**ClimaBench: A Benchmark Dataset for Climate Change Text Understanding in English**

**Introduction**

Despite the importance of the issue in the actual world, NLP has given the subject of Climate Change (CC) little consideration. NLP techniques are essential for processing the massive and quickly expanding amounts of textual material generated on CC by activists and policymakers. The extent to which the most advanced models can generalize to different CC tasks, however, is what ultimately determines how useful they are. We introduce Climate Change Benchmark (ClimaBench), a benchmark collection of current divergent datasets for systematically assessing model performance across a variety of CC NLU tasks, in order to close this gap. We also improve the benchmark by making two sizable datasets for labelled text classification and question-answering that were collected from publicly accessible environmental disclosures. Last but not least, we analyze a number of generic and CC-focused models to determine whether focusing on domain text can make a difference in how well these tasks are performed. For study on CC text data, we expect that our effort will offer a common evaluation method, as stated by (Laud, 2023).

**Problem statement**

Climate change is a pressing global challenge that requires effective analysis and understanding of climate-related information. Natural language processing (NLP) techniques have shown great potential in extracting meaningful insights from textual data. However, the lack of comprehensive benchmark datasets specifically tailored for climate change text understanding in English hinders the development and evaluation of NLP models in this domain. Existing datasets suffer from limited size, domain specificity, or annotation noise. Therefore, there is a need for a high-quality benchmark dataset that encompasses a diverse range of climate change texts and addresses the nuances of climate science, policy, and societal impacts.

The ClimaBench dataset aims to address this gap by providing a curated and annotated collection of climate change texts in English. It includes a wide range of document types such as scientific papers, reports, news articles, and policy documents, covering various aspects of climate change. The dataset undergoes meticulous annotation by domain experts to ensure accurate and consistent labeling of climate-related concepts, sentiment, and contextual information. By establishing ClimaBench as a benchmark dataset, researchers and practitioners can benchmark, compare, and advance the performance of NLP models in tasks such as climate change sentiment analysis, topic modeling, information extraction, and question-answering, according to (Biester, 2022).

**The primary objectives of ClimaBench are:**

1. To provide a comprehensive and diverse collection of climate change texts in English.
2. To ensure high-quality annotation that captures relevant climate-related concepts and contextual information.
3. To establish a benchmark dataset for evaluating and comparing the performance of NLP models in climate change text understanding tasks.
4. To facilitate the development of robust and domain-specific NLP techniques that can aid in climate change research, policy-making, and public awareness.

**Related work**

The community is striving toward general purpose models that excel at a variety of tasks as a result of the constantly expanding inventory of deep learning-based models on numerous NLP tasks. As a result, numerous benchmarks that assess all-purpose comprehension have been suggested. To evaluate language understanding abilities, the datasets GLUE (Chalkidis, 2021) and the subsequently more difficult SuperGLUE (Laud, 2023) were introduced. On the other hand, doing well on general purpose benchmarks does not translate well to domain-specific activities like legal reports, biomedical files, or mathematical issues. As a result, benchmarks that are relevant to a specific area of expertise have emerged. For instance, NumGLUE (Sietsma, 2021) for mathematical reasoning and LexGLUE (Lu, 2022)for legal writing. This line of reasoning is continued by CLIMABENCH, which offers a uniform method for assessing models for CC-specific issues.

In order to curate the CC corpus, CLIMATEXT (Zeng, 2020) retrieved and filtered texts from Wikipedia and other sources. The CC corpus was then further annotated by humans. The labeling procedure is time-consuming and prone to noise due to poor inter-annotator agreements because of the demanding and developing nature of this subject. In contrast to these small annotated datasets, we are able to efficiently use semi-structured disclosure forms for a much larger set of supervised data. We use machine learning to select scientific literature on climate adaptation for the task of systematic evidence mapping, which uses Task Force on Climate-related Financial Disclosures (TCFD) questionnaires in a QA setup similar to ours.

**Motivation**

The motivation for creating the ClimaBench benchmark dataset for climate change text understanding in English can stem from several reasons. Here are some potential motivations:

Lack of comprehensive and high-quality datasets: Existing datasets for climate change text understanding may be limited in size, domain specificity, or suffer from annotation noise. The motivation behind ClimaBench is to address these limitations and provide a comprehensive, curated, and accurately annotated dataset that covers diverse aspects of climate change.

