

CLIMATEBERT: A Pretrained Language Model for Climate-Related Text

Nicolas Webersinke[◇] Mathias Kraus[◇] Julia Anna Bingler[†] Markus Leippold[‡]

[◇]FAU Erlangen-Nuremberg, Germany

[†]ETH Zurich, Switzerland

[‡]University of Zurich, Switzerland

nicolas.webersinke@fau.de, mathias.kraus@fau.de, binglerj@ethz.ch, markus.leippold@bf.uzh.ch

Abstract

Over the recent years, large pretrained language models (LM) have revolutionized the field of natural language processing (NLP). However, while pretraining on general language has been shown to work very well for common language, it has been observed that niche language poses problems. In particular, climate-related texts include specific language that common LMs can not represent accurately. We argue that this shortcoming of today’s LMs limits the applicability of modern NLP to the broad field of text processing of climate-related texts. As a remedy, we propose CLIMATEBERT, a transformer-based language model that is further pretrained on over 1.6 million paragraphs of climate-related texts, crawled from various sources such as common news, research articles, and climate reporting of companies. We find that CLIMATEBERT leads to a 46% improvement on a masked language model objective which, in turn, leads to lowering error rates by 3.57% to 35.71% for various climate-related downstream tasks like text classification, sentiment analysis, and fact-checking.

1 Introduction

Researchers working on climate change-related topics increasingly use natural language processing (NLP) to automatically extract relevant information from textual data. Many studies in this domain apply traditional NLP methods, such as bag-of-words approaches or simple extensions thereof (e.g., Gruning, 2011). However, such analyses face considerable limitations, since climate-related wording could vary substantially by source (Kim and Kang, 2018). Deep learning techniques that promise higher accuracy are gradually replacing these approaches (e.g., Köbel et al., 2020; Luccioni et al., 2020; Bingler et al., 2021; Callaghan et al., 2021; Wang et al., 2021;

Friederich et al., 2021). Indeed, it has been shown in related domains that deep learning in NLP allows for impressive results, outperforming traditional methods by large margins (Varini et al., 2020).

These deep learning-based approaches make use of language models (LMs), which are trained on large amounts of textual and unlabelled data. This training on unlabelled data is called *pretraining* and leads to the model learning representations of words and patterns of common language. One of the most prominent language models is called BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) with its successors ROBERTA (Liu et al., 2019), Transformer-XL (Dai et al., 2019) and ELECTRA (Clark et al., 2020). These models have been trained on huge amounts of text which was crawled from an unprecedented amount of online resources.

After the pretraining phase, most LMs are trained on additional tasks, the *downstream task*. For the downstream tasks, the LM builds on and benefits from the word representations and language patterns learned in the pretraining phase. The pre-training benefit is especially large on downstream tasks for which the collection of samples is difficult and, thus, the resulting training datasets are small (hundreds or few thousands of samples). Furthermore, it has been shown that a model that was pretrained on the downstream task-specific text exhibits better performance, compared to a model that has been pretrained solely on general text (Araci, 2019; Lee et al., 2020).

Hence, a straightforward extension to the standard combination of pretraining is the so-called domain-adaptive pretraining (Gururangan et al., 2020). This approach has recently been studied for various tasks and basically comes in the form of pretraining multiple times — in particular pretraining in the language domain of the downstream

task, i.e.,

pretraining (general domain)
+ **domain-adaptive**
pretraining (downstream domain)
+ training (downstream task).

To date, regardless of the increase in using NLP for climate change related research, a model with climate domain-adaptive pretraining has not been publicly available, yet. Research so far rather relied on models pretrained on general language, and fine-tuned on the downstream task. To fill this gap, our contribution is threefold. First, we introduce CLIMATEBERT, a state-of-the-art language model that is specifically pretrained on climate-related text corpora of various sources, namely news, corporate disclosures, and scientific articles. This language model is designed to support researchers of various disciplines in obtaining better performing NLP models for a manifold of downstream tasks in the climate change domain. Second, to illustrate the strength of CLIMATEBERT, we highlight the performance improvements using CLIMATEBERT on three standard climate-related NLP downstream tasks. Third, to further promote research at the intersection of climate change and NLP, we make the weights of all trained language models publicly available at climatebert.ai.

2 Background

As illustrated in Figure 1, our LM training approach for CLIMATEBERT comprises all three phases — using an LM pretrained on a general domain, the domain-adaptive pretraining on the climate domain, and the training phase on climate-related downstream tasks.

2.1 Pretraining on General Domain

As of 2018, pretraining became the quasi-standard for learning NLP models. First, a neural language model, often with millions of parameters, is trained on large unlabeled corpora in a semi-supervised fashion. By learning on multiple levels which words/word-sequences/sentences appear in the same context, an LM can represent a semantically similar text by similar vectors. Typical objectives for training LMs are the prediction of masked words or the prediction of a label indicating whether two sentences occurred consecutively in the corpora (Devlin et al., 2018).

