Climate Policy Transformer: Utilizing NLP to track Climate Commitments in Climate Policy Documents in the Context of the Paris Agreement

Prashant Pratap Singh, Erik Lehmann, Mark Tyrrell
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ)
Digital Innovation Unit - GFA Consulting
prashant.singh@giz.de
erik.lehmann@giz.de
mark.tyrrell@gfa-group.de

Abstract

Climate policy implementation is pivotal in global efforts to mitigate and adapt to climate change. In this context, this paper explores the use of Natural Language Processing (NLP) as a tool for policy advisors to efficiently track and assess climate policy and strategies, such as Nationally Determined Contributions (NDCs). These documents are essential for monitoring coherence with the Paris Agreement, yet their analysis traditionally demands significant labor and time. We demonstrate how to leverage NLP on existing climate policy databases to transform this process by structuring information extracted from these otherwise unstructured policy documents and opening avenues for a more in-depth analysis of national and regional policies. Central to our approach is the creation of a dataset 'CPo-CD' for training text classifiers, based on data provided by the International Climate Initiative (IKI) and Climate Watch (CW). The CPo-CD dataset is utilized to fine-tune pre-trained Transformer Models on classifying climate targets, actions, policies, and plans, along with their sector, mitigation-adaptation, and greenhouse gas (GHG) components. We publish our model and dataset at the GIZ Hugging Face repository (GIZ, 2024).

1 Introduction

The 2015 UN Climate Change Conference in Paris produced a landmark agreement whereby all signatories agreed to hold “the increase in the global average temperature to well below 2 °C above pre-industrial levels” (UNFCCC, 2016). The means for effecting this change are left to the countries, but each signatory is required to report progress every 5 years via nationally determined contributions (NDCs). Signatories are also encouraged to periodically communicate long-term strategies (LTS) to address climate change. The recent COP28 meeting in the UAE was the first global stocktake (GST) making use of this reporting (UNFCCC, 2023).

As the most frequent mandated reporting mechanism under the Paris Agreement, NDCs provide a consistent basis for tracking each country's progress and commitments. Consequently, analysts utilize these reports to gauge global efforts towards climate goals. Additionally, the agreement permits countries to revise their NDCs at any time (C2ES, 2017). Therefore, frequent review of these documents is important for holding signatories to account.

The Paris Agreement and follow-up COPs prescribed no standardized reporting framework. As a result, there is significant variation in the scope, format, and coverage of NDCs and LTSs over jurisdiction and reporting periods (UNEP, 2018). This variation is evident in numerous aspects of the reports, such as the articulation of mitigation contributions and the incorporation of adaptation strategies. Additionally, the documents are often not intuitively structured, and in some cases voluminous. Combined, these factors pose substantial challenges to aggregating and analyzing the data, thereby complicating the assessment of global and national efforts in addressing climate change.

Natural language processing (NLP) techniques based on deep learning have become increasingly viable for producing high-quality automated analyses in recent years, particularly with the advent of the transformer - and BERT models built upon its architecture (Vaswani et al., 2017; Devlin et al., 2019). The accessibility of pre-trained large language models via Huggingface has further lowered barriers to entry, allowing easy fine-tuning on various NLP tasks, as well as model deployment. These tools add value to analytical workflows by providing analysts with the ability to extract knowledge from unstructured data to a much higher level than was previously possible.

In this work, we seek to address the challenges with climate policy document analysis by applying sequence classification to the unstructured text in NDCs and LTSs. Our contribution consists of three main components. We first build and publish a training dataset 'CPo-CD' (Climate Policy Classification Dataset) derived from an agglomeration of two existing datasets: 1) the NDC Transport Tracker from the Advancing Transport Climate Strategies project of the International Climate Initiative (IKI TraCS)[1]; and 2) the NDC Sector Data from ClimateWatch.org[2]. We then fine-tune 2 LLMs on the dataset to classify text according to binary, multi-class and multi-label domain categories aligning with the UNFCCC hierarchical taxonomy: Targets, Actions, Policies and Plans, Mitigation / Adaptation, Sectors, Target types and Conditionality (see page 28 of Bakkegaard et al. (2015) for a breakdown of the taxonomy). Finally, we publish CPo-CD and the fine-tuned models making them accessible via a web application[3] on Huggingface - allowing analysts to upload and derive ad hoc insights from climate policy documents.

