Climate-NLI: A Model for Natural Language Inference and Zero-Shot Classification on Climate-Related Text

Anonymous ACL submission

Abstract

Climate change is one of the most significant challenges of our era, necessitating innovative solutions across multiple fields. Advancements in Natural Language Processing (NLP) offer a promising pathway, particularly through the development of generalized models applicable to various tasks. Despite recent progress, current specialized NLP models excel in individual tasks but require substantial domain-specific training data and fail to generalize well to new scenarios. This paper introduces the Climate-NLI, an approach that utilizes Natural Language Inference (NLI) models to create a versatile NLP framework. Experiment results on 10 climate-related datasets show that our proposed model obtained comparable results to the models that have been fine-tuned on task-specific datasets. Our proposed model can significantly reduce the use of computational resources by training only one general model that can be applied to different tasks.

1 Introduction

Climate change represents one of the most pressing challenges of our time, demanding innovative and efficient solutions across various domains. One such promising avenue lies in leveraging advancements in NLP to create generalized models that can be applied to a wide range of tasks. NLP has witnessed tremendous growth in recent years, with specialized models achieving state-of-the-art performance on individual tasks like sentiment analysis, machine translation, and question answering (Khurana et al., 2022; Maulud et al., 2021; Jiang and Lu, 2020; Tan et al., 2020; Yang et al., 2020; Patil et al., 2022). However, these models often require significant domain-specific training data and struggle to generalize to unseen scenarios (Torralba and Efros, 2011; Arjovsky et al., 2020). This presents a critical challenge: developing efficient and adaptable NLP systems capable of handling diverse tasks with limited resources.

This paper proposes the Climate-NLI[1] that leverages the power of NLI model to build a general-purpose NLP framework. NLI models determine the entailment relationship between a premise and a hypothesis sentence (Storks et al., 2020). We posit that the core reasoning capabilities of NLI models can be exploited to build a foundation for various NLP tasks. By learning to understand the semantic relationships between sentences, the model can be adapted to diverse applications without extensive task-specific training.

2 Related Works

NLI is a well-studied subtask of NLP with numerous applications. Recent work has explored methods that leverage automatically generated, label-specific natural language explanations to produce more reliable labels (Kumar and Talukdar, 2020). Beyond methods, specific datasets have been created for NLI tasks, such as the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) and its explained variant, e-SNLI (Camburu et al., 2018). The extensive research focus on NLI is understandable considering its usage in many things. NLI serves as a foundation for various tasks, including question answering (Jeong et al., 2021), textual entailment (Bowman et al., 2015; Camburu et al., 2018), and even text classification using few-shot and zero-shot settings (Schick and Schütze, 2021; Kim et al., 2020).

Zero-shot learning is one of the methods that has gained traction in text classification for automatic labeling. It is a technique that transfers knowledge from labeled classes to unseen ones (Wang et al., 2019). This approach often utilizes pre-trained language models (PLMs) like BERT and RoBERTa (Chen et al., 2022; Gao et al., 2023; Alcoforado et al., 2022; Gonsior et al., 2020; Bujel et al., 2021). However, most studies combined

[1] Our code is publicly available at https://github.com/fjoeda/climate-nli.

PLMs with other methods. Some studies enhanced the performance of the language models by incorporating domain knowledge to do zero-shot classification. For instance, the work by Chen et al. (2022) combined sentence BERT with knowledge graph embedding, achieving better results compared to PLMs alone. Gao et al. (2023) also utilized additional data containing label descriptions fed to RoBERTa as input, leading to significant accuracy improvements of up to 17% compared to using the original RoBERTa only. This highlights the importance of a model's ability to understand the relationships between words and concepts, which aligns with the core principles of NLI.

Yin et al. (2019) proposed a benchmark and a textual entailment framework that leverages NLI for zero-shot text classification. Wei et al. (2021) also explored the ability of the language models to perform zero-shot tasks, including zero-shot classification, by using inference on unseen task type. By leveraging pre-trained models with strong NLI capabilities, zero-shot learning can achieve robust performance even with limited labeled data.

