Deep Climate Change: A Dataset and Adaptive domain pre-trained Language Models for Climate Change Related Tasks

Saeid A. Vaghefi
University of Zurich
Switzerland
saeid.vaghefi@geo.uzh.ch

Veruska Muccione
University of Zurich
Switzerland
veruska.muccione@geo.uzh.ch

Christian Huggel
University of Zurich
Switzerland
christian.huggel@geo.uzh.ch

Hamed Khashehchi
2w2e GmbH, Switzerland
Switzerland
h.khashehchi@2w2e.com

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

Abstract

The quantity and quality of literature around climate change (CC) and its impacts are increasing yearly. Yet, this field has received limited attention in the Natural Language Processing (NLP) community. With the help of large Language Models (LMs) and transfer learning, NLP can support policymakers, researchers, and climate activists in making sense of large-scale and complex CC-related texts. CC-related texts include specific language that general language models cannot represent accurately. Therefore we collected a climate change corpus consisting of over 360 thousand abstracts of top climate scientists' articles from trustable sources covering large temporal and spatial scales. Comparison of the performance of GPT-2 LM and our 'climateGPT-2 models', fine-tuned on the CC-related corpus, on claim generation (text generation) and fact-checking, downstream tasks show the better performance of the climateGPT-2 models compared to the GPT-2. The climateGPT-2 models decrease the validation loss to 1.08 for claim generation from 43.4 obtained by GPT-2. We found that climateGPT-2 models improved the masked language model objective for the fact-checking task by increasing the F1 score from 0.67 to 0.72.

1 Introduction

Two of the most established Language Models (LMs) are GPT (Generative Pre-trained Transformer) Brown et al. (2020); Radford et al. (2019) and BERT (Bidirectional Encoder Representations from

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

Transformers) Devlin et al. (2018). They are the same in that they are both based on the transformer architecture, but they are fundamentally different in that BERT has just the encoder blocks from the transformer, while GPT-2 has just the decoder blocks from the transformer Devlin et al. (2018); Radford et al. (2019).Following the general pre-training phase, the LMs are further fine-tuned on downstream tasks. However, researchers have shown the better performance of LMs pre-trained on general and target domains (domain adaptive pre-training) prior to the fine-tuning step compared to LMs pre-trained only on general texts before fine-tuning Konlea and Jannidisa (2020); Lee et al. (2020); Bambroo and Awasthi (2021). While LMs have been fine-tuned to achieve state-of-the-art results on a large number of tasks, the idea of further pre-training of the LMs to incorporate more domain knowledge has been explored less. Moreover, BERT and its successors have been used more for domain adaptive pre-training compared to GPT-based LMs. In climate science, one of the major contributions was climateBert Webersinke et al. (2021) which DistilBERT model Sanh et al. (2019) was further pre-trained and then fine-tuned on downstream tasks, i.e., classification and fact-checking. To the knowledge of the authors of this paper, there is not a publicly available transformer-based LM with climate domain-adaptive pre-training from the GPT family. The contributions of our paper are as follows: 1) 'climateGPT-2 models' are the first domain-specific GPT-based models pre-trained on climate change corpora for 3 days on two NVIDIA Gforce 2800 GPUs. 2) We show that pre-training GPT-2 on climate change corpus improves its performance on two downstream tasks: text generation (claim generation) and fact-checking. 3) We make our pre-processed datasets, the pre-trained weights of climateGPT-2, and the source code for fine-tuning climateGPT-2 models publicly available.

2 Related Work

2.1 Pretraining on General versus Specific Domain

Transfer learning and pre-trained language models in NLP facilitated language understanding and generation. In this context, LMs are functions that map text to text. Using a large-scale corpus for unsupervised pre-training models is a norm for LMs, making them ready for zero-and few-shot learning Zhang et al. (2022). The latest breakthroughs in LMs during the last few years are: BERT Devlin et al. (2019), GPT-2 Radford et al. (2019), Xlnet Yang et al. (2019), RoBERTa Liu et al. (2019), ALBERT Lan et al. (2019), DistilBERT Sanh et al. (2019), T5 Raffel et al. (2020), GPT-3 Brown et al. (2020), DeBERTa He et al. (2020), PaLM Chowdhery et al. (2022), OPT Zhang et al. (2022). In this work, we focus on GPT-2; therefore, it should be noted that the main difference between the GPT family models is their size. The original transformer model had around 110 million parameters. GPT-1 adopted the size and GPT-2 has 1.5 billion, and GPT-3 has 175 billion parameters.