2 Related Work

The use of Natural Language Processing (NLP) in document analysis has gained significant momentum in recent years, marking a transformative shift from cumbersome manual methods of knowledge discovery using unstructured text data. In an earlier paper, Grimmer and Stewart (2013) succinctly points out the benefit in the domain of policy analysis, where NLP techniques at the time had leveled the playing field, providing independent researchers and smaller teams of analysts the ability to perform “systematic analysis of large-scale text collections without massive funding support”.

Encoder-based masked-language models trained on large text corpora have demonstrated high performance on downstream NLP tasks such as classification since Devlin et al. (2019) introduced the Bidirectional Encoder Representations from Transformers (BERT). Subsequent variations have improved on the original architecture. With RoBERTa, Zhuang et al. (2021) improved performance, resulting in higher performance across multiple NLP tasks. These base models are trained on a generalized task of next-word prediction and can be fine-tuned for a domain-specific context and downstream tasks such as the classification of targets, actions, policies, and plans.

Specific challenges to NLP tasks in the domain of policy analysis are inherent to its lexical properties - i.e. technical and domain-specific jargon. In domains with similar highly-specialized lexicons, applications have involved adapted approaches (Beltagy et al., 2019; Lee et al., 2020; Chalkidis et al., 2020. Concerning NLP application in the climate domain, the literature is surprisingly sparse. Recent work by Gonzalez et al. (2023) provides tangible evidence of this, finding that in over 76k ACL Anthology[4] NLP papers, "hardly [any papers] address other important goals such as poverty and climate", with only 50 climate-relevant NLP papers in the entire corpus since 1980 (cf. 2753 for health). Meanwhile Sietsma et al. (2024) noted 54 papers in the literature that either used or substantially discussed the use of NLP for climate adaptation specifically.

However some recent efforts are quite prominent. Concerning technical approaches, the field is quite active, with approaches using some combination of encoder/decoder architectures and pre-trained models primarily. Peña et al. (2023) present a system for multi-class classification of policy documents using RoBERTa coupled with an SVM classifier. Their results demonstrate that the combination with SVM classifiers can achieve high accuracy (over 85%) over 30 classes, even in under-represented categories.

Other recent work in the climate domain involves domain-adaptive pre-training with RoBERTa on climate-relevant corpora, before fine-tuning on downstream tasks including text classification to create ClimateBERT (Webersinke et al., 2022; Schimanski et al., 2023b). Training on a dataset of climate-related literature (i.e. news reports, news, corporate ESG disclosures, and scientific abstracts) resulted in ClimateBERT outperforming a base-DistilROBERTA model on cross-entropy loss and F1 (Webersinke et al., 2022). Building on ClimateBERT, Schimanski et al. (2023a) recently released ClimateBERT-NetZero which fine-tunes ClimateBERT to classify net zero and emissions reduction targets in corporate communications using a dataset of 3.5K expert-annotated text samples. Classification using ClimateBERT-NetZero resulted in marginally better performance than larger BERT base models.

Juhasz et al. (2024) showcases an approach

[1] https://changing-transport.org/tracker-expert/

[2] https://www.climatewatchdata.org

[3] https://huggingface.co/collections/GIZ/cpu-paper-65e7c7700dd74ca7b008a109

[4] https://aclanthology.org

for extracting mentions of net zero and other targets from national laws and policies. Building on ClimateBert and manually annotated data they fine-tune a classification model. Our work closely aligns with Schimanski et al. (2023a) and Juhasz et al. (2024) while leveraging existing policy databases to create a comprehensive training dataset and an array of classifiers corresponding to multiple UNFCCC mitigation contribution types.