3 Dataset

We performed the experiment on several datasets representing both text classification and natural language inference tasks limited to climate-related domain, including: Climate-Fever (Leippold and Diggelmann, 2020), ClimateStance, ClimateEng (Vaid et al., 2022), SciDCC (Mishra and Mittal, 2021), Climate Sentiment, Climate Detection (Webersinke et al., 2022), Climate Commitment, Climate Environmental Claim, Climate Specificity, and TCFD Recommendation (Bingler et al., 2022) as shown in Table 4. All datasets except Climate-Fever are for text classification task. We used each training, validation, and testing sets provided on each dataset. If the validation set is not provided, we split the validation set from the training data for each dataset with 90:10 proportion. Since the SciDCC dataset was published in a single CSV file, we split the dataset into training, validation and testing set with 80:10:10 proportion.

We performed additional pre-processing on the Climate-Fever and SciDCC datasets. The Climate-Fever dataset contains 1.5K climate change-related claims where each claim has five evidences. We converted the dataset into pairs of claim and evidence where each pair is labeled as "support", "refutes", or "not_enough_info". Following Webersinke et al. (2022), we filtered out the evidence sentences with "not_enough_info" label and focusing our model only on deciding whether a claim is supported or refuted. The SciDCC dataset contains 11,539 news articles taken from Science Daily, classified into 20 classes such as Earthquake, Hurricane, Pollution, etc. Each article consist of title, summary, and body content. In this work, we concatenated the title, summary, and body as the text input.

4 Methodology

The proposed model, Climate-NLI, was developed to handle both fact-checking and classification tasks for general climate-related text. The model was trained on the NLI setting. Using NLI, the model can solve the fact-checking task, and in the same time address the text classification problem using an entailment-based zero-shot classification. The development processes of the model are presented in this section.

4.1 Dataset Preparation

As mentioned earlier, we used an entailment-based approach for zero-shot classification. Therefore, all text classification datasets were converted into NLI task-setting in the preparation step by generating the entailment and contradiction samples. NLI takes two sentences as the premise and hypothesis and then decides whether those sentences are an entailment, neutral, or a contradiction.

Selecting Entailment Samples. The entailment samples from the text classification dataset are selected by adding the text data as the premise with the corresponding class label as the hypothesis. Beside the class label, the hypothesis is constructed from a template such as "The text is about <class name>" (e.g., "The text is about agriculture", "The text is about environment"). In terms of zero-shot classification tasks, the model will be provided with the text input along with its candidate labels. The label hypothesis which receives the highest entailment score will be selected as the predicted label for the text input.

Selecting Contradiction Samples. The contradiction samples are added to make the zero-shot classification model able to differentiate between labels. We followed Gera et al. (2022), who used the contrast-random approach for generating the contradiction samples. Contrast-random is the best performing setting along with contrast-all setting,

but in terms of computational cost, contrast-random setting is more preferred. Contrast-random approach will add the contradiction samples for each entailment samples with replaced class name on the hypothesis.

Adding Label Variation. We implemented label variation to introduce the model to the unseen labels. The addition of label variation to the hypothesis was done by replacing the corresponding label with its synonym. We used WordNet from NLTK package to find the list of the synonyms for the corresponding label. The label is then replaced with one of the synonym randomly. We applied the label variation specifically on topic classification datasets, including ClimateEng and SciDCC.

The Hypothesis Templates. When it comes to zero-shot classification task, the entailment-based models such as *bart-large-mnli*[2] use the default hypothesis template like "The example is <class name>". In our case, since we used different datasets from various domains, we specified the hypothesis template based on the dataset as shown in the Table 4. Referring to that table, some hypothesis templates use a yes-no question format (e.g., "Does the text related to climate? c") to handle the binary classification tasks where the class names only consist of "yes" and "no".

