Domain-Focused Versus General Model Efficacy in NLP Tasks on Climate Change

Anonymous ACL submission

Abstract

Global warming is a critical concern that requires both scientific understanding and public support for effective policy action. Stance detection using deep learning technologies, particularly large language models (LLMs) like GPT and BERT, can help analyze public and policy opinions on climate change. This study assesses the effectiveness of domain-specific pretraining versus general pretraining for stance detection tasks related to climate change, using a pretrained model named ClimateBERT. The aim is to determine if incorporating climate-specific knowledge into LLMs improves stance detection accuracy in climate-related discourse. The study compares the performance of ClimateBERT with general models like RoBERTa across various climate-related datasets. Results indicate that while domain-specific models offer some advantages, general-purpose models like RoBERTa often achieve higher accuracy and F1 scores, especially in fine-tuning settings. This suggests that robust general-purpose models are often sufficient for specialized tasks, highlighting the need to balance model architecture and domain adaptation for optimal performance in natural language processing applications.

1 Introduction

Global warming remains a critical concern, with wide-reaching impacts on natural and human systems (Grimm et al., 2015). To mitigate these challenges, deep learning-based global weather forecasting models such as KARINA, Graphcast, and FourcastNet have been developed, offering advanced predictive capabilities to better understand and respond to climate patterns (Cheon et al., 2024)(Pathak et al., 2022)(Lam et al., 2022).

The primary objective of this paper is to assess and compare the effectiveness of domain-specific pretraining versus general pretraining for stance detection tasks related to global warming and climate change through the pretrained model, named ClimateBERT (Webersinke et al., 2021). The study focuses on determining whether incorporating domain-specific knowledge on climate change into the pretraining of LLMs can improve the accuracy of stance detection in climate-related discourse. By enhancing the performance of stance detection models, this research aims to provide more effective tools for gauging public opinion and improving engagement strategies in the fight against global warming (Maibach et al., 2011). This contribution is essential for leveraging NLP technologies in environmental science, thereby aiding efforts to address one of the most pressing global challenges (Kawintiranon and Singh, 2021a).

2 Related Works

Kawintiranon and Singh introduced a novel approach, termed Knowledge Enhanced Masked Language Modeling (KE-MLM), integrated stance-specific knowledge by selectively masking words that are statistically significant in distinguishing between stances in the context of the 2020 US Presidential election. The researchers used two datasets for stance detection: one unlabeled dataset with over 5 million tweets from the 2020 US Presidential election and another labeled dataset of 2,500 tweets, divided equally between Joe Biden and Donald Trump, annotated for support, opposition, or neutrality. KE-MLM outperformed both the original BERT and fine-tuned BERT models in stance detection, achieving F1 macro scores of 0.7577 for Biden and 0.7877 for Trump, compared to lower scores achieved by the other models (Kawintiranon and Singh, 2021b).

Inkpen and Caragea developed a substantial dataset consisting of 21,574 English tweets related to political figures such as Donald Trump, Joe Biden, and Bernie Sanders. This dataset is designed for the task of stance detection, where tweets are annotated to indicate whether the sentiment expressed is in favor of, against, or neutral towards the targeted political figure. To handle this dataset, the team employed a variety of deep learning models, particularly focusing on the BERTweet model, which achieved a macro-average F1-score of 80.21%. This result showed that the BERTweet outperformed the original BERT which yielded 76.27 %. This performance underscores the effectiveness of using advanced language models fine-tuned on large, domain-specific datasets for improved stance detection in social media texts (Li et al., 2021).

Grasso et al. evaluated various BERT-based models on the stance detection task using the EcoVerse dataset, which includes 3,023 English tweets related to environmental issues. The evaluation showed that RoBERTa and its distilled version, DistilRoBERTa, performed the best with accuracy scores of 81.29% each. The specialized ClimateBERT models showed varied performance, with ClimateBertF scoring 69.60%, ClimateBertS at 72.51%, and ClimateBertS+D at 75.44% in accuracy (Grasso et al., 2024).