In the earlier NLP pretraining days, LMs traditionally used neural networks (Collobert and Weston, 2008) or later Long-Short-Term-Memory (LSTM) networks to process text (Howard and Ruder, 2018). Today’s LMs mostly build on transformer models (e.g., Devlin et al., 2018; Dai et al., 2019; Liu et al., 2019). One of the latter is named ROBERTA (Liu et al., 2019) which was trained on 160GB of various English-language corpora - data from BOOKCORPUS (Zhu et al., 2015), WIKIPEDIA, a portion of the CCNEWS dataset (Nagel, 2016), OPENWEBTEXT corpus of web content extracted from URLs shared on Reddit (Gokaslan and Cohen, 2019), and a subset of CommonCrawl that is said to resemble the story-like style of WINOGRAD schemas (Trinh and Le, 2019). While these sources are valuable to build a model working on general language, it has been shown that domain-specific, niche language (such as climate-related text) poses a problem to current state-of-the-art language models (Araci, 2019).

2.2 Domain-Specific Pretraining

As a remedy to inferior performance of general language models when applied to niche topics, multiple language models have been proposed which build on the pretrained models but continue pretraining on their respective domains. FinBERT, LegalBERT, MedBERT are just a few language models that have been further pretrained on the finance, legal, or medical domain (Araci, 2019; Chalkidis et al., 2020; Rasmy et al., 2021). In general, this domain-adaptive pretraining yields more accurate models on downstream tasks (Gururangan et al., 2020).

It is still an open debate which sample strategy improves performance best for domain-adaptive pretraining. Various strategies can be applied to extract the text samples on which the LM is further pretrained. For example, while traditional pretraining uses all samples from the pretraining corpus, similar sample selection (SIM-SELECT) uses only a subset of the corpus, in which the samples are similar to the samples in the downstream task (Ruder and Plank, 2017). In contrast, diverse sample selection (DIV-SELECT) uses a subset of the corpus, which includes dissimilar samples compared to the downstream dataset (Ruder and Plank, 2017). Previous research has investigated the benefit of these approaches, yet no final conclusion about the efficiency has been obtained. Conse-

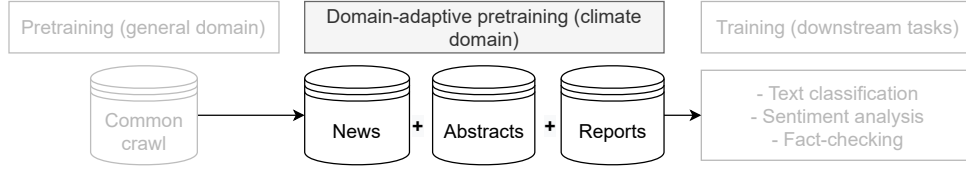


Figure 1: Sequence of training phases. Our main contribution is the continued pretraining of language models on the climate domain. In addition, we evaluate the obtained climate domain-specific language models on various downstream tasks.

quently, we compare these approaches in our experiments.

2.3 NLP on Climate-Related Text

In the past, climate-related textual analysis often used pre-defined dictionaries of presumably relevant words and then simply searched for these words within the documents. For example, [Cody et al. \(2015\)](#) use such an approach for climate-related tweets. Similarly, [Sautner et al. \(2020\)](#) use a keyword-based approach to capture firm-level climate change exposure. However, these methods do not account for context. The lack of context is a significant drawback, given the ambiguity of many climate-related words such as "environment," "sustainable," or "climate" itself ([Varini et al., 2020](#)).

Only recently, BERT has been used for NLP in climate-related text. The transformers-based BERT models are capable of accounting for the context of words and have outperformed traditional approaches by large margins across various climate-related datasets ([Kölbel et al., 2020](#); [Luccioni et al., 2020](#); [Varini et al., 2020](#); [Bingler et al., 2021](#); [Callaghan et al., 2021](#); [Wang et al., 2021](#); [Friederich et al., 2021](#)). However, this research has also shown that extracting climate-related information from textual sources is a challenge, as climate change is a complex, fast-moving, and often ambiguous topic with scarce resources for popular text-based AI tasks.

While context-based algorithms like BERT can detect a variety of complex and implicit topic patterns in addition to many trivial cases, there remains great potential for improvement in several directions. To our knowledge, none of the above cited work has examined the effects of domain-adaptive pretraining on their specific downstream tasks. Therefore, we investigate whether domain-adaptive pretraining will improve performance for climate-change-related downstream tasks such as text classification, sentiment analysis, and fact-

checking.

3 CLIMATEBERT

In the following, we describe our approach to train CLIMATEBERT. We first list the underlying data sources before describing the our sample selection techniques and, finally, the vocabulary augmentation we used for training the language model.