3 Data

The creation of the training dataset 'CPo-CD' was the most extensive task in this project and is the main contribution alongside the models. For this reason, we describe the creation in detail below.

CPo-CD is comprised of labeled text passages extracted from policy documents (NDCs and LTS) with accompanying labels. The data is sourced originally from 2 climate policy datasets: ClimateWatch NDC Sector data[5] (CW) and IKI TraCS Climate Strategies for Transport Tracker[6] (IKI). Both datasets include text extracts from NDC/LTS documents labeled by human annotators (domain experts), as well as the accompanying climate category labels in the form of metadata. However, the labeled text from both sources is not natively useful for text classification. The length of the labeled text differs from 2 up to 250 words. While some text passages are very focused and limited to short phrases with no peripheral context and often missing information relevant to determining all climate category labels; other contain more than one item of interest but are only annotated for one. Additionally, the labeled text passages are often condensed, summarized versions of the original text and can appear multiple times in the document in varying contexts. Therefore, we identify and retrieve the original source text from the policy documents.

A sample of a short text observation is taken from the Indian NDC "75 GW by 2022" which comes with additional meta-information: sectoral mitigation policy and energy sector. From the original NDC document, we extend the text so the metadata relevant context is included: "Green Generation for Clean Energy Secure India: more than 5 times increase in Renewable Capacity from 35 GW (up to March 2015) to 175 GW by 2022. National Solar Mission scaled up five-fold from 20 GW to 100 GW by 2022. Kochi Airport is the World's first airport to fully run on solar power." This paragraph now includes a second sectoral policy information, the *upscaling of the National Solar Mission*. This missing metadata will be later added by grouping by paragraph (step 5). The whole process is described as follows:

Step 1: Text Processing: The required information is distributed across different files for both datasets. In the first step, we link the text passages with the metadata labels. For CW we utilize the Sector file from the 'NDC Content' dataset, taking the text passages and associated sector labels. We then merge with fields from the Metadata file, which allows us to add additional labels (e.g. Target, Actions, Plans, etc.) in subsequent steps. In some cases, the text has conjoined sequences between a separator character. In such cases, each unique sequence is broken out into its own sample.

The IKI data is structured much more simply - consisting of tables for categories such as Target, Netzero, Mitigation, and Adaptation. In this case, we join all tables together and retain the table name as the label.

We next combine both the CW and IKI data. A basic cleaning process is applied to the dataset, involving the removal of duplicates and erroneous samples. During Step 1, we also produce text length statistics for each country represented in the dataset. This object is used to calibrate the split strategy in Step 2 for the source text.

Step 2: Document Processing: The text extractions from both CW and IKI are narrowly focused and require expansion using the source text to make them usable for text classification. We collect the original NDC documents from the CW-associated WRI repository repo[7] in HTML format. For further documents from the IKI dataset, we source the original PDF versions of the documents from the UNFCCC website[8] using the document names provided in the IKI dataset. After downloading, the IKI pdf files are processed into raw text. Both sets of source documents are then chunked into 60, 85, and 150-word sequences which respect sentence boundaries and include an overlap to ensure the labeled text passages from the datasets are fully covered. The inclusion of multiple sequence lengths allows for greater versatility in downstream NLP tasks (in this case, text classification). The arbi-

[5] https://www.climatewatchdata.org

[6] https://changing-transport.org/tracker-expert/

[7] https://github.com/wri/ndc

[8] https://unfccc.int/sites/default/files/NDC

trary choice of the sequence lengths reflects our informed estimate of the lower and higher limits of utility, based on knowledge of the dataset.

Step 3: Secondary Label Processing and Harmonization: Various metadata accompanies the text passages for both CW and IKI that can be used to apply further labels to the text. Curating and harmonizing these metadata so that they can serve as useful labels is complex as both source datasets utilize slightly different methodologies.