Schimanski1 et al. described the development and application of ClimateBERT-NetZero, a specialized NLP model for detecting net zero and emission reduction targets in text. The model was trained using a dataset of 3,500 expert-annotated text samples focused on sustainability commitments. ClimateBERT-NetZero achieved an impressive accuracy of 96.6% with a standard deviation of 0.004, outperforming both DistilRoBERTa and RoBERTa-base models in similar tests. Furthermore, the study described how this model can analyze the ambitiousness of these targets in real-world texts, such as earnings call transcripts, highlighting its practical applications for tracking corporate and institutional climate actions (Schimanski et al., 2023).

Webersinke et al. described the development of ClimateBERT, a language model specifically pretrained on over 2 million paragraphs of diverse climate-related texts sourced from news, corporate disclosures, and scientific articles. This model significantly enhanced performance on NLP tasks by incorporating domain-specific pretraining, which is crucial because traditional models trained on a general text show limited effectiveness in handling specialized climate-related terminology and contexts. By adapting the model to this niche, the authors achieved a 48% improvement on a masked language model objective, leading to significant error rate reductions between 3.57% and 35.71% across various downstream tasks such as text classification, sentiment analysis, and fact-checking. This demonstrated the model's capability to provide more accurate analyses of climate-related texts, supporting deeper insights into environmental discourse (Webersinke et al., 2021).

Whereas specialized models like ClimateBERT frequently outperform general models like BERT in stance detection tasks related to politics, our observations with ClimateBERT applied to climate-related tasks did not show a significant improvement over the original BERT model. This surprising outcome motivates more research into the possible causes of this performance disparity. To gain further insight into the subtleties of ClimateBERT's performance, we intend to apply it to a wider range of natural language processing jobs. We hope to pinpoint particular domains in which ClimateBERT performs particularly well or poorly by expanding the range of tasks and situations it is evaluated in. By using this method, we can improve the model's training and fine-tuning procedures and possibly identify important variations in the data.

3 Materials and Methods

3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) was developed by researchers at Google and introduced in their 2018 paper. BERT is unique for its deep bidirectional training, where it learns information from both the left and right context of a token within all layers of its architecture. The model's architecture is built on the Transformer mechanism, and utilizes only encoder parts of the Transformer (Vaswani et al., 2017). For pre-training, BERT was trained on the BookCorpus with 800 million words and a version of the English Wikipedia containing 2,500 million words. BERT also utilizes two innovative training strategies: Masked Language Model (MLM) and Next Sentence Prediction (NSP), which help it understand language context and relationships between sentences (Devlin et al., 2018).

3.2 RoBERTa

RoBERTa (Robustly optimized BERT approach) is an enhanced version of BERT, designed for enhanced performance through several main optimizations. It involves training the model for longer durations with larger batches over more extensive datasets, enabling it to learn from a more diverse range of data. Unlike BERT, RoBERTa removes the Next Sentence Prediction (NSP) objective, simplifying the training process. It also trains on longer sequences, allowing it to capture more context within texts. Additionally, RoBERTa employs a dynamically changing masking pattern, ensuring that the masked tokens vary with each epoch, preventing the model from seeing the same masked sequence twice (Liu et al., 2019).

3.3 DistilBERT

DistilBERT uses knowledge distillation during the pre-training phase, reducing the model size by 40% and increasing speed by 60%, while only sacrificing about 3% of BERT's performance. It leverages a technique called knowledge distillation, where the smaller DistilBERT model (the student) is trained to mimic the larger BERT model (the teacher). This process involves learning not just from the final outputs but also from the intermediate layers of BERT. DistilBERT achieves its compactness by reducing the number of layers from 12 to 6, significantly decreasing computational requirements. Despite being 60% smaller and 60% faster than BERT, it retains 97% of BERT's performance on various NLP benchmarks (Sanh et al., 2019).