3.1 Text Corpus

Our goal was to collect a large corpus of text that included general and domain-specific climate-related language. We decided to include the following three sources: news articles, research abstracts, and corporate climate reports.¹ In total, we collected 1 662 206 paragraphs of climate-related text (see Table 1).

The NEWS dataset is mainly retrieved from Refinitiv Workspace and includes articles tagged with climate change topics such as climate politics, climate actions, and floods and droughts. In addition, we crawled climate-related news articles from the web.

The ABSTRACTS dataset includes abstracts of climate-related research articles crawled from the Web of Science, primarily published between 2000 and 2019.²

The REPORTS dataset comprises corporate climate and sustainability reports of more than 600 companies from the years 2015-2020 retrieved from Refinitiv Workspace and the respective company websites.

Given the nature of the datasets, we find a large heterogeneity between the paragraphs in terms of

¹We decided not to include full research articles because this language is likely too specific and does not represent general climate language. We also did not include Twitter data, as we assume that these texts are too noisy

²We are very thankful to Jan Minx and Max Callaghan from the Mercator Research Institute on Global Commons and Climate Change (MCC) Berlin for providing us with the data, which is a subset of the data they used in [Berrang-Ford et al. \(2021\)](#) and [Callaghan et al. \(2021\)](#).

| Dataset | Num. of paragraphs | Avg. num. of words | | |
|-----------|--------------------|--------------------|------|-----|
| | | Q1 | Mean | Q3 |
| News | 641,095 | 30 | 48 | 57 |
| Abstracts | 530,819 | 165 | 218 | 260 |
| Reports | 490,292 | 34 | 65 | 79 |
| Total | 1,662,206 | 36 | 107 | 168 |

Table 1: Corpus CORP used for pretraining CLIMATEBERT. Q1 and Q3 stand for the 0.25 and 0.75 quantiles, respectively.

number of words. Unsurprisingly, on average, the paragraphs with the least words come from the NEWS and the REPORTS datasets. In contrast, ABSTRACTS includes paragraphs with the most words. Table 1 lists these descriptives.

To estimate the benefit from domain-adaptive pretraining, we compare the similarity of our text corpus with the one used for pretraining ROBERTA. Following Gururangan et al. (2020), we consider the vocabulary overlap between both corpora. The resulting overlap of 57.05% highlights the dissimilarity between the two domains and the need to add specific vocabularies. Therefore, we expect to see considerable performance improvements of domain-adaptive pretraining.

3.2 Sample Selection

Prior work has shown that specific selections of the samples used for pretraining can foster the performance of the LM. In particular, incorporating information from the downstream task by selecting similar or diverse samples has been shown to yield favorable results compared to using all samples from the dataset. We follow both approaches and select samples that are similar or diverse to climate-text using our text classification task (see 5.1). We experiment with three different strategies from Ruder and Plank (2017) for the selection of samples from our corpus:

- In the most traditional sample selection strategy, FULL-SELECT, we use all paragraphs from CORP to train CLIMATEBERT_F.
- In SIM-SELECT, we select the 70% of samples from CORP, which are most similar to the samples of our text classification task. We use a Euclidean similarity metric for this sample selection strategy. We call this LM CLIMATEBERT_S.

- In DIV-SELECT, we select the 70% of samples from CORP, which are most diverse compared to the samples from our text classification task. We use the sum between the type-token-ratio and the Shannon-entropy for measuring diversity (Ruder and Plank, 2017). This LM is named CLIMATEBERT_D.

- In DIV-SELECT + SIM-SELECT, we use the same diversity and similarity metrics as before. We then compute a composite score by summing over their scaled values. We keep the 70% of the samples with the highest composite score to train CLIMATEBERT_{D+S}.

3.3 Vocabulary Augmentation

We extend the existing vocabulary of the original model to include domain-specific terminology. This allows CLIMATEBERT to explicitly learn representations of terminology that frequently occur in a climate-related text but not in the general domain. In particular, we add the 235 most common tokens as new tokens to the tokenizer, thereby extending the size of the vocabulary for our basis language model (DistilROBERTA) from 50 265 to 50 500.³

3.4 Model Selection

For all our experiments, we use DistilROBERTA, a distilled version of ROBERTA from Huggingface,⁴ as our starting point for training (Sanh et al., 2019). All our language models are trained with a masked language modeling objective (i.e., cross-entropy loss on predicting randomly masked tokens). We report all hyperparameters in Table 2. The large batch size of 2016 for training the LM is achieved using gradient accumulation.

3.5 Training on Downstream Task

After pretraining DistilROBERTA on CORP, we follow standard practice (Devlin et al., 2018) and pass the final layer [CLS] token representation to a task-specific feedforward layer for prediction. We report all hyperparameters of this feedforward layer in Table 2.

