dictates the pace of these enrollment campaigns.
We are looking at opportunities to expand our commitment to renewable diesel while continuing to optimize the efficiency of our fleet of traditional biodiesel plants.	And so when we get these biopsy data published, which we're aggressively working on, we think we will have sufficient information to begin to approach payers, including Medicare.
We plan to continue our low risk growth strategy by building our core business with rate base infrastructure, while maintaining the commitment to renewable energy initiatives and to reducing emissions.	And I guess first of all, I would say the thesis which we have at FERC here for precedent is no different than what takes place right now for the LDC companies, where the LDC companies pay for pipeline infrastructure that's developed by a pipeline operator.
We just completed $1 billion of capital projects to expand, upgrade and modernize and improve the environmental footprint of an important industry in Russia.	But as Jon points out, the thing that they really seem to be focused on is we claim a five-year life, and they want to make sure that that's a reasonable claim on our batteries for AED Plus.
And we also announced that BHGE is committed to reduce its carbon footprint by 50% by 2030, and also net 0 by 2050.	They're critical to reimbursement, meaning you just simply can't get revenue unless you've done things like enroll it, and you have to have accurate data to get providers enrolled.

Table 5: Environmental Claims and Negative Examples Predicted in Quarterly Earning Calls Answer Sections.

Minimum card	
Information	**Unit**
1. Is the resulting model publicly available?	Yes
2. How much time does the training of the final model take?	< 5 min
3. How much time did all experiments take (incl. hyperparameter search)?	60 hours
4. What was the energy consumption (GPU/CPU)?	0.3 kW
5. At which geo location were the computations performed?	*Switzerland*
Extended card	
6. What was the energy mix at the geo location?	89 gCO2eq/kWh
7. How much CO2eq was emitted to train the final model?	2.2 g
8. How much CO2eq was emitted for all experiments?	1.6 kg
9. What is the average CO2eq emission for the inference of one sample?	0.0067 mg
10. Which positive environmental impact can be expected from this work?	This work can help detect and evaluate environmental claims and thus have a positive impact on the environment in the future.
11. Comments	-

Table 6: Climate performance model card following [23]