4 Experiments

4.1 Dataset Description

For the experiment, a total of five different datasets were utilized, all sourced from Hugging Face: Climate Environmental Claims, Climate Detection, Climate Sentiment, Climate Commitment Actions, and Climate Specificity. The Climate Environmental Claims dataset supports a binary classification task, determining whether a given sentence constitutes an environmental claim. The Climate Detection dataset supports a binary classification task of identifying whether a given paragraph is climate-related. The Climate Sentiment dataset involves a ternary sentiment classification task, categorizing climate-related paragraphs as expressing opportunity, neutrality, or risk. The Climate Commitment Actions dataset supports a binary classification task, identifying whether a paragraph discusses climate commitments and actions. Lastly, the Climate Specificity dataset supports a binary classification task, assessing whether a climate-related paragraph is specific (Team, 2024). Examples from each dataset are detailed in the table below.

Based on the findings from the existing studies, we can conclude that both model architecture and domain adaptation are crucial for the performance of large language models (LLMs). Advanced architectures like BERT and RoBERTa provide a robust foundation, but pretraining and fine-tuning within a specific domain significantly enhance their effectiveness. Domain adaptation, particularly for political contexts, is especially important, as demonstrated by the superior performance of models like PoliBERTweet over general models such as RoBERTa and BERTweet in tasks like stance detection. This highlights that domain-specific pretraining can lead to substantial gains in accuracy and reliability, underscoring the importance of considering both architecture and domain adaptation in developing LLMs (Burnham, 2024)(Burnham, 2023).

4.2 Experiment Description

Based on the findings from previous studies, which highlight the critical role of both model architecture and domain adaptation, we hypothesize that these factors will similarly influence performance on climate-related datasets. Specifically, the prior research demonstrates that domain adaptation, especially in politically sensitive areas, significantly enhances model performance. To test whether these results are consistent with climate-related tasks, we will conduct experiments using various LLMs, including those with general architectures like RoBERTa and those adapted to specific domains. By comparing the performance of these models on climate environment claims, climate detection, climate sentiment, climate commit action, and climate specificity datasets, we aim to verify if domain-specific pretraining leads to similar gains in accuracy and reliability in the context of climate-related data. This experiment will help determine whether the importance of domain adaptation observed in political contexts extends to other specialized domains, such as climate science.

5 Results

The following experiments were evaluated using the F1 score to ensure a robust comparison of model performance across different tasks. We first conducted experiments on Climate Stance datasets. Since Webersinke et al. already performed the same experiment, we brought the re-

Datasets	Example
Climate Environmental Claims	The project will make a significant contribution to the German and European hydrogen strategy and hence to achievement of the climate targets.
Climate Detection	A material portion of this network is still relatively immature and there are risks that may develop over time. For example, it is possible that branches may not be able to sustain the level of revenue or profitability that they currently achieve (or that it is forecasted that they will achieve).
Climate Sentiment	We emitted 13.4 million tonnes CO2 of Scope 2 (indirect emissions), being emissions arising from our consumption of purchased electricity, steam or heat. Our Scope 3 emissions include emissions from a broad range of sources, including shipping and land transportation. More details on our Scope 3 emissions will be available in our 2014 report.
Climate Commitments actions	The Group is not aware of any noise pollution that could negatively impact the environment, nor is it aware of any impact on biodiversity. With regards to land use, the Group is only a commercial user, and the Group is not aware of any local constraints with regards to water supply. The Group does not believe that it is at risk with regards to climate change in the near-or mid-term.
Climate Specificity	Climate change is a challenge faced by the entire P&C insurance industry. In particular, our home insurance business has been affected due to changing climate patterns and an increase in the number and cost of claims associated with severe storms. Water damages now make up more than half of our home insurance claims.

Table 1: Examples of datasets and their contents used in the experiment

sults from that paper. DistilRoBERTa achieved an F1 score of 0.825, while the different variants of ClimateBERT, namely ClimateBERT_F, ClimateBERT_S, ClimateBERT_D, and ClimateBERT_D+S, scored 0.838, 0.836, 0.835, and 0.834, respectively. RoBERTa, when fine-tuned for stance detection, achieved the highest F1 score of 0.84375. Despite the domain-specific adaptations of the ClimateBERT models, they did not surpass the performance of the general-purpose RoBERTa in the fine-tuning setting. Additionally, the loss values for these models ranged from 0.138 to 0.150, with DistilRoBERTa exhibiting the highest loss.