³We also experimented with language models that do not use vocabulary augmentation or add more tokens. However, overall we find improvements using this technique and, thus, apply it to all language models which we pretrain on the climate domain.

⁴huggingface.co/distilroberta-base

| | Downstream domain-adaptive pretraining | Downstream tasks training |
|--------------------------|--|---------------------------|
| Hyperparameter | Value | |
| Batch size | 2016 | 32 |
| Learning rate | 5e-4 | 5e-5 |
| Number of epochs | 12 | 1000 |
| Patience | — | 4 |
| Class weight | — | Balanced |
| Feedforward nonlinearity | — | tanh |
| Feedforward layer | — | 1 |
| Output neurons | — | Task dependent |
| Optimizer | Adam | |
| Adam epsilon | 1e-6 | |
| Adam beta weights | (0.9, 0.999) | |
| Learning rate scheduler | Warmup linear | |
| Weight decay | 0.01 | |

Table 2: Hyperparameters used for the downstream domain-adaptive pretraining and the downstream tasks training of CLIMATEBERT.

4 Performance Analysis of Language Model

Table 3 lists the results after pretraining DistilROBERTA on CORP with various sample selection strategies. For evaluation, we split CORP randomly into 80% training data and 20% validation data. The reported loss is the cross-entropy loss on predicting randomly masked tokens from the validation data. We find that CLIMATEBERT_F leads to the lowest validation loss. This performance is followed by the other CLIMATEBERT LMs, which all show similar results. Overall, we find that our domain-adaptive pretraining decreases the cross-entropy loss by 46–47% compared to the basis DistilROBERTA, cutting the loss almost in half.

| Model | Val. loss |
|----------------------------|-----------|
| DistilROBERTA | 2.238 |
| CLIMATEBERT _F | 1.188 |
| CLIMATEBERT _S | 1.205 |
| CLIMATEBERT _D | 1.204 |
| CLIMATEBERT _{D+S} | 1.203 |

Table 3: Loss of our language models on a validation set from our text corpus CORP.

5 Performance Analysis for Climate-Related Downstream Tasks

For our experiments, we used the following downstream tasks: text classification, sentiment analysis, and fact-checking. Table 4 lists basic statistics about the downstream tasks. We repeated the training and evaluation phase 60 times for each experiment, each time with a random 90% set of

samples for training and the remaining 10% set for validation.

| Downstream task | Num. of samples | Labels | Label distribution |
|---------------------|-----------------|--------------------------|--------------------|
| Text classification | 1220 | climate-related: yes/no | 1000/220 |
| Sentiment analysis | 1000 | opportunity/neutral/risk | 250/408/342 |
| Fact-checking | 2745 | claim: support/refute | 1943/802 |

Table 4: Overview of our downstream tasks used for evaluating CLIMATEBERT.

5.1 Text Classification

For our text classification experiment, we use a dataset consisting of hand-selected paragraphs from companies’ annual reports or sustainability reports. All paragraphs were annotated as *yes* (climate-related) or *no* (not climate-related) by at least four experts from the field using the software prodigy.⁵ See Appendix B.1 for our annotation guidelines. In case of a close verdict or a tie between the annotators, the authors of this paper discussed the paragraph in depth before reaching an agreement.

In the following, Table 5 reports the results of the language models when trained on our text classification task, i.e., whether the text is climate-related or not. Overall, we find that all CLIMATEBERT LMs outperform a non-pre-trained DistilROBERTA across both metrics for the text classification task. Most notably, domain-adaptive pretraining with similar samples to our downstream tasks (CLIMATEBERT_S) leads to improvements of 32.64% in terms of cross-entropy loss and a reduction in the error rate of the F1 score by 35.71%.

| Model | Text classification | |
|----------------------------|------------------------|------------------------|
| | Loss | F1 |
| DistilROBERTA | 0.242 _{0.171} | 0.986 _{0.010} |
| CLIMATEBERT _F | 0.191 _{0.136} | 0.989 _{0.010} |
| CLIMATEBERT _S | 0.163 _{0.132} | 0.991 _{0.008} |
| CLIMATEBERT _D | 0.197 _{0.153} | 0.988 _{0.009} |
| CLIMATEBERT _{D+S} | 0.217 _{0.153} | 0.988 _{0.009} |

Table 5: Results on our text classification task. Reported are the average cross-entropy loss and the average weighted F1 score on the validation sets across 60 evaluation runs. Value subscripts report the standard deviations.

⁵prodigy

5.2 Sentiment Analysis

Our next task studies the sentiment behind the climate-related paragraphs, using the same dataset as in the previous section. In our context, we use the term ‘sentiment’ to distinguish whether an entity reports on climate-related developments as negative *risk*, as positive *opportunity*, or as *neutral*.