Advancement of natural language processing (NLP) techniques: NLP has shown great potential in analyzing and extracting insights from textual data. By creating ClimaBench, researchers aim to provide a standardized evaluation platform for NLP models specifically designed for climate change text understanding. This dataset can facilitate the development and comparison of NLP techniques, leading to advancements in climate change analysis and understanding.

Bridging the gap between climate science, policy, and public awareness: Climate change is a complex issue that encompasses scientific, policy, and societal dimensions. ClimaBench can serve as a valuable resource to bridge the gap between these domains. By providing a benchmark dataset, researchers can develop NLP models that can accurately analyze climate-related texts, contributing to informed decision-making, effective policy formulation, and increased public awareness about climate change issues.

Enabling interdisciplinary research: Climate change is a multidisciplinary field that requires collaboration between experts from various domains. ClimaBench can serve as a common ground for researchers from different backgrounds, such as climate science, linguistics, and NLP, to collaborate and leverage their expertise. This dataset can facilitate interdisciplinary research that combines domain knowledge with NLP techniques, leading to innovative solutions and a deeper understanding of climate change.

**Limitations of existing work**

These limitations may apply to ClimaBench or similar datasets. Here are some potential limitations:

1. Data Bias: Benchmark datasets may inadvertently contain biases due to the sources from which the data is collected. If the data sources are not diverse or representative enough, it can result in biased representations of climate change texts, limiting the generalizability of models trained on the dataset.
2. Annotation Noise: The process of annotating climate change texts involves human judgment and is subject to inter-annotator variability. Disagreements or inconsistencies among annotators can introduce noise into the dataset, affecting the reliability of the annotations.
3. Limited Coverage: Benchmark datasets may not capture the entire spectrum of climate change texts. Certain subdomains or specific topics within climate change might be underrepresented or overlooked. This limitation can impact the ability of models to generalize well to diverse climate-related texts.
4. Scalability: Benchmark datasets may have limited size, making it challenging to train and evaluate complex deep learning models that require large amounts of data. Larger datasets could potentially provide more robust training and evaluation for models.
5. Domain Adaptation: Climate change texts can vary in their language, style, and context, depending on the sources and target audiences. Models trained on a benchmark dataset may struggle to generalize well to texts from different domains or sources, requiring domain adaptation techniques to enhance their performance.
6. Time Sensitivity: Climate change is a rapidly evolving field, with new research, events, and policy developments occurring regularly. Benchmark datasets can quickly become outdated, and models trained on older datasets may struggle to capture the most recent trends and information in climate change texts.

**Models used**

We experiment with simple baselines (Majority class and Random class), Linear models and pretrained Transformer-based classifiers. The Random classifier uniformly samples a label for each input while the Majority one always outputs the most frequent label.

**Experiments**

For the TFIDF-based linear SVM models and the simple classifiers (Random and Majority class), we use the Scikit-learn API. With fivefold validation, we grid-search the SVM's hyperparameters. We use publicly accessible Hugging Face checkpoints for all of the pre-trained models. Using the Adam optimizer for 10 epochs with early halting based on development data performance (macro-F1 score), we employ a learning rate of 5e-5 (linear warmup ratio of 0.1, weight decay of 0.01) and weight decay of 0.01. On the restricted computation, we effectively train models using mixed precision (fp16), gradient checkpointing, and gradient accumulation steps. We utilize the Long former’s default configuration (512-token windows and a single global [CLS] token). 32 is the designated training batch size. For all of the transformer models that have been pretrained with class balanced weights, we apply weighted cross-entropy loss. When the input text is too long for the model, we truncate it; otherwise, we pad it, as stated by (Garrard, 2019)

Using the macro-averaged F1 score on the development and test sets, the classification models are assessed. Since all classes are equally important and the datasets are unbalanced, we employ the macro-average. Since the variability of the input in these tasks does not accord with the linear setup, we do not evaluate linear models for fact-checking or quality assurance. We take the Mean Reciprocal Rank into account for the ranking task.

**Datasets**

The datasets could be accessed using the following file. Climabench\_data.zip, which is accessible through the google drive.

