# CliMedBERT: A Pre-trained Language Model for Climate and Health-related Text

Babak Jalalzadeh Fard[1], Sadid Hasan[2], Jesse E. Bell[1]

1. University of Nebraska Medical Center, College of Public Health

2. Microsoft AI

# Introduction

## Climate change - the **biggest health threat** facing humanity[1]

- Between 2030 and 2050, climate change is expected to cause approximately 250,000 additional deaths per year, from malnutrition, malaria, diarrhea and heat stress.

## With **timely and effective adaptation**, many of these threats or their effects can be reduced or avoided[2]

- But this require comprehensive studies and policies that are multi-sectoral, multi-system, and collaborative at different scales

[1] World Health Organization, "Cop24 Special Report Cop24 Special Report Health and Climate," *WHO*, 2015.

[2] G. Cissé *et al.*, "Health, Wellbeing, and the Changing Structure of Communities. In: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner," Cambridge, UK and New York, NY, USA, 2022.

# Introduction

## Such collaboration requires evidence synthesis

- But exponential growth in literature makes manual methods unattainable and biased toward more restrictive inclusion criteria[3]

# CliMedBERT

can serve as a foundation for various
downstream NLP tasks

[3] L. Berrang-Ford *et al.*, "Systematic mapping of global research on climate and health: a machine learning review," *Lancet Planet. Heal.*, vol. 5, no. 8, pp. e514–e525, Aug. 2021, doi: 10.1016/S2542-5196(21)00179-0

# Related Work

## ClimateBERT[4]

- pretraining of DistilROBERTA over 1.6 million paragraphs of climate-related texts.
- Showed considerable improvements compared to the base model over climate-related NLP tasks
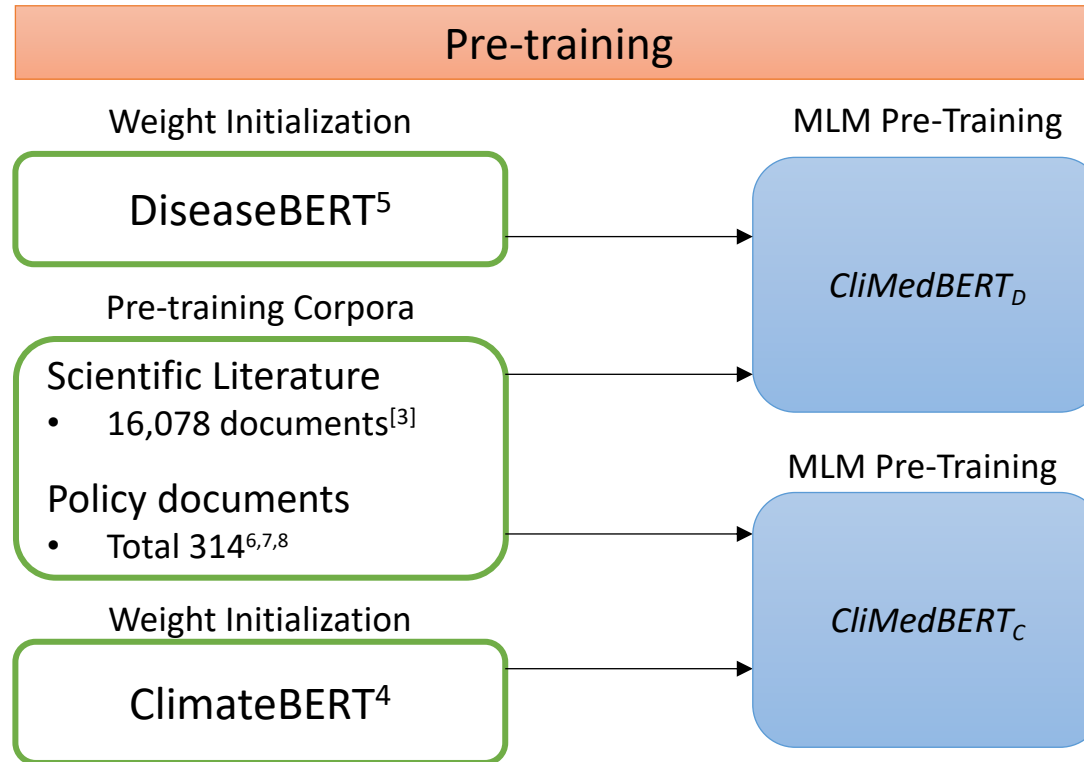
## diseaseBERT[5]

- It created the model through disease knowledge infusion training of BERT, ALBERT, and four pre-trained biomedical LMs
- showed improvements in three biomedical NLP tasks over their base models

[4]    N. Webersinke, M. Kraus, J. Anna Bingler, M. Leippold, F. Erlangen-Nuremberg, and E. Zurich, "ClimateBert: A Pretrained Language Model for Climate-Related Text," Oct. 2021, doi: 10.48550/arxiv.2110.12010.

[5]    Y. He, Z. Zhu, Y. Zhang, Q. Chen, and J. Caverlee, "Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition," *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 4604–4614, Oct. 2020, doi: 10.48550/arxiv.2010.03746.

# Proposed Model – First Step



Pre-training

Weight Initialization

DiseaseBERT[5]

MLM Pre-Training

$CliMedBERT_D$

Pre-training Corpora

Scientific Literature
- 16,078 documents[3]

Policy documents
- Total 314[6,7,8]

MLM Pre-Training

$CliMedBERT_C$

Weight Initialization

ClimateBERT[4]

[3] L. Berrang-Ford *et al.*, "Systematic mapping of global research on climate and health: a machine learning review," *Lancet Planet. Heal.*, vol. 5, no. 8, pp. e514–e525, Aug. 2021, doi: 10.1016/S2542-5196(21)00179-0

[6] L. Bornmann, R. Haunschild, and W. Marx, "Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents?," *Scientometrics*, vol. 109, no. 3, pp. 1477–1495, Dec. 2016, doi: 10.1007/S11192-016-2115-Y/TABLES/7.