In the CW dataset, we take the broad "Overview-Category" to define Adaptation and Mitigation related text. We further use a subcategory "Question-Text" to define text relating to Policies, Targets, Actions and Plans (TAPP), as well as Conditional and Unconditional commitments. A full mapping of CW QuestionText subcategories to TAPP is available on the CW website[9].

The structure of the IKI dataset is less extensive and is processed to include labels using the associated tab from the original Excel file. This includes 3 categories: Target, Adaptation-Mitigation, and Netzero. An additional 2 categories are defined from the "Parameter" subcategories within the Target spreadsheet: GHG and Conditionality. The IKI data presents a specific sectoral focus (i.e. transport) and differing nomenclature compared to Climate Watch. IKI also contains no (mitigation) Actions, nor the daughter categories of Policies and Plans, as found in CW. Therefore these labels are not represented in samples sourced from IKI.

Step 4: Context Extraction: We now perform matching of the text passages from IKI and CW with the source policy documents to build out a larger text window so that the text can be used to train a text classifier. As the text passages from CW and IKI usually only partially correspond to the original text and can appear multiple times, we retrieve the top 3 paragraphs from the processed policy documents using a BM25Okapi[10] retriever. In case of language mismatch between the text and source NDC documents (French and Spanish), translated paragraphs are used. We further use fuzzy matching of retrieved candidate passages as a quality check to ensure the relevant information is included and finalize the 'context' for each labeled text sample. This step not only ensures an accurate match for the labeled text but also provides a large number of negative samples where retrieved candidates do not match the relevant information from the CW/IKI datasets.

Step 5: Final dataset In the last step, we merge the matched text candidates with the main dataset. We then group by the final text field and remove duplicates. The dataset now contains a text field including full context, rather than short extracts -and multiple labels. Additionally, the dataset contains negative samples of unlabeled text taken from the source documents. These samples are labeled as 'None'.

The IKI dataset exhibits sub-categorization by target (GHG Target, Netzero Target, Non-GHG Target), however very few samples are available for these sub-categories. We therefore augment the dataset for these categories via manual annotation to increase positive samples and collect negative samples.

The final CPo-CD dataset contains 13,728 samples for each sequence length, split into 12,538 training and 1,190 test samples.

CPo-CD Dataset: Structure

CPo-CD is created to train classifiers (multilabel or binary), which allow policy documents to be analyzed as per the schema presented in Figure 1.

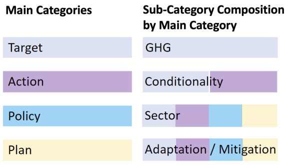

Figure 1: Classification Schema

CPo-CD Dataset: Characterisation of Label Classes

1) The first four principal categories are Target, Action, Policy, and Plan (TAPP). The data has a multilabel structure, a paragraph can entail a combination of TAPP or none of them. The training and test data for the TAPP categories in multilabel setting is presented in Table 1.

Dataset	Target	Action	Policy	Plans
Train	2,911	5,416	1,396	2,140
Test	256	513	122	198

Table 1: Number of positive samples for TAPP labels split by train and test

2) When a paragraph discusses a 'Target', it is

[9] https://wri-sites.s3.us-east-1.amazonaws.com/climatewatch.org/www.climatewatch.org/climate-watch/wri_metadata/NDC_methodology.pdf

[10] https://en.wikipedia.org/wiki/Okapi_BM25

further assessed by the 'Greenhouse Gas (GHG)' classifier to determine whether it specifies objectives relating to GHG emissions. In the CPo-CD dataset structure, a 'Non-GHG' label signifies a 'Target' relating to energy efficiency, road building, etc (in keeping with the UNFCCC taxonomy). Such labels should not be inferred as negative examples of GHG. Indeed, a paragraph can be labeled true for both categories. The number of samples for GHG targets is presented in Table 2.