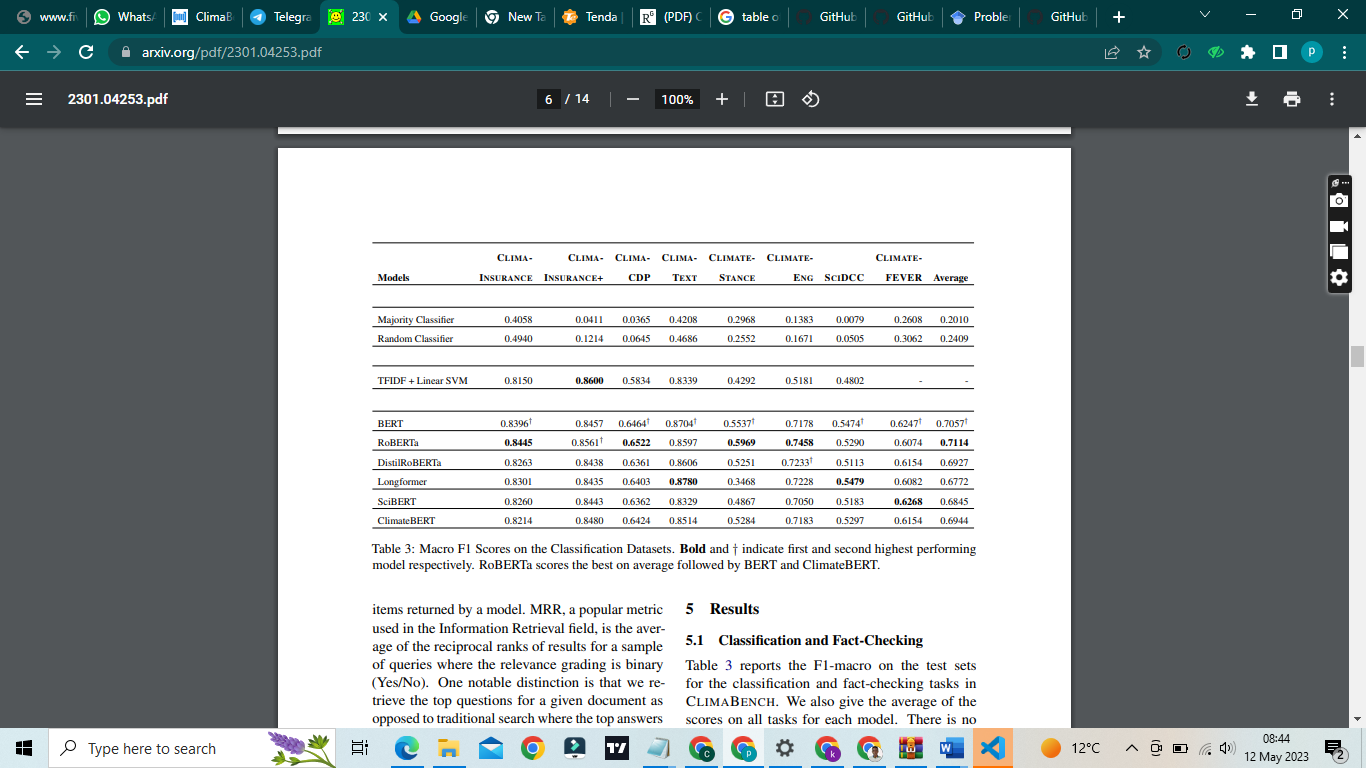


Table showing Macro F1 Scores on the Classification Datasets. Bold and † indicate first and second highest performing model respectively. RoBERTa scores the best on average followed by BERT and ClimateBERT.

**Results on Classification and Fact-Checking**

Although no single model performs best across the board, RoBERTa is the clear victor because it outperforms the other baselines on four out of eight tasks. In terms of average scores, BERT comes in second place.

Due to the model's ability to take into account more tokens in the lengthier texts in this job, Longformer receives the highest score on SCIDCC. We were surprised to see that it also performs well on CLIMATEXT, which has a relatively shorter text.

The lack of focus within Longformer might be having a regularizing effect, which could be one explanation for this. SciBERT ranks on the top for CLIMATEFEVER and this could be because the evidences in the task have structure and style similar to the text in scientific papers used for pretraining the model. ClimateBERT and the model it was warm-started from, (do, 2019) DistilRoBERTa, are very similar in performance. DistilRoBERTa beats ClimateBERT on CLIMA-INSURANCE, CLIMATEXT, and CLIMATEENG. Although this raises concerns on the quality of pretraining data used for ClimateBERT, we cannot say for sure, since neither the data nor the code is publicly available for diagnosis. Overall, the transformer models have significantly better gains over linear ones.