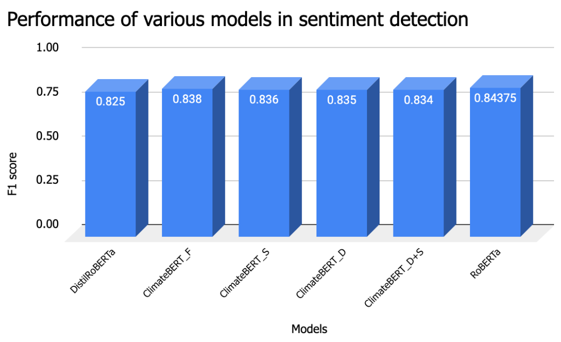

Figure 1: Performance of various models in stance detection based on ClimateBERT

The experimental results provide a nuanced view regarding the hypothesis that both model architecture and domain adaptation are important for the performance of LLMs. The data show that fine-tuning generally results in superior performance compared to zero-shot learning across various climate-related datasets. However, the results indicate that the general-purpose RoBERTa model often outperforms the domain-adapted models, especially in fine-tuning contexts. For instance, in the Climate Specificity and Climate Commitment Actions datasets, RoBERTa achieved the highest fine-tuning scores, surpassing both Distil-RoBERTa and ClimateBERT. Notably, ClimateBERT showed the lowest scores in several fine-tuning tasks, such as in the Climate Detection and Climate Environment Claim datasets, where it failed to outperform even the distilled version of RoBERTa.

Furthermore, using the net zero datasets introduced in the related work section, ClimateBERT achieved a score of 0.966, DistilRoBERTa yielded 0.959, and RoBERTa-base gained 0.963. Although ClimateBERT attained first place among the three models, the differences are quite small (Schimanski et al., 2023).

This suggests that while domain adaptation can enhance performance in zero-shot settings, it does not consistently provide an advantage over well-architected general models when fine-tuning is applied. Moreover, the time and resources required for domain adaptation might not always result in proportional gains in performance efficiency. Thus, while domain-specific pretraining has its merits, particularly in zero-shot contexts, the overall effectiveness and efficiency of using domain-adapted models versus robust general models like RoBERTa should be carefully evaluated based on the specific requirements and constraints of the task at hand. These findings highlight the importance of considering both model architecture and the practicality of domain adaptation in achieving optimal performance for specialized tasks such as climate-related analyses. The detailed summary of the results is summarized in Table 2.

6 Discussion

One potential reason behind this result is that the general-purpose language models, such as RoBERTa, are already sufficiently robust and versatile to handle a wide range of topics, including climate-related content, without needing extensive domain-specific adaptations. The relatively small differences in performance among ClimateBERT, DistilRoBERTa, and RoBERTa-base suggest that the underlying model architecture and general language understanding capabilities play a more critical role than the specialized domain knowledge for these tasks. This indicates that while domain adaptation can provide some benefits, the gains may not be substantial enough to justify the additional complexity and resources required for domain-specific pretraining in certain contexts.

Furthermore, comprehensive surveys on domain specialization techniques suggest that while domain-specific adaptations can improve performance, the benefits are sometimes marginal compared to the robust baseline provided by general-purpose models. These insights indicate that for climate-related tasks, the general architecture and pretraining of models like RoBERTa are sufficiently powerful, making extensive domain-specific pretraining less critical. This highlights the importance of balancing the need for domain adaptation with the inherent strengths of general-purpose language models (Zhao et al., 2023).