Dataset	Netzero	GHG	Non-GHG
Train	120	440	259
Test	11	49	30

Table 2: Number of positive samples which include a GHG component

3) If a paragraph encompasses elements of a 'Target' or an 'Action', it requires a 'Conditionality' assessment to ascertain whether the described commitments are unconditional or dependent on external support or circumstances. Table 3 displays the number of conditionality samples.

An "unconditional contribution" refers to actions that countries can take independently, using their resources and abilities, without relying on any external conditions. On the other hand, a "conditional contribution" describes the efforts countries are willing to make if they receive international support or if certain criteria are fulfilled. Labeling conditionality is especially complex because conditional and unconditional statements often co-occur in the same paragraph, reference a group of targets and actions, or appear outside of the paragraph context.

Dataset	Conditional	Unconditional
Train	1,986	1,312
Test	192	136

Table 3: Number of positive samples with information on conditionality

3) Regardless of the TAPP category, every paragraph can be assessed to identify the economic or social sectors addressed, as well as the 'Adaptation/Mitigation' aspect. Adaptation/Mitigation discerns whether the content pertains to adaptive strategies or mitigation efforts against environmental challenges.

The sector labels encompass 16 different sectors which are distributed as follows (Train, Test): Agriculture: (2235,200); Buildings: (169,18); Coastal Zone: (698,71); Cross-Cutting Area: (1853,180); Disaster Risk Management (DRM): (814,85); Economy-wide: (873,85); Education: (180,23); Energy: (2847,254); Environment: (905,91); Health: (662,68); Industries: (419,41); LULUCF/Forestry: (1861,193); Social Development: (507,56); Tourism: (192,28); Transport: (1173,107); Urban: (558,51); Waste: (714,59); Water: (1207,106)

The number of Adaptation and Mitigation is presented in Table 4.

Dataset	Mitigation	Adaptation
Train	6,659	5,439
Test	604	533

Table 4: Number of positive samples labeled as Adaptation and Mitigation

4 Methodology

To address the challenges of efficiently analyzing voluminous and complex climate policy documents, we adopt two distinct but complementary NLP methodologies: fine-tuning a generic LLM embedding for classification tasks (Xiao et al., 2023), and further fine-tuning a pre-trained domain-specific LLM, ClimateBERT Webersinke et al. (2022).

In case of a sparsity of positive examples, we fine-tune using SetFit (Tunstall et al., 2022). SetFit represents an efficient few-shot learning framework based on Sentence Transformers (Reimers and Gurevych, 2019) which has proven to achieve high accuracy with a minimal number of samples. Its process involves first fine-tuning a Sentence Transformer embedding model on a set of labeled examples through contrastive learning. Following this, a classification head, in our case logistic regression model, was trained on these embeddings to classify new unseen data.

Our choice of the 'BAAI/bge-base-en-v1.5' (Xiao et al., 2023) - a recent 109M parameter model provided by the Beijing Academy of Artificial Intelligence - as the foundation for the generic LLM was based on its superior performance in classification tasks and ranking on the Hugging Face leaderboard.

For comparison purposes, we made use of ClimateBERT, a climate domain-specific adaptation of the DistilRoBERTa, 82.4M params (Sanh et al., 2020) transformer model. ClimateBERT was pre-trained on a large corpus of climate-related texts, imbuing it with a nuanced understanding of climate discourse. This makes it particularly suitable for classifying texts based on climate policy content. Our methodology involves further fine-tuning ClimateBERT on CPo-CD, leveraging its domain-specific pre-training to enhance classification performance. This approach is validated by its demonstrated superiority in net-zero classification tasks over larger models, including GPT-3.5-turbo, as reported by recent studies Schimanski et al. (2023a).