Therefore, we created a second labeled dataset on climate-related sentiment, for which we asked the annotators to label the paragraphs by one of the three categories — *risk*, *neutral*, or *opportunity*. See Appendix B.2 for our annotation guidelines. Similarly, as before, in case of a close verdict or a tie between the annotators, the authors of this paper discussed the paragraph in depth before reaching an agreement.

Table 6 shows the performance of our models in sentiment prediction. Again, all CLIMATEBERT LMs outperform the DistilROBERTA baseline model in terms of F1 score and average cross-entropy loss. The largest improvements can be observed with CLIMATEBERT_F, which amount to a 7.33% lower cross-entropy loss and a 7.42% lower error rate in terms of average F1 score compared to the DistilROBERTA baseline LM.

| Model | Sentiment analysis | |
|----------------------------|------------------------|------------------------|
| | Loss | F1 |
| DistilROBERTA | 0.150 _{0.069} | 0.825 _{0.046} |
| CLIMATEBERT _F | 0.139 _{0.042} | 0.838 _{0.036} |
| CLIMATEBERT _S | 0.140 _{0.057} | 0.836 _{0.033} |
| CLIMATEBERT _D | 0.138 _{0.043} | 0.835 _{0.040} |
| CLIMATEBERT _{D+S} | 0.139 _{0.043} | 0.834 _{0.036} |

Table 6: Results on our sentiment analysis task in terms of average validation loss and average weighted F1 score across 60 evaluation runs. Subscripts report the standard deviations.

5.3 Fact-Checking

We now turn to the fact-checking downstream task. We apply our model to a dataset that was proposed by Diggelmann et al. (2020) and comprises 1.5k sentences that make a claim about climate-related topics. This CLIMATE-FEVER dataset is to the best of our knowledge to date the only dataset that focuses on climate change fact-checking. CLIMATE-FEVER adapts the methodology of FEVER, the largest dataset of artificially designed claims, to real-life claims on climate

change collected online. The authors of CLIMATE-FEVER find that the surprising, subtle complexity of modeling real-world climate-related claims provides a valuable challenge for general natural language understanding. Working with this dataset, Wang et al. (2021) recently introduced a novel semi-supervised training method to achieve a state-of-the-art (SotA) F1 score of 0.7182 on the fact-checking dataset CLIMATE-FEVER.

| | |
|------------------------------|---|
| Claim: | 97% consensus on human-caused global warming has been disproven. |
| Evidence: REFUTE | In a 2019 CBS poll, 64% of the US population said that climate change is a ""crisis"" or a ""serious problem"", with 44% saying human activity was a significant contributor. |
| Claim: | The melting Greenland ice sheet is already a major contributor to rising sea level and if it was eventually lost entirely, the oceans would rise by six metres around the world, flooding many of the world’s largest cities. |
| Evidence : SUPPORT | The Greenland ice sheet occupies about 82% of the surface of Greenland, and if melted would cause sea levels to rise by 7.2 metres. |

Table 7: Examples taken from CLIMATE-FEVER.

Each claim in CLIMATE-FEVER is supported or refuted by evidence sentences (see Table 8), and an evidence sentence can also be classified as giving not enough information.⁶ The objective of the model is to classify an evidence sentence to *support* or *refute* a claim. To feed this combination of claim and evidence into the model, we concatenate the claims with the related evidence sentences, with a [SEP] token separating them.

Table 8 lists the results of our experiments on the CLIMATE-FEVER dataset. In line with our previous experiments, we find similar or better results for all CLIMATEBERT LMs across all metrics. Our CLIMATEBERT_{D+S} LM achieves similar cross-entropy loss compared to the basis DistilROBERTA model, yet pushes the average F1 score from 0.748 to 0.757, which outperforms Wang et al. (2021)’s previous SotA F1 score of 0.7182, and is hence, to the best of our knowledge, the new SotA on this dataset.

6 Carbon Footprint

Training deep neural networks in general and large language models in particular, has a significant carbon footprint already today. If the LM research

⁶As in Wang et al. (2021), and for comparison with their results, we filter out all evidence sentences with the label *NOT_ENOUGH_INFO* in the CLIMATE-FEVER dataset.

| Model | Fact-checking | |
|----------------------------|------------------------|------------------------|
| | Loss | F1 |
| DistilROBERTA | 0.135 _{0.017} | 0.748 _{0.036} |
| CLIMATEBERT _F | 0.134 _{0.020} | 0.755 _{0.037} |
| CLIMATEBERT _S | 0.133 _{0.017} | 0.753 _{0.042} |
| CLIMATEBERT _D | 0.135 _{0.016} | 0.752 _{0.042} |
| CLIMATEBERT _{D+S} | 0.135 _{0.018} | 0.757 _{0.044} |