[7] N. Dasandi, W. Cai, P. Friberg, S. Jankin, J. Kuylenstierna, and M. Nilsson, "The inclusion of health in major global reports on climate change and biodiversity," BMJ Glob. Heal., vol. 7, no. 6, p. e008731, Jun. 2022, doi: 10.1136/BMJGH-2022-008731.

[8] L. Bornmann, R. Haunschild$, K. Boyack, W. Marx$, and J. C. Minx, "How relevant is climate change research for climate change policy? An empirical analysis based on Overton data," Mar. 2022, doi: 10.48550/arxiv.2203.05358.

# Proposed Model – Second Step
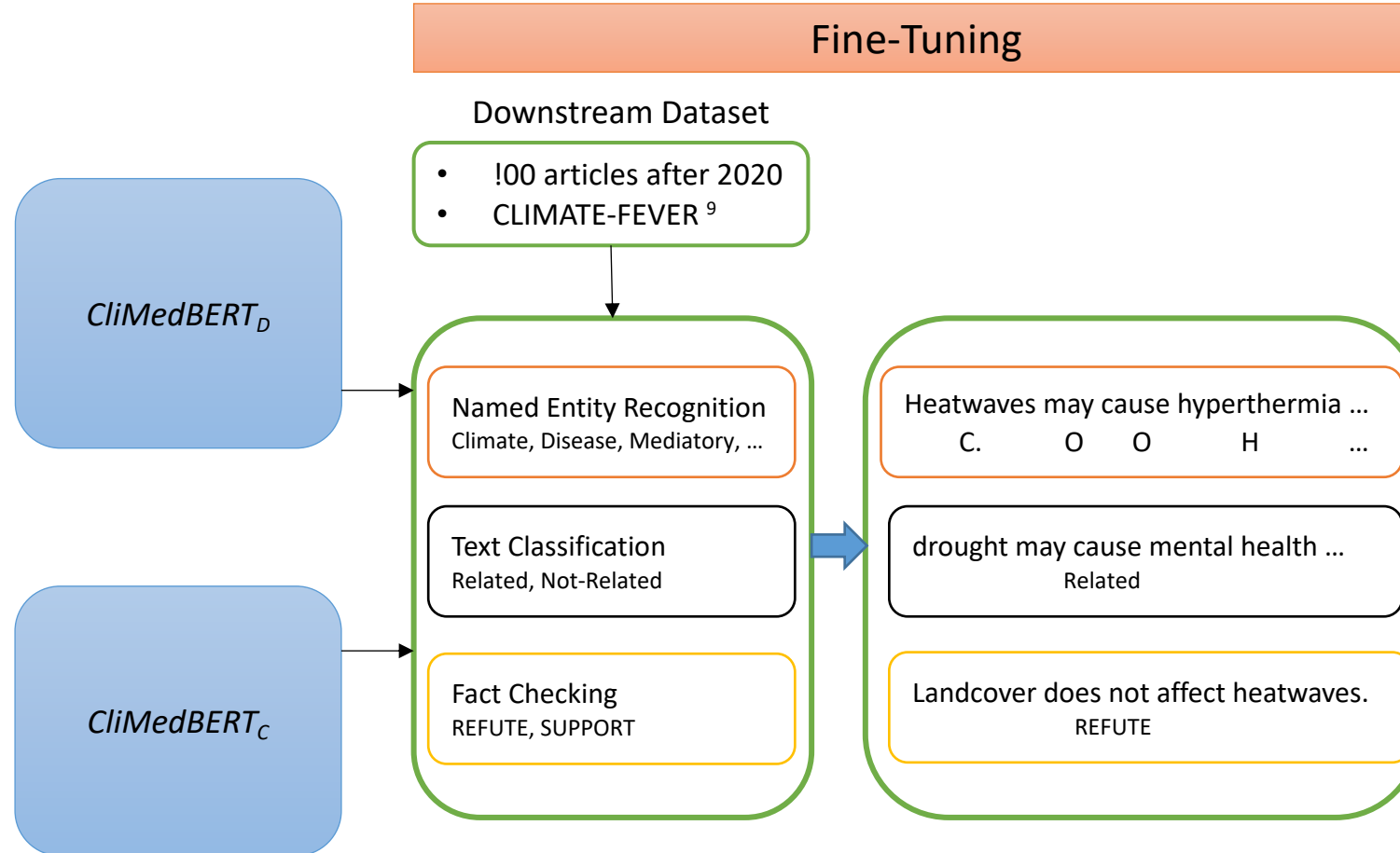
[9]    T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, "CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims," Dec. 2020, doi: 10.48550/arxiv.2012.00614.

# Expected Outcomes

- A freely accessible repository of the codes and developed models.

- An article about the process and results, including validation metrics and test results.

- A recreation of a recent evidence synthesis study[3] using our Language model and comparing the results.

- A classification engine for evaluating related texts and suggesting the inclusion of overlooked health or climate-related variables based on factors such as the location of the study.

[3] L. Berrang-Ford *et al.*, "Systematic mapping of global research on climate and health: a machine learning review," *Lancet Planet. Heal.*, vol. 5, no. 8, pp. e514–e525, Aug. 2021, doi: 10.1016/S2542-5196(21)00179-0

# Thank you for your attention