Given the prevalence of imbalanced classes, we chose the F1 score as the primary metric to assess model performance. The F1 score, a harmonic mean of precision and recall, provides a more comprehensive measure of a model's accuracy, especially in scenarios where class distribution is skewed. To address the inherent class imbalances within our dataset, we employed stratified sampling in the train-test split. This approach ensures that the class proportions are mirrored in the test set. Additionally, we disclose the count of test samples (support) which account for 10% of the data. Furthermore, to overcome class imbalance in a multi-label setting, we have used positive class weights in the loss function.

The paragraphs extracted from the NDCs and LTS climate policies often cover several topics, such as different climate actions or targets. Acknowledging this, we train our models in a multi-label setup, that can recognize multiple topics in one paragraph. This method is more complex than the simpler multi-class classification where one one label per paragraph is attached. The additional complexity usually results in lower performance scores. However, multi-label is a better match for this use case, ensuring we accurately capture the wide range of climate policy discussions within a single paragraph.

Carbon Emissions Monitoring

To monitor and publish the carbon emissions associated with running our models, we integrate CodeCarbon, a lightweight software tool (Schmidt et al., 2024). CodeCarbon estimates CO_2 emissions based on the electricity consumption of computing resources and the carbon intensity of the region where the computations are performed.

This transparency aligns with our commitment to environmentally responsible research, encouraging us and others in the field to consider the carbon footprint of AI and machine learning projects

5 Results

Following the described classification schema, the categorical labels 'Target', 'Action', 'Policy', and 'Plans' identify the relevant content from the policy text.

Model	Label	F1 Score	Support
bge-base-en	Target	0.84	256
ClimateBert	Target	0.81	256
bge-base-en	Action	0.85	513
ClimateBert	Action	0.82	513
bge-base-en	Policy	0.76	122
ClimateBert	Policy	0.76	122
bge-base-en	Plan	0.65	198
ClimateBert	Plan	0.63	198

Table 5: Comaparison of model performance for bge-base-en-v1.5 (BAAI) and Climate Bert fine-tuned on TAPP paragraphs

The results (ref. Table 5) show that both the generic LLM embedder (bge-base-en-v1.5) and ClimateBert models performed relatively well on the task of classifying TAPP within climate policy documents. Specifically, both models achieved their highest F1 Scores on the 'Target' and 'Action' labels, followed by 'Policy', and 'Plan'. Where the least performing class 'Plan' is the one with the fewest samples and least concrete definition. Interestingly, a classifier based on BGE Embeddings overall outperforms ClimateBert even in this data-rich scenario. In an initial comparison, ClimateBert was evaluated against a fine-tuned MPNET model, which is comparable in both age and size to BERT. In this comparison, ClimateBert demonstrated superior performance, suggesting that domain-specific adaptation does enhance performance. However, it appears that advancements in model size and technical capabilities since ClimateBert was pre-trained, may offer even greater benefits. As this pattern is repeated in the following classifications, we only report the generic fine-tuned model results specifically bge-base-en-v1.5 (LLM embedder).

Identified targets are classified for their GHG components in the next step (ref. Table 6).

Table 7 illustrates results for the conditionality classifier. The relatively poor performance reflects the challenges relevant to this category (ref. Section 7).

The sector classification results once again highlight the constraints imposed by the dataset, reveal-

Label	F1 Score	Support
GHG	0.91	49
NetZero	0.92	11
Non GHG	0.92	30

Table 6: Performance of bge-base-en-v1.5 fine-tuned using SetFit on greenhouse gas (GHG) paragraphs

Label	F1 Score	Support
Conditional	0.60	192
Unconditional	0.62	136

Table 7: Performance of bge-base-en-v1.5 fine-tuned on conditional and unconditional paragraphs

ing variable performance across different classes (ref. Table 8). Generally, a clearer distinction between classes and more definitive training data correlates with improved performance. In particular, classes such as 'cross-cutting' and 'economy-wide' proved challenging to differentiate. Despite these challenges, our evaluation reveals a commendable overall F1 score of 0.76, indicating a favorable outcome under the circumstances.

