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

Faturahman Yudanto[1], Yunita Sari[2], Maeve Zahwa Adriana Crown Zaki[1]
Department of Computer Science and Electronics
Universitas Gadjah Mada, Yogyakarta, Indonesia
[1]{f.yudanto, maeve.zahwa.adriana.crown.zaki}@mail.ugm.ac.id,
[2]yunita.sari@ugm.ac.id

Abstract

Climate change is one of the most significant challenges of our era, necessitating innovative solutions across multiple fields. Advancements in NLP offer a promising pathway, particularly through the development of generalized models applicable to various tasks. Despite recent progress, 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 NLI models to create a versatile NLP model that can be used for fact-checking and text classification on climate-related text. 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 model improves adaptability to new classes by adding training samples without full retraining but struggles with certain classes due to limited related samples and similar but distinct concepts.

1 Introduction

Climate change represents one of the most pressing challenges of our time, demanding innovative and efficient solutions across various domains. A promising approach involves using Natural Language Processing (NLP) advancements to develop versatile models for various tasks. NLP has witnessed tremendous growth, with specialized models achieving state-of-the-art performance on individual tasks such as 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 various tasks with limited resources.

This paper proposes the Climate-NLI[1] that leverages the power of the Natural Language Inference (NLI) model to build a general-purpose NLP framework. NLI models are designed to determine the entailment between a premise and a hypothesis sentence (Storks et al., 2020). We posit that the 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 classification is one of the methods that has gained traction in text classification. 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) such as BERT and RoBERTa (Chen et al.,

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

2022; Gao et al., 2023; Alcoforado et al., 2022; Gonsior et al., 2020; Bujel et al., 2021). However, most studies combined 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. These tasks involve determining the entailment relationship between a premise and a hypothesis by essentially asking whether the hypothesis logically follows from the provided information (Storks et al., 2020).

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 types. 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 1. All datasets except Climate-Fever are for text classification tasks. We used each training, validation, and testing set provided on each dataset. If the validation set is not provided, we split the validation set from the training data for each dataset with a 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 an 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 and 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 the "not_enough_info" label and focused 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 consists of a 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 at 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 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. Besides 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 that receives the highest entailment score will be selected as the predicted label for the text input.

Selecting Contradiction Samples. The contra-

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 1: The list of datasets 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 an 80:20 train-test proportion and filtered out the "not_enough_info" label in the data preprocessing step.

diction 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 preferred setting in terms of performance and computational cost. The contrast-random approach will add the contradiction samples for each entailment sample with a 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 the NLTK package to find the list of the synonyms for the corresponding label. The label is then replaced with one of the synonyms 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 tasks, 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 Table 1. 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

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

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 1 were used to fine-tune the model. In total, there are 45,802 pairs of premises and hypotheses 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 2). The text length for each premise and hypothesis was limited to 256 each, to fit the overall limit of 512.

Hyperparameter	Values
Max. sequence 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 2: 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 the 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 a 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 the NLI-based model. We also adjust the hypothesis template for each dataset as shown in 1. For additional comparison, we also trained several ClimateBert models. Each model was individually fine-tuned on their corresponding task-specific training dataset. Macro-averaged F1 were used as the evaluation metrics.

Dataset	Climate-NLI	Bart-Large-MNLI	FT Climate-Bert
ClimateEng	0.66	0.45	0.67
ClimateStance	0.42	0.37	0.52
SciDCC	0.40	0.25	0.49
Climate Commitment	0.74	0.24	0.78
Climate Env Claim	0.84	0.21	0.90
Climate Sentiment	0.73	0.25	0.80
Climate Specificity	0.75	0.42	0.79
TCFD Recomm	0.69	0.17	0.74
Climate Detection	0.90	0.46	0.94
Climate-Fever	0.77	0.39	0.81
Average	0.69	0.32	0.74

Table 3: The F1 scores of Climate-NLI (Ours), Bart-Large-MNLI, and fine-tuned (FT) ClimateBert on each test set of the dataset. The Climate-NLI model was trained with all datasets combined, meanwhile fine-tuned Climatebert was trained on each dataset individually.