Table 8: Results on our fact-checking task on CLIMATE-FEVER in terms of average validation loss and average weighted F1 score across 60 evaluation runs. Subscripts report the standard deviations.

trends continue, this detrimental climate impact will increase considerably. We acknowledge that our work is part of this trend. In total, training CLIMATEBERT caused 94.75 kg CO₂ emissions.⁷ Nevertheless, we decided to carry out this project, as we see the high potential of NLP to support action against climate change. Given our awareness of the carbon footprint of our research, we address this sensitive topic as follows:

1. We specifically decided to focus on DistilROBERTA, which is a considerably smaller model in terms of number of parameters compared to the non-distilled version and, thus, requires less energy to train. Moreover, we do not crawl huge amounts of data without considering the quality. This way, we try to take into account the issues mentioned by Bender et al. (2021).
2. Hyperparameter tuning yields considerably higher CO₂ emissions in the training stage due to tens or hundreds of different training runs.⁸ We therefore refrain from exhaustive hyperparameter tuning. Rather, we build on previous findings. We systematically experimented with a few hyperparameter combinations and found that the hyperparameters proposed by Gururangan et al. (2020) lead to the best results.

⁷We use two energy efficient NVIDIA RTX A5000 GPUs: 0.7 kW (power consumption of GPU server) x 288 hours (combined training time of all experiments) x 470 gCO₂e/kWh (emission factor in Germany in 2018 according to umweltbundesamt.de/publikationen/entwicklung-der-spezifischen-kohlendioxid-7) = 94752 gCO₂e.

⁸Note that our multiple training runs on the downstream task are not causing long training times as the downstream datasets are very small compared to the dataset used for training the language model.

3. We would have liked to train and run our model on servers powered by renewable energy. This first best option was unfortunately not available. In order to speed up the energy system transformation required to achieve the global climate targets, we contribute our part by donating Euro 100 to atmosfair.⁹ We explicitly refrain from calling this donation a CO₂ compensation, and we refrain from a solution that is based on afforestation. See Appendix A for a more detailed statement on the matter.

7 Conclusion

We propose CLIMATEBERT, the first language model that was pretrained on a large scale dataset of over 1.6 million climate-related paragraphs. We study various selection strategies to find samples from our corpus which are most helpful for later tasks. Our experiments reveal that our domain-adaptive pretraining leads to considerably lower masked language modeling loss on our climate corpus. We further find that this improvement is also reflected in predictive performance across three essential downstream climate-related NLP tasks: text classification, the analysis of risk and opportunity statements by corporations, and fact-checking climate-related claims.

References

- Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623.
- Lea Berrang-Ford, Anne J Sietsma, Max Callaghan, Jan C Minx, Pauline FD Scheelbeek, Neal R Haddaway, Andy Haines, and

⁹atmosfair was founded in 2005 and is supported by the German Federal Environment Agency supports the initiative. atmosfair offsets carbon dioxide in more than 20 locations: from efficient cookstoves in Nigeria, Ethiopia and India to biogas plants in Nepal and Thailand to solar energy in Senegal and Brazil and renewable energies in Tansania and Indonesia. See <https://www.atmosfair.de/en/of-fset/fix/>.

- Alan D Dangour. 2021. Systematic mapping of global research on climate and health: a machine learning review. *The Lancet Planetary Health*, 5(8):e514–e525.
- Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2021. Cheap talk and cherry-picking: What CLIMATEBERT has to say on corporate climate risk disclosures. *Available at SSRN 3796152*.
- Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J Brecha, et al. 2021. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change*, pages 1–7.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The mup-pets straight out of law school. *arXiv preprint arXiv:2010.02559*.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELEC-TRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Emily M. Cody, Andrew J. Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M. Danforth. 2015. Climate change sentiment on twitter: An unsolicited public opinion poll. *PLOS ONE*, 10(8):e0136092.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.
- David Friederich, Lynn H Kaack, Alexandra Luccioni, and Bjarne Steffen. 2021. Automated identification of climate risk disclosures in annual corporate reports. *arXiv preprint arXiv:2108.01415*.
- Aaron Gokaslan and Vanya Cohen. 2019. [Open-WebText Corpus](#).
- Michael Grüning. 2011. Artificial intelligence measurement of disclosure (aimd). *European Accounting Review*, 20(3):485–519.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 328–339.
- Do-Yeon Kim and Sung-Won Kang. 2018. Analysis of Recognition of Climate Changes using Word2Vec. *International Journal of Pure and Applied Mathematics*, 120(6):5793–5807.
- Julian F Kölbl, Markus Leippold, Jordy Rillaerts, and Qian Wang. 2020. Ask BERT: How regulatory disclosure of transition and physical climate risks affects the cds term structure. *Available at SSRN 3616324*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