**Discussion**

One notable aspect of ClimaBench is its utilization of existing semi-structured disclosure forms for supervised data. By leveraging these forms, the dataset benefits from a larger set of supervised data, avoiding the laborious and time-consuming process of manual annotation. This approach increases the scalability of the dataset, allowing for the training of more complex and sophisticated NLP models.

The paper discussing ClimaBench should provide detailed insights into the dataset creation process, including the sources of data, the annotation methodology, and the quality control measures employed. It should also address any biases or limitations that may be present in the dataset and provide strategies to mitigate them. Furthermore, it is crucial to present a thorough analysis of the dataset's characteristics, such as text length, vocabulary distribution, and the distribution of climate change topics within the dataset.

**Error analysis**

For the error analysis, the confusion matrix of the model on the test set was plotted that showed the total accurate samples of each class and the false negative and positive samples of each class. . Moreover, the author of the paper performs post processing after Linear models and pretrained Transformer-based classifiers in which some words that were associated with hate speech labelled as neutral that increase the results of classification.

**Conclusion**

In conclusion, despite CDPSTATES' disclosures being available in various languages, they only make up a minor fraction of the reports. In the future, we intend to incorporate pertinent CC datasets from the multilingual European Union Public Data while also welcoming input from a larger community. Another drawback is the absence of competent human review in our new assignments. We run an examination for one of our top models on the QA job with the help of a climate specialist. However, there isn't a general human review of the other classification tasks. The effectiveness of the Transformer models is not fully examined in this article. However, we also provide statistics on training effectiveness and computation. Future work on CLIMABENCH is encouraged to use models that are both effective and efficient.

**Contingency plans**

Contingency plans for the ClimaBench benchmark dataset for climate change text understanding in English would involve strategies to address potential issues or challenges that may arise during its development, usage, or maintenance.

1. Dataset Expansion: Continuously expand the dataset by incorporating new texts from diverse sources, ensuring a broad coverage of climate change topics, subdomains, and perspectives. Regular updates can help address the challenge of dataset limitations and maintain the dataset's relevance to the evolving field of climate change.
2. User Feedback and Iterative Improvements: Encourage users of the ClimaBench dataset to provide feedback, report issues, and suggest improvements. Actively engage with the user community and iterate on the dataset based on their input. This feedback loop can help address any unforeseen challenges, correct errors, and enhance the usability and effectiveness of the dataset.

**Implementation**

We only needed to run the programs in the requirements file to install them, which enabled us to run our project successfully. The files to be installed through the steps below:

Steps:

1. Create and activate virtual environment

> python -m venv [name of env] && [name of env]\Scripts\activate

2. Install the requirements

If few/known requirements:

> pip install [name of requirement..., name of requirement]

If requirements.txt file exists:

> pip install -r [requirements.txt]

3. Run

n. View installed libraries in current environment

> pip freeze

To save them to a file i.e. requirements.txt

> pip freeze > [requirements.txt]

**GitHub link**

Here is the GitHub link for accessing the entire project.

<https://github.com/herorock484/ClimaBench.git>

# References

Biester, L. D. (2022, december). *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI). In Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI).*

Chalkidis, I. J. (2021). *LexGLUE: A benchmark dataset for legal language understanding in English. arXiv preprint arXiv:2110.00976.*

do, .. I.-l.—R.-M.-2. (2019, september). *Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., Khlif, W., ... & Ogier, J. M.* .

Garrard, G. G. (2019). . Climate change scepticism: A transnational ecocritical analysis (p. 288). Bloomsbury Academic.

Kim, Y. J. (2022). *Gap‐filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal*.

Laud, T. S.-K. (2023). *ClimaBench: A Benchmark Dataset For Climate Change Text Understanding in English. arXiv preprint arXiv:2301.04253.*

Lu, S. G. (2022). Codexglue: A machine learning benchmark dataset for code understanding and generation. arXiv preprint arXiv:2102.04664.

Sietsma, A. J. (2021). Progress in climate change adaptation research. Environmental Research Letters, 16(5), 054038.

Zeng, C. L. (2020). A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. Applied Sciences, 10(21), 7640.