4.2 Model Training

The Climate-NLI model was developed by fine-tuning ClimateBert (Webersinke et al., 2022) on NLI-task setting. ClimateBert is a transformer-based language model that has been pre-trained on over 2 million paragraphs of climate-related texts, such as common news, research articles, and climate reporting of companies. ClimateBert used DistilRoBERTa-base[3], a distilled version of RoBERTa containing 82M parameters, as the starting point of training (Sanh et al., 2020). Climate-Fever and all the converted text classification datasets as shown in Table 4 were used to fine-tune the model. In total, there are 45,802 pairs of premise and hypothesis along with their labels that were used as the training data. In addition to that, 5,498 pairs were used as validation set. The best model was selected based on the best validation accuracy. The Climate-NLI model was trained with specific hyperparameter settings (see Table 1). The text length for each premise and hypothesis was limited to 256 each, to fit the overall limit of 512.

Hyperparameter	Values
Max. sequcence length	512
Batch size	16
Optimizer	AdamW
Learning rate	$5 \cdot 10^{-5}$
Max. num. of epochs	50
Num. of early stopping patience	5

Table 1: Hyperparameter for NLI model training.

We also conducted different experiments by fine tuning ClimateBert on each task-specific dataset with similar hyperparameter settings. Moreover, as the baseline comparison for NLI-based task, we used *bart-large-mnli*, a pre-trained model with 409M parameters, trained on the Multi-Genre Natural Language Inference (MultiNLI) corpus which contains crowd-sourced collection of 433K sentence pairs annotated with textual entailment information. All experiments were performed on a single NVIDIA A100 GPU and the random state was set to 42.

4.3 Model Evaluation

We evaluated the Climate-NLI model on the test set for each task-specific dataset. For the fact-checking tasks on the Climate-Fever, we directly used the NLI setting for the inference process and mapped the label, specifically "Support" to entailment and "Refutes" to contradiction. In this work, we only focused on how good the model is in determining whether evidence supports or refutes a claim. Meanwhile, for all classification tasks, we use a zero-shot classification procedure to predict the final label. The Climate-NLI model will be presented with a text input as the premise and a set of label candidates prepended with a template as a hypothesis. In the model output, we took the entailment and contradiction score and applied a softmax function. The label with the highest entailment score will be chosen as the final label.

With the same procedure, we also evaluate the pretrained *bart-large-mnli* model as the baseline comparison for NLI-based model. We also evaluate the fine-tuned ClimateBert models on their corresponding task-specific training dataset. Accuracy and macro-averaged F1 were used as the evaluation metrics.

[2] https://huggingface.co/facebook/bart-large-mnli

[3] https://huggingface.co/distilbert/distilroberta-base

Dataset	Climate-NLI Model		Bart-Large-MNLI		Fine-tuned ClimateBert	
	Acc.	F1	Acc.	F1	Acc.	F1
ClimateEng	0.79	0.66	0.56	0.45	0.78	0.67
ClimateStance	0.78	0.42	0.44	0.70	0.81	0.52
SciDCC	0.52	0.40	0.31	0.25	0.61	0.49
Climate Commitment	0.78	0.74	0.31	0.24	0.80	0.78
Climate Env Claim	0.86	0.84	0.26	0.21	0.92	0.90
Climate Sentiment	0.74	0.73	0.36	0.25	0.81	0.80
Climate Specificity	0.77	0.75	0.44	0.42	0.82	0.79
TCFD Recomm	0.75	0.69	0.21	0.17	0.79	0.74
Climate Detection	0.94	0.90	0.80	0.46	0.96	0.94
Climate-Fever	0.81	0.77	0.57	0.39	0.85	0.81
Average	0.774	0.69	0.426	0.321	0.815	0.744

Table 2: Evaluation results.