Label	F1 Score	Support
Agriculture	0.79	200
Buildings	0.65	18
Coastal Zone	0.64	71
Cross-Cutting	0.63	180
DRM	0.67	85
Economy-wide	0.48	85
Education	0.65	23
Energy	0.81	254
Environment	0.63	91
Health	0.77	68
Industries	0.74	41
LULUCF/Forestry	0.78	193
Social Develop	0.71	56
Tourism	0.60	28
Transport	0.77	107
Urban	0.48	51
Waste	0.76	59
Water	0.68	106

Table 8: Performance of bge-base-en-v1.5 fine-tuned on sectoral information

Differentiation of TAPP paragraphs between mitigation and adaptation is handled well by the classifier as illustrated by Table 9.

Label	F1 Score	Support
Mitigation	0.92	604
Adaptation	0.92	533

Table 9: Performance of bge-base-en-v1.5 model fine-tuned on mitigation and adaptation paragraphs

Model	Label	CO_2
bge-base-en-v1.5	TAPP	71.45
ClimateBert	TAPP	23.35
bge-base-en-v1.5	GHG	26.8
bge-base-en-v1.5	Conditional	28.45
bge-base-en-v1.5	Sector	58.19
bge-base-en-v1.5	Adaptation	40.45

Table 10: Comparison of CO_2 consumption in grams during the training process

Human Annotation

To assess the dataset creation process and enhance the robustness of our evaluation, we manually annotated certain paragraphs with two independent human reviewers. This provides a realistic benchmark on model performance when it comes to the analysis and classification of climate policy documents. The results are presented below (ref. Table 11).

Label	Agreement Score (%)
Target	90
Action	72
Policy	89
Plans	77
NetZero Target	98
GHG Target	96
Non GHG Target	85
Adaptation	97
Mitigation	92

Table 11: Agreement Score between two human annotators on 325 sampled paragraphs

The 'Target' category surfaced as the most consistently identified element, as evidenced by a substantial 90% concurrence among human annotators. The 'Policy' category also demonstrated notable clarity, with 89% agreement. Conversely, the 'Action' and 'Plans' categories showcased less than 80% agreement among manual annotators, revealing a relative subjectivity and interpretative flexibility within these classifications.

Carbon Emissions Results

In our analysis of model efficiency, we observe that the larger size of the BGE embedding base also results in higher emissions for fine-tuning of the TAPP classifier with 71.45 g compared to 23.35 g of CO_2 for ClimateBert. ClimateBert took 15.79 Kg of CO_2 emissions for pre-training, indicating that our fine-tuning of the ClimateBert base model for classification tasks accounts for less than 1% compared to the domain adaptation.

6 Conclusion

In conclusion, this paper explores the application of Natural Language Processing (NLP) techniques to enhance the analysis and classification of climate policy documents, with a focus on Nationally Determined Contributions (NDCs) and Long-term Strategies (LTS). We show how existing policy databases can be used to create a machine-learning-ready dataset (CPo-CD) and fine-tune pre-trained transformer models for policy analysis. We have developed a methodology that significantly streamlines the process of structuring information from these critical documents. The use of our models has been shown to markedly reduce the time required for policy analysis, enhance the effectiveness of policy examination, and enable the inclusion of a broader array of documents in the analytical process. Our approach facilitates the efficient assessment of climate targets, actions, policies, and plans (TAPP), along with their associated mitigation/adaptation, greenhouse gas (GHG), and sector components. By achieving noteworthy accuracy in TAPP, GHG, adaptation/mitigation as well as useful accuracy in sector classification, our research underscores the potential of NLP to offer meaningful insights into the alignment of international climate commitments with the Paris Agreement's objectives and support evidence-based policymaking. The release of our dataset 'CPo-CD' and model contributions marks a significant step forward towards advancing the capacity to monitor and analyze international climate commitments at scale, enhancing transparency, accountability, and informed decision-making in climate policy evaluation.