- 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. *arXiv preprint arXiv:1907.11692*.
- Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. Analyzing sustainability reports using natural language processing. *arXiv preprint arXiv:2011.08073*.
- Sebastian Nagel. 2016. [CC-NEWS](#).
- Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.
- Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Zacharias Sautner, Laurence van Lent, Grigory Vilkov, and Ruisheng Zhang. 2020. Firm-level climate change exposure. *TRR 266 Accounting for Transparency Working Paper Series No. 33*.
- Trieu H. Trinh and Quoc V. Le. 2019. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.
- Francesco S Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. In *Tackling Climate Change with Machine Learning (Climate Change AI) workshop at NeurIPS*.
- Gengyu Wang, Lawrence Chillrud, and Kathleen McKeown. 2021. Evidence based automatic fact-checking for climate change misinformation. *International Workshop on Social Sensing on The International AAAI Conference on Web and Social Media*.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Appendix

A CLIMATEBERT's Climate Protection Contribution

We donate Euro 100 to atmosfair. We donated many times more than we directly emitted, as we are aware of the controversial and often misleading debate about carbon offsetting strategies and the amount of carbon pricing that would be required for the climate targets. Therefore, we apply a precautionary principle in the sense that if we err, we err in favor of the planet.



Figure 2: Certificate for donation.

Technically, our donation offsets 4,350 kg CO₂ via additional mitigation activities in the supported atmosfair projects. However, we would rather like to consider our donation in the upper range of the price for 1 tCO₂ emissions that would be required according to the report of the High Level Commission on Carbon Prices, where Economic Nobel Prize laureate Joseph E. Stiglitz together with Nicolas Stern and other researchers assembled the state of the art on the magnitude of carbon pricing that would be required to achieve the climate targets. In their final report, the commission

states that “the explicit carbon-price level consistent with achieving the Paris temperature target is at least USD 40–80/tCO₂ by 2020 and USD 50–100/tCO₂ by 2030, provided a supportive policy environment is in place.”¹⁰ We chose an upper bound, given that this lies naturally within the principle of “Common But Differentiated Responsibility and Respective Capabilities” (CBDR-RC) of the Paris Agreement, and given that this range might still be a relatively conservative estimate. Most recently, the IEA estimated in its World Energy Outlook 2021 that to achieve the 1.5 degrees target in their net zero emissions (NZE) scenario, a carbon price of USD 250/tCO₂ by 2050 in advanced economies would be required - together with further accompanying policy measures.¹¹

In addition, we refrain from calling our donation a compensation, and do not support afforestation projects in this context, for three main reasons rooted in climate science and economics. First, the IPCC and various international organisations like the IEA state that mitigation activities are key. Compensation activities will be required for hard-to-abate-sectors, once all other technological solutions have been implemented, and where mitigation is not (yet) feasible. The energy sector is explicitly not part of these hard-to-abate-sectors.¹²

Second, afforestation measures can be accompanied by land use conflicts, might be prone to carbon releases in the case of wildfires, and do not avoid additional deforestation activities elsewhere. Furthermore, afforestation activities are usually not considered as *additional* carbon sinks at global scale, since afforestation is required to store parts of the carbon released in the past, and deals with the carbon that is part of the fast natural carbon cycle.¹³

Third, economic dynamic efficiency requires investments in decarbonisation technologies in order to keep the climate targets within reach. Compensation activities, especially in the afforestation

¹⁰See <https://www.carbonpricingleadership.org/report-of-the-highlevel-commission-on-carbon-prices>

¹¹See <https://www.iea.org/reports/world-energy-outlook-2021>.

¹²See for example <https://www.ipcc.ch/sr15/> and <https://www.iea.org/reports/world-energy-outlook-2021>.

¹³See, e.g., <https://www.climatechangenews.com/2020/12/11/10-myths-net-zero-targets-carbon-offsetting-busted/>.

area, delay the required investments. This delay might exacerbate the likelihood of crossing climate tipping points, and/or yields to a disorderly transition to a decarbonised economy¹⁴.

B Annotation Guidelines

For our annotation procedure, we implemented the following general rules. The annotators had to label climate-relevant paragraphs. If the paragraph was climate-relevant, then they had to attach a sentiment to a paragraph. Annotators were asked to apply common sense, e.g., when a given paragraph might not provide all the context, but the context might seem obvious. Moreover, annotators were informed that each annotation should be a 0-1 decision. Hence, if an annotator was 70% certain, then this was rounded up to 100%. We asked, on average, five researchers to annotate the same tasks to obtain some measure of dispersion. In case of a close verdict or a tie between the annotators, the authors of this paper discussed the paragraph in depth before reaching an agreement.