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

5 Result and Analysis

In this study, we compared three kinds of model specifically the *bart-large-mnli*, fine-tuned ClimateBert model on each dataset, and the Climate-NLI model (ours). We evaluated both fact-checking using the NLI approach and text classification tasks. For the NLI-based model such as *bart-large-mnli* and Climate-NLI, we use zero-shot classification approach to do the classification tasks.

The performance of all models is detailed in Table 3. Notably, Climate-NLI outperforms *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 on each dataset, Climate-NLI obtained slightly lower performances in all datasets. These performances are in line with Patadia et al. (2021) experiment results, where the entailment-based zero-shot classification model still failed to outperforms the text classification models trained on the task-specific datasets.

5.1 Text Classification Result

In this section, we discuss the Climate-NLI model performance on the zero-shot text classification task. The text classification datasets used to train the model are generally divided into binary and multi-class classifications. Figure 1 shows the distribution of the F1 scores for all classes across each dataset.

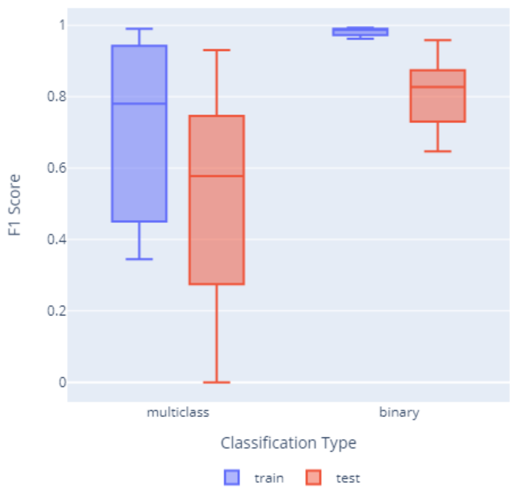

Figure 1: The distribution of F1 scores for all classes across each dataset.

As shown in Fig. 1 that the Climate-NLI model still struggles on the multi-class classification task. Compared to binary classification, the distribution of F1 scores on multi-class is wider than the binary classification even in the train set. This indicates the greater variability in performance across different datasets. The median F1 score in the multi-class classification is also lower, suggesting that the model has difficulty differentiating among multiple classes, as opposed to the simpler binary task. The lowest F1 score in the multi-class classification is 0, which reflects the model's inability to predict certain classes, leading to class imbalance issues. We will discuss the class imbalance issue further in the next section.

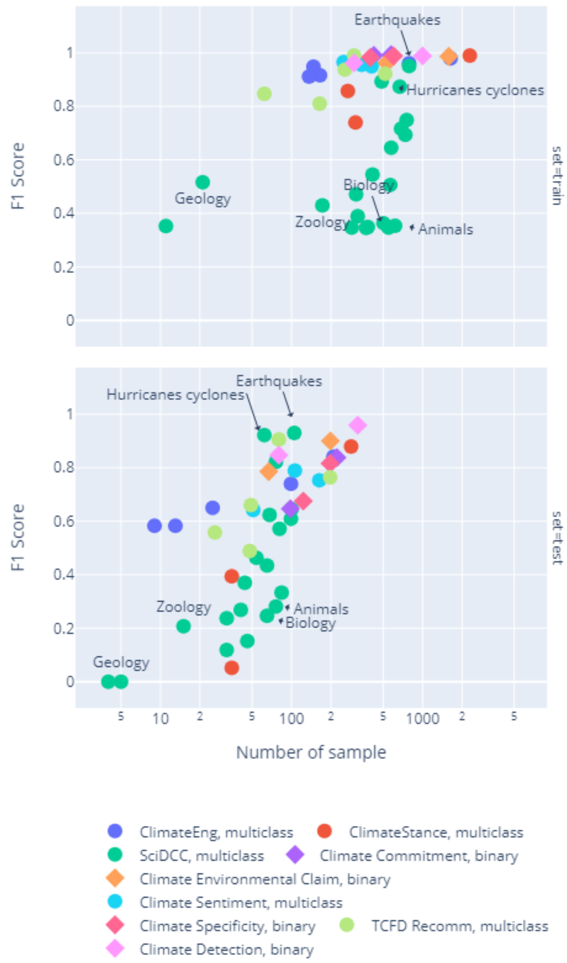

Figure 2: Relationship between the number of samples and the F1 Score for each class on climate-related datasets.