There is a consensus that further training of a pre-trained LM could improve the performance of the LM on the downstream tasks. When the target sector is a technical domain such as legal, finance, medical, or climate science, the frequency of the common words is way different from the general domain, which makes the use of available LMs more limited Bambroo and Awasthi (2021); Lin et al. (2021). BioBERT Lee et al. (2020) for biomedical tasks and SciBERT Beltagy et al. (2019) for science domains are just two examples showing the effectiveness of LMs pre-training for downstream tasks in specific domains.

2.2 NLP on Climate Change Related Text

From financial climate disclosure analyses Bingler et al. (2022); Friederich et al. (2021); Luccioni and Palacios (2019), topic modeling at the intersection of climate and health Berrang-Ford et al. (2021); Callaghan et al. (2021), to climate claims fact-checking Webersinke et al. (2021), NLP techniques

have been widely used in the literature. In general, the BERT has been used more than GPT models. For example, original BERT was used to identify and classify studies on observed climate impacts and produced a comprehensive machine-learning-assisted evidence map Callaghan et al. (2021). Also, a domain-adaptive version of DistillBERT was implemented to assess companies' climate-risk disclosures along the primary climate-related Financial Disclosures (TCFD) categories Bingler et al. (2022).

3 Approach (ClimateGPT-2 models)

In this article, we introduce 'climateGPT-2 models', which are domain adaptive pre-trained LMs for climate change related topics.To analyze the effectiveness of our approach in climate change text mining, climateGPT-2 models are fine-tuned and evaluated on two text mining tasks: 1) text (claim) generation and 2) fact-checking. First, we initialize climateGPT-2 with weights from GPT-2, which was pre-trained on general domain corpora (the number of parameters in GPT-2 LM families varies and reaches 1.5 billion in GPT-2 XL). Second, climateGPT-2 models are trained on climate change domain corpora (over 360 thousand abstracts of articles from leading climate scientists). Lastly, we evaluate the climateGPT-2 models for two downstream tasks in climate change domain.

3.1 Climate Change corpus for domain adaptive pre-training

Climate change domain texts contain many domain-specific nouns (e.g., RCPs, SSPs, CO2) and terms (e.g., mitigation, adaptation), which are understood mostly by climate researchers. As a result, NLP models designed for general purpose language understanding often obtain poor performance in climate change text mining tasks. To overcome this challenge, we decided to collect a large corpus of text from the publication of well-known scientists in the climate field. Therefore we used the 1000 hot scientists list of climate change published by the Reuters press. Details of the corpus are presented in Table 1.

Table 1: Climate performance model card following

Corpus	Num of paragraphs	Domain	Reference
climateGPT-2	360,233	Climate Change	https://www.reuters.com/investigates/special-report/climate-change-scientists-list

Details of the ranking procedure of Reuters have been described in the link provided in Table 1.

3.2 Fact-checking dataset

We used the Climate Fever dataset for the fact-checking task Diggelmann et al. (2020). Climate Fever consists of roughly 1,500 claims in the climate domain. Annotators have classified claims in the climate fever dataset as supported, refuted, or not enough by evidence sentences.

3.3 Training of climateGPT-2 on 'text (claim) generation' task

One big challenge with text generation is the limited control in the direction it goes when sentences are added. To overcome this issue, we propose using a title and a list of keywords so the models can better recognize the direction. Therefore we will check if climateGPT-2 models are able to generate meaningful climate-related texts given a title with and without a list of keywords. In the standard text generation, the next token is predicted based on the text it has seen. Therefore, the labels are just the

shifted encoded tokenized input. As we want more control over the climateGPT-2, we give the model a title and a list of keywords in the customized dataset class to improve the next predicted token.