5 Result and Discussion

The performance of all models is detailed in Table 2. Notably, Climate-NLI surpasses *bart-large-mnli* on every dataset despite having fewer parameters. This is likely because Climate-NLI was trained using climate-focused data, whereas *bart-large-mnli* was trained on a broader range of information. However, compared to the fine-tuned ClimateBert model, Climate-NLI obtained slightly lower performances in all datasets.

Despite the lower performance, we argue that our proposed model can significantly reduce the use of computational resources, by training only one general model that can be applied for different tasks. Moreover, in terms of adaptability to the new classes, entailment-based zero-shot classification model is capable to adapt to any newly added class by adding the new training samples. Meanwhile, the fine-tuned classification model needs to be retrained when a new class is introduced, since the number of classes is already defined before the training process (Patadia et al., 2021).

Dataset	Acc.	F1
ClimateEng	0.76	0.66
ClimateStance	0.80	0.46
SciDCC	0.43	0.35
Climate Commitment*	0.32	0.26
Climate Env Claim*	0.63	0.62
Climate Sentiment	0.80	0.78
Climate Specificity*	0.46	0.46
TCFD Recomm*	0.23	0.20
Climate Detection	0.84	0.67
Climate-Fever	0.84	0.80

* Unseen dataset

Table 3: Climate-NLI evaluation results on unseen dataset.

The Capability of Predicting The Unseen Dataset. We performed an experiment by excluding four datasets: Climate Commitment, Climate Environmental Claim, Climate Specificity, and TCFD Recommendation. As shown in the Table 3, the model's performance on those datasets decreased significantly. The drop is likely due the lacks of training samples for NLI tasks which only relies on Climate-Fever dataset. This makes the model unable to decide an entailment between premises and hypotheses which are completely unseen.

Potential Implementation. Zero-shot classification has capability of being used across unseen dataset and unseen labels (Pushp and Srivastava, 2017). Despite the mediocre performance on the unseen datasets (see Table 3), zero-shot classification model can be implemented for automatic data labeling through weak supervision where the model is expected to provide hints about the desired class from the defined candidate labels (Åslund, 2021; Wang et al., 2021). This could reduce the time needed to develop a dataset related to climate change.

6 Conclusion

In this paper, we presented Climate-NLI, an NLI-based model specifically designed for fact-checking and zero-shot classification tasks. Evaluation results show that Climate-NLI successfully outperformed *bart-large-mnli*, the NLI model trained on more general text, while obtained slightly lower performance compared to the task-specific fine-tuned ClimateBert model. Our proposed model has better adaptability to new classes by adding the training samples instead of retraining the model with the whole training samples. Moreover, the general model, significantly reduced the use of computational resources.

Limitations

In terms of the fact-checking task, we only tested how good the model was at deciding whether a claim is supported or refuted by evidence, which is just one of the parts of the fact-checking pipeline. A further test of the Climate-NLI model on the whole fact-checking pipeline from evidence retrieval to entailment prediction can be done as future work.

To simplify the training pipeline in the model training process, we only use the yes-no question template followed by a "yes" or "no" label for the binary classification tasks. Instead of relying on a yes-no question as a template, we may extend the "yes" and "no" labels to a sentence that shows the complete context related to the label.

Ethics Statement

We ensure that our work complies with the ACL Ethics Policy.

References

Alexandre Alcoforado, Thomas Palmeira Ferraz, Rodrigo Gerber, Enzo Bustos, André Seidel Oliveira, Bruno Miguel Veloso, Fabio Levy Siqueira, and Anna Helena Reali Costa. 2022. *ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling*, page 125–136. Springer International Publishing.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization.

Jacob Åslund. 2021. *Zero/Few-Shot Text Classification : A Study of Practical Aspects and Applications*. Ph.D. thesis, KTH, School of Electrical Engineering and Computer Science, Stockholm, Sweden.