B.1 Text classification

The first task was to label climate-relevant paragraphs. The labels are *Yes* or *No*. As a general rule, we determined that just discussing nature/environment can be sufficient, and mentioning clean energy, emissions, fossil fuels, etc., can also be sufficient. It is a *Yes*, if the paragraph includes some wording on a climate-change or environment related topic (including transition and litigation risks, i.e., emission mitigation measures, energy consumption and energy sources etc.; and physical risks, i.e., increase in risk of floods, coastal area exposure, storms etc.). It is a *No*, if the paragraph is not related to climate policy, climate change or an environmental topic at all. For some examples, see Table 9.

B.2 Sentiment Analysis

For the sentiment analysis, annotators had to provide labels as to whether a (climate-change related) paragraph talks about a *Risk* or threat that negatively impacts an entity of interest, i.e. a company (negative sentiment), or whether an entity is

¹⁴See for example <https://op.europa.eu/en/publication-detail/-/publication/0dcfc3f3-3138-11e6-b497-01aa75ed71a1/language-en> and again <https://www.iea.org/reports/world-energy-outlook-2021> and <https://www.ipcc.ch/sr15/>

| Label | Examples |
|-------|---|
| Yes | Sustainability: The Group is subject to stringent and evolving laws, regulations, standards and best practices in the area of sustainability (comprising corporate governance, environmental management and climate change (specifically capping of emissions), health and safety management and social performance) which may give rise to increased ongoing remediation and/or other compliance costs and may adversely affect the Group's business, results of operations, financial condition and/or prospects. |
| Yes | Scope 3: Optional scope that includes indirect emissions associated with the goods and services supply chain produced outside the organization. Included are emissions from the transport of products from our logistics centres to stores (downstream) performed by external logistics operators (air, land and sea transport) as well as the emissions associated with electricity consumption in franchise stores. |
| No | Risk and risk management Operational risk and compliance risk Operational risk is the risk of loss resulting from inadequate or failed internal processes, people and systems, or from external events including legal risk but excluding strategic and reputation risk. It also includes, among other things, technology risk, model risk and outsourcing risk. |

Table 9: Examples for the annotation task climate (Yes/No).

referring to some *Opportunity* arising due to climate change (positive sentiment). The paragraph can also make just a *Neutral* statement.

To be more precise, we consider a paragraph relating to risk, if the paragraph mainly talks about 1) business downside risks, potential losses and adverse developments detrimental to the entity 2) and/or about negative impact of an entity's activities on the society/environment 3) and/or associates specific negative adjectives to the anticipated, past or present developments and topics covered.

We consider a paragraph relating to opportunities, if the paragraph mainly talks about 1) business opportunities arising from mitigating climate change, from adapting to climate change etc. which might be beneficial for a specific entity 2) and/or about positive impact of an entity's activities on the society/environment 3) and/or associates specific positive adjectives to the anticipated, past or present developments and topics covered.

Lastly, we consider a paragraph as neutral if it

| Label | Examples |
|-------------|--|
| Opportunity | Grid & Infrastructure and Retail – today represent the energy world of tomorrow. We rank among Europe’s market leaders in the grid and retail business and have leading positions in renewables. We intend to spend a total of between Euro 6.5 billion and Euro 7.0 billion in capital throughout the Group from 2017 to 2019. |
| Opportunity | We want to contribute to the transition to a circular economy. The linear economy is not sustainable. We discard a great deal (waste and therefore raw materials, experience, social capital and knowledge) and are squandering value as a result. This is not tenable from an economic and ecological perspective. As investor we can ‘direct’ companies and with our network, our scale and our influence we can help the movement towards a circular future (creating a sustainable society) further along. |
| Neutral | A similar approach could be used for allocating emissions in the fossil fuel electricity supply chain between coal miners, transporters and generators. We don’t invest in fossil fuel companies, but those investors who do should account properly for their role in the production of dangerous emissions from burning fossil fuels. |
| Neutral | Omissions: Emissions associated with joint ventures and investments are not included in the emissions disclosure as they fall outside the scope of our operational boundary. We do not have any emissions associated with heat, steam or cooling. We are not aware of any other material sources of omissions from our emissions reporting. |
| Risk | We estimated that between 36.5 and 52.9 per cent of loans granted to our clients are exposed to transition risks. If the regulator decides to pass ambitious laws to accelerate the transition towards a low-carbon economy, carbon-intensive companies would incur in higher costs, which may prevent them from repaying their debt. In turn, this would weaken our bank’s balance sheets. . |
| Risk | American National Insurance Company recognizes that increased claims activity resulting from catastrophic events, whether natural or man-made, may result in significant losses, and that climate change may also affect the affordability and availability of property and casualty insurance and the pricing for such products. |

Table 10: Examples for the annotation task sentiment (Opportunity/Neutral/Risk).

mainly states facts and developments 1) without putting them into positive or negative perspective for a specific entity and/or the society and/or the environment, 2) and/or does not associate specific positive or negative adjectives to the anticipated, past or present facts stated and topics covered. For some examples, see Table 10.