3.4 Training of climateGPT-2 on 'fact-checking' task

For fact-checking task, we use the approach suggested by Webersinke et al., (2021). The combination of concatenated claim and related evidence sentences and [SEP] token separator is fed to the model. The goal is to correctly classify if an evidence sentence is supported or refuted by a claim.

4 Results

4.1 Text Generation (Climate Change claims)

Table 2 summarizes the results of our experiments for the claim generation task. We see that climateGPT-2 models have improved the performance significantly (lowering the validation loss). Table 4 in the Apendix, summarizes Inputs/Output samples of climateGPT-2 models for text generation. In the generated texts, we could realize that individual sentences are semantically coherent, and the first three to four sentences are related to the title and/or keywords.

Table 2: Loss of GPT-2 vs climateGPT-s models on the CC-related corpus

Model	Val Loss
GPT-2	43.45
climateGPT-2 (cGPT-2)	1.08
climateGPT-2 (ckeyGPT-2)	1.56

4.2 Fact-checking

Table 3 summarizes the results of our experiments on the CLIMATE-FEVER dataset. The F1 score results show that climateGPT-2 models outperform GPT-2 by lowering F1 from 1.17 to 0.83.

Table 3: Results, average Validation Loss and average weighted F1 score on the fact-checking task on CLIMATE-FEVER dataset

Model	Val Loss	F1
GPT-2	1.17	0.67
climateGPT-2	0.83	0.72

By comparing our results with the fact-checking results performed on climate fever data with climateBERT, we realized that our results align with that study. However, our in-house corpus for adaptive pre-training is different from the climateBERT.

5 Conclusion

We developed climateGPT-2, the first language model from GPT families that was pre-trained on a large dataset of 360,233 climate-related scientific abstracts derived from articles of 1000 hot scientists' list of Reuters. We found that our language model lowers the loss on our climate-related corpus. We checked this model across two different climate change NLP downstream tasks. We showed that pre-training GPT-2 on climate change corpus largely improved its performance. climateGPT-2 obtains higher F1 scores in climate change fact-checking (0.72) compared to GPT-2 (0.67) and a lower validation loss (1.08) in climate change text generation task compared to the original GPT-2 (43.45).

References

Purbid Bambroo and Aditi Awasthi. 2021. https://doi.org/10.1109/ICAECT49130.2021.9392558 LegalDB: Long DistilBERT for Legal Document Classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. https://doi.org/10.18653/v1/D19-1371 SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Lea Berrang-Ford, A R Siders, Alexandra Lesnikowski, Alexandra Paige Fischer, Max W Callaghan, Neal R Haddaway, Katharine J Mach, Malcolm Araos, Mohammad Aminur Rahman Shah, Mia Wannewitz, Deepal Doshi, Timo Leiter, Custodio Matavel, Justice Issah Musah-Surugu, Gabrielle Wong-Parodi, Philip Antwi-Agyei, Idowu Ajibade, Neha Chauhan, William Kakenmaster, Caitlin Grady, Vasiliki I Chalastani, Kripa Jagannathan, Eranga K Galappaththi, Asha Sitati, Giulia Scarpa, Edmond Totin, Katy Davis, Nikita Charles Hamilton, Christine J Kirchhoff, Praveen Kumar, Brian Pentz, Nicholas P Simpson, Emily Theokritoff, Delphine Deryng, Diana Reckien, Carol Zavaleta-Cortijo, Nicola Ulibarri, Alcade C Segnon, Vhalinavho Khavhagali, Yuanyuan Shang, Luckson Zvobgo, Zinta Zommers, Jiren Xu, Portia Adade Williams, Ivan Villaverde Canosa, Nicole van Maanen, Bianca van Bavel, Maarten van Aalst, Lynée L Turek-Hankins, Hasti Trivedi, Christopher H Trisos, Adelle Thomas, Shinny Thakur, Sienna Templeman, Lindsay C Stringer, Garry Sotnik, Kathryn Dana Sjostrom, Chandni Singh, Mariella Z Siña, Roopam Shukla, Jordi Sardans, Eunice A Salubi, Lolita Shaila Safaee Chalkasra, Raquel Ruiz-Díaz, Carys Richards, Pratik Pokharel, Jan Petzold, Josep Penuelas, Julia Pelaez Avila, Julia B Pazmino Murillo, Souha Ouni, Jennifer Niemann, Miriam Nielsen, Mark New, Patricia Nayna Schwerdtle, Gabriela Nagle Alverio, Cristina A Mullin, Joshua Mullenite, Anuszka Mosurska, Mike D Morecroft, Jan C Minx, Gina Maskell, Abraham Marshall Nunbogu, Alexandre K Magnan, Shuaib Lwasa, Megan Lukas-Sithole, Tabea Lissner, Oliver Lilford, Steven F Koller, Matthew Jurjonas, Elphin Tom Joe, Lam T M Huynh, Avery Hill, Rebecca R Hernandez, Greeshma Hegde, Tom Hawxwell, Sherilee Harper, Alexandra Harden, Marjolijn Haasnoot, Elisabeth A Gilmore, Leah Gichuki, Alyssa Gatt, Matthias Garschagen, James D Ford, Andrew Forbes, Aidan D Farrell, Carolyn A F Enquist, Susan Elliott, Emily Duncan, Erin Coughlan de Perez, Shaugn Coggins, Tara Chen, Donovan Campbell, Katherine E Browne, Kathryn J Bowen, Robbert Biesbroek, Indra D Bhatt, Rachel Bezner Kerr, Stephanie L Barr, Emily Baker, Stephanie E Austin, Ingrid Arotoma-Rojas, Christa Anderson, Warda Ajaz, Tanvi Agrawal, and Thelma Zulfawu Abu. 2021. https://doi.org/10.1038/s41558-021-01170-y A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change*, 11(11):989–1000.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. https://doi.org/https://doi.org/10.1016/j.frl.2022.102776 Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J Brecha,