Julia Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. How Cheap Talk in Climate Disclosures relates to Climate Initiatives, Corporate Emissions, and Reputation Risk.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Kamil Bujel, Helen Yannakoudakis, and Marek Rei. 2021. Zero-shot sequence labeling for transformer-based sentence classifiers. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 195–205, Online. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. E-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen. 2022. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal*, 9(12):9205–9213.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-Shot Text Classification with Self-Training.

Julius Gonsior, Maik Thiele, and Wolfgang Lehner. 2020. Weakal: Combining active learning and weak supervision. In *Discovery Science*, pages 34–49, Cham. Springer International Publishing.

Minbyul Jeong, Mujeen Sung, Gangwoo Kim, Donghyeon Kim, Wonjin Yoon, Jaehyo Yoo, and Jaewoo Kang. 2021. Transferability of Natural Language Inference to Biomedical Question Answering.

Kai Jiang and Xi Lu. 2020. Natural language processing and its applications in machine translation: A diachronic review. In *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, pages 210–214.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.

Youngwoo Kim, Myungha Jang, and James Allan. 2020. Explaining Text Matching on Neural Natural Language Inference. *ACM Transactions on Information Systems*, 38(4):39:1–39:23.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Markus Leippold and Thomas Diggelmann. 2020. Climate-FEVER: A Dataset for Verification of Real-World Climate Claims. In *Climate Change AI*. Climate Change AI.

Dastan Maulud, Subhi Zeebaree, Karwan Jacksi, Mohammed M.Sadeeq, and Karzan Hussein. 2021. A state of art for semantic analysis of natural language processing. *Qubahan Academic Journal*, 1.

Prakamya Mishra and Rohan Mittal. 2021. NeuralNERE: Neural Named Entity Relationship Extraction for End-to-End Climate Change Knowledge Graph Construction. In *Climate Change AI*. Climate Change AI.

Devika Patadia, Shivam Kejriwal, Pashva Mehta, and Abhijit R. Joshi. 2021. Zero-shot approach for news and scholarly article classification. In *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)*, pages 1–5.

Spandan Pankaj Patil, Lokshana Chavan, Janhvi Mukane, Deepali Rahul Vora, and Vidya Chitre. 2022. State-of-the-art approach to e-learning with cutting edge nlp transformers: Implementing text summarization, question and distractor generation, question answering. *International Journal of Advanced Computer Science and Applications*.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train Once, Test Anywhere: Zero-Shot Learning for Text Classification.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards Fine-grained Classification of Climate Change related Social Media Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2).

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-Class: Text Classification with Extremely Weak Supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3043–3053, Online. Association for Computational Linguistics.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A Pretrained Language Model for Climate-Related Text.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

A Dataset Description

The detailed description of the datasets are shown in Table 4.

Dataset	Task	Data Composition	Num. of Classes	Hypothesis Template
ClimateEng	Classification	Train: 2781; Val: 354; Test: 355	5	This example is about c
Climate Stance	Classification	Train: 2781; Val: 354; Test: 355	3	The stance of this tweet regarding to climate change is c
SciDCC	Classification	Train: 11539	20	This example is about c
Climate Commitment	Classification	Train: 1000; Test: 320	2	Does text talk about climate commitment action? c
Climate Environmental Claim	Classification	Train: 2117; Test: 265	2	Does the claim relate to environment? c
Climate Sentiment	Classification	Train: 1000; Test: 320	3	The text sentiment regarding climate change is c
Climate Specificity	Classification	Train: 1000; Test: 320	2	The text is climate change c
TCFD Recommendation	Classification	Train: 1300; Test: 400	5	Regarding climate recommendation, the text is about c
Climate Detection	Classification	Train: 1300; Test: 400	2	Does the text related to climate? c
Climate-Fever	Fact-checking (NLI)	Train: 2196; Test: 549	2	-

Table 4: The list of dataset used in the training phase along with their task, composition, the number of classes, and the hypothesis template. The class label in the hypothesis template is represented with "c". For the Climate-Fever dataset, we split the dataset with 80:20 train-test proportion and filter out the "not_enough_info" label in the data preprocessing step.