5.2 Class Imbalance Issue

We trained the Climate-NLI model with imbalanced datasets which likely influence the model performance. In Fig. 2, we showed the relationship between the F1 score and the number of training samples for each class based on the dataset and the classification type. Fig. 2 also presents the relationship on both train and test set.

Fig 2 shows that classes with smaller training samples tend to have a lower F1 score, especially in multi-class classification. In the SciDCC dataset, while the model can classify minority classes in the training set, it struggles in the test set. Moreover, with 20 classes in the SciDCC dataset, classification becomes challenging, especially for minority classes with very few samples—such as only 11 out of 9,231 for the "global warming" class or 21 out of 9,231 for the "geology" class. As a result, the model fails to generalize well on these minority classes. Similar results occur in the hardly distinguished majority classes such as "Animals", "Zoology", and "Biology". We present the comparison of the prediction results on both majority and minority classes of the SciDCC dataset in Table 4 and Table 5.

Example	Seismic activity of New Zealand's alpine fault more complex than suspected A rupture ...
Prediction **Ground Truth**	Earthquakes Earthquakes
Example	How has society adapted to hurricanes? A look at New Orleans over 300 years ...
Prediction **Ground Truth**	Hurricanes Cyclones Hurricanes Cyclones
Example	How do mantis shrimp find their way home?. Patel, a Ph.D. candidate in biological sciences at UMBC, found that the species of...
Prediction **Ground Truth**	Biology Animals
Example	Spectacular bird's-eye view? Hummingbirds see diverse colors humans can only imagine To find food...
Prediction **Ground Truth**	Zoology Animals

Table 4: Climate-NLI prediction samples on the majority class of SciDCC dataset

Table 4 shows that the model predicts the majority and specific classes correctly such as "Earthquakes" and "Hurricanes Cyclones". However, it struggles with distinguishing between overlapping classes, such as "Biology" versus "Animals" or "Zoology" versus "Animals," where the distinction is more nuanced. These errors highlight a limitation in handling closely related classes although the "Animals" class is considered as a majority class.

Example	Volcanic growth 'critical' to the formation of Panama Yet for scientists the exact process by ...
Prediction **Ground Truth**	Earthquake Geology
Example	Fishing for a theory of emergent behavior Some of the most difficult questions in science today ...
Prediction **Ground Truth**	Zoology Zoology
Example	Songbirds, like people, sing better after warming up Researchers at Duke University say there may be a good reason why birds ...
Prediction **Ground Truth**	Animals Zoology

Table 5: Climate-NLI prediction samples on the minority class of SciDCC dataset

In terms of predicting the minority classes, the model performed variably, correctly identifying some labels while struggling with others. Table 5 shows that the model correctly classified a text about "emergent behavior" under "Zoology", demonstrating its ability to match specific scientific content with the correct label. However, in another case, it incorrectly predicted "Earthquake" instead of "Geology" for a text on volcanic growth, and also incorrectly predicted "Animals" instead of the more specific "Zoology" on the songbird text. Those incorrect predictions are likely due to the model focusing on related but distinct concepts. Moreover, the class "Earthquake" has significantly more training samples than "Geology" class which makes the model tend to classify on the majority class over the minority ones.

The results suggest that while zero-shot classification is promising, further refinement or more context-specific candidate labels could improve its accuracy in specialized fields like scientific classification. Additionally, the zero-shot classification model can be applied to multi-label classification tasks when the labels are not highly distinctive.

5.3 Potential Implementation

Despite the lower performance compared to the fine-tuned model, the entailment-based zero-shot classification model is capable of adapting 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).

Zero-shot classification also has the capability of being used across unseen datasets and unseen labels (Pushp and Srivastava, 2017). Despite the mediocre performance on the minority classes and the difficulty in distinguishing certain similar classes, 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 obtaining 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. However, our model still struggles to classify certain classes due to limited training samples for related classes and the presence of similar but distinct concepts.

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 in the 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. Currently, we leave this as an open question.

Ethics Statement

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

Acknowledgements

This research is supported by a 2024 type A Grants from the Faculty of Mathematics and Natural Science, Universitas Gadjah Mada and also partially supported by the Department of Computer Science and Electronics, Universitas Gadjah Mada under the Publication Funding Year 2024.

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.