Michael Hegarty, Chelsea Jones, Kaylin Lee, Agathe Lucas, Nicole van Maanen, Inga Menke, Peter Pfleiderer, Burcu Yesil, and Jan C Minx. 2021. https://doi.org/10.1038/s41558-021-01168-6 Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change*, 11(11):966–972.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian Gehrmann. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. https://doi.org/10.18653/v1/N19-1423 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. Association for Computational Linguistics.

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*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Leonard Konlea and Fotis Jannidisa. 2020. Domain and Task Adaptive Pretraining for Language Models. *Proceedings http://ceur-ws. org ISSN*, 1613:73.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. https://doi.org/10.1093/bioinformatics/btz682 BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Haomin Lin, Jr-Shian Wu, Yu-Shiang Huang, Ming-Feng Tsai, and Chuan-Ju Wang. 2021. NFinBERT: A Number-Aware Language Model for Financial Disclosures (short paper). In *SwissText*.

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 Sasha Luccioni and Héctor Palacios. 2019. Using Natural Language Processing to Analyze Financial Climate Disclosures. In *Using Natural Language Processing to Analyze Financial Climate Disclosures*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. In *proceedings*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

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*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. https://doi.org/10.48550/ARXIV.2110.12010 Climatebert: A pretrained language model for climate-related text.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, and Xi Victoria Lin. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Supplementary Materials

A climateGPT-2 models, Inputs/Outputs samples

Table 4: Comparison of Climate Change related generated claims by GPT-2 vs climateGPT-2 models

	Claim generation downstream task	
GPT-2	**c GPT-2**	**ckey GPT-2**
Given title: Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns **Given Keywords:** NA	**Given title:** Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns **Given Keywords:** NA	**Given title:** Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns **Given Keywords:** climate change', 'hydropower', 'energy', 'mitigation', 'adaptation
Generated Claims: Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns, including drought. However this may not be enough to offset all impacts on water resources that occur over a period or several decades due primarily for some reason from climate variability such as increases across regions within each state's watersheds	**Generated Claims:** Hydroelectricity generation is expected to be affected by climate variability, which may have significant impacts on water resources systems. This study evaluates possible implications for reservoir operation under changing hydroclimate conditions using a high-resolution global river discharge model driven with two regional models.	**Generated Claims:** Climate-induced shifts will alter energy systems throughout much higher latitudes as well; however climate mitigation is not always possible for some regions or even globally at present because there has been little work to date about how these responses vary across different levels within a given basin due primarily to human activities.