Dataset Name	Metric	Roberta	Distil-Roberta	Climate-Bert
Climate Specificity	Fine-tuning	0.8375	0.80625	0.825
	Zero-shot	0.4125	0.5875	0.5875
Climate Commitment Actions	Fine-tuning	0.84375	0.8875	0.85
	Zero-shot	0.34375	0.65625	0.65625
Climate Detection	Fine-tuning	0.975	0.965	0.965
	Zero-shot	0.77	0.23	0.23
Climate Environment Claim	Fine-tuning	0.886364	0.924242	0.901515
	Zero-shot	0.265152	0.265152	0.265152

Table 2: Performance results by diverse model and dataset for each circumstance.

7 Limitation

The datasets included in this study may not fully represent the range of discourse associated with climate change because they were retrieved from particular sources. The findings' applicability to other settings or domains relevant to climate change may be impacted by this constraint. Discussions about climate change varies greatly throughout various platforms, such as social media, policy-making, scientific research, and the media (Mavrodieva et al., 2019). A model that performs well on training data but finds it difficult to generalize to new, unseen data from other settings may result from the datasets' restricted emphasis. Further investigations ought to integrate a wider range of information in order to enhance the model's resilience and suitability for a variety of climate-related discourses.

8 Conclusion

This paper explored the performance of general-purpose and domain-specific language models on climate-related tasks, with a focus on models such as RoBERTa, DistilRoBERTa, and various Climate-BERT variants. The results indicate that while domain-specific pretraining can offer some performance benefits, these gains are often marginal compared to the robust performance of well-architected general-purpose models. RoBERTa, in particular, consistently performed well across different datasets, both in fine-tuning and zero-shot settings, highlighting its versatility and robust architecture. The findings underscore the importance of balancing the inherent strengths of general-purpose models with the targeted improvements offered by domain adaptation. While domain-specific models can provide benefits in certain contexts, the versatility and robustness of models like RoBERTa are sufficiently powerful for many specialized tasks, including those related to climate. This points to a more nuanced approach in leveraging both general-purpose and domain-specific strategies to achieve optimal performance in natural language processing applications.

References

Michael Burnham. 2023. Stance detection with supervised, zero-shot, and few-shot applications. *arXiv preprint arXiv:2305.01723*.

Michael Burnham. 2024. Stance detection: A practical guide to classifying political beliefs in text.

Minjong Cheon, Yo-Hwan Choi, Seon-Yu Kang, Yumi Choi, Jeong-Gil Lee, and Daehyun Kang. 2024. Karina: An efficient deep learning model for global weather forecast. *arXiv preprint arXiv:2403.10555*.

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*.

Francesca Grasso, Stefano Locci, Giovanni Siragusa, and Luigi Di Caro. 2024. Ecoverse: An annotated twitter dataset for eco-relevance classification, environmental impact analysis, and stance detection. *arXiv preprint arXiv:2404.05133*.

Nancy B Grimm, Peter Groffman, Michelle Staudinger, and Heather Tallis. 2015. Climate change impacts on ecosystems and ecosystem services in the united states: process and prospects for sustained assessment. In *The US National Climate Assessment: Innovations in Science and Engagement*, pages 97–109. Springer.

Kornraphop Kawintiranon and Lisa Singh. 2021a. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735.

Kornraphop Kawintiranon and Lisa Singh. 2021b. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. 2022. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

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*.

Edward W Maibach, Anthony Leiserowitz, Connie Roser-Renouf, and CK Mertz. 2011. Identifying like-minded audiences for global warming public engagement campaigns: An audience segmentation analysis and tool development. *PloS one*, 6(3):e17571.

Aleksandrina V Mavrodieva, Okky K Rachman, Vito B Harahap, and Rajib Shaw. 2019. Role of social media as a soft power tool in raising public awareness and engagement in addressing climate change. *Climate*, 7(10):122.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

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*.

Tobias Schimanski, Julia Bingler, Camilla Hyslop, Mathias Kraus, and Markus Leippold. 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. *arXiv preprint arXiv:2310.08096*.

ClimateBERT Team. 2024. Climate specificity dataset. https://huggingface.co/datasets/climatebert/climate_specificity. Accessed: 2024-05-23.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Li Yun, Hejie Cui, Zhang Xuchao, Tianjiao Zhao, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.