7 Limitations

Our research encountered several limitations, with the most significant challenges stemming from the nature of the original data utilized for analysis. These limitations underscore the complexities inherent to the standardization of climate policy analysis and data extraction, highlighting the need for enhanced data preparation and methodological refinement.

In addressing the limitations of our methodology, a critical point of discussion is the absence of a standardized approach to the analysis of climate policies. The heterogeneity in taxonomies and classification schemas across various databases and initiatives presents a substantial challenge. In our research, we navigated this complexity by adapting existing standards from the International Climate Initiative (IKI) and Climate Watch (CW) to establish a coherent framework for our analysis. This adaptation, while necessary for the integrity and applicability of our work, inherently limits the scalability of our methodology to other labels and databases.

The diversity in policy document formats and the varied terminologies used across different geographical and institutional contexts mean that any attempt at standardization must account for a wide range of variables. Consequently, our approach, though robust within the confines of the standards we adopted, may not seamlessly apply to analyses that rely on different sets of labels or databases. This limitation underscores a broader challenge in the field of climate policy analysis: the need for a universally accepted framework that can accommodate the nuances of global climate policy documentation. The reliance on IKI and CW standards, while enabling a structured and systematic analysis within this study, suggests that further efforts are necessary to enhance the adaptability and scalability of NLP methodologies in this domain.

Another significant challenge is the inherent complexity and subjectivity of classifying climate policy documents, as evidenced by the discrepancies in annotation. Our methodology faced limitations due to the non-distinct nature of classification categories and the variability in annotator interpretations. Even with our manual annotation benchmarking (ref. Section 5), an exact match was attained in as little as 72% of cases for some categories, highlighting the difficulties in achieving consistent and accurate data classification even for human annotators. This issue not only underscores the challenges of subjective interpretation but also signals a broader problem in harmonizing classification systems across diverse data sources.

A further limitation we encountered during the creation of machine-learning-ready training data was the fidelity of annotated context to the original source documents. The text excerpts for targets, actions, policies, and plans in existing databases varied greatly in length — from single words to multiple sentences — and were often not direct copies but rather concatenated snippets or summaries. This variance presented significant challenges in the matching process to the original context. The statistical matching introduced potential sources of error. To ensure robustness, we decided on a high matching threshold, which resulted in a substantial loss of samples. Even still, some areas of the training data potentially suffer from quality issues. Consequently, although the large existing databases represented a valuable resource, we were only able to partially utilize them for CPo-CD. This experience underscores the need for - and potential benefits of - incorporating standardized criteria, with a focus on automation, into the dataset creation process.

A notable limitation of our approach is its focus on English-language documents and specific types, primarily NDCs and LTS. This restricts our analysis to a narrow linguistic range and does not yet cover the diversity of global climate policies documented in other languages. Additionally, by concentrating on NDCs and LTS, we miss out on evaluating the performance of our model on other crucial document types like local policies and laws, which play a significant role in the practical implementation of climate strategies.

Expanding our models to include multilingual capabilities and a broader spectrum of document types would enhance its utility, allowing for a more comprehensive analysis of global climate actions. Such improvements would offer an even more detailed understanding of international efforts to address climate change, though this expansion remains a notable rather than a critical limitation in our current research scope.

The classification of conditionality within climate policy documents proved to be a complex task that our current model and the provided context struggled to adequately address. This complexity arises from the nuanced nature of conditionality clauses, which require a deep understanding of the text to accurately classify. Generative language models with advanced reasoning capabilities over larger context windows could potentially offer improved performance in this area leveraging recent work from Thulke et al. (2024) with the trade-off of higher costs.

Acknowledgements

We extend our sincerest gratitude to the International Climate Initiative (IKI), specifically the TraCS-Project for their provision of data and their support throughout the evaluation process. Their collaboration not only facilitated our research but also enriched it by providing a pertinent use case that significantly contributed to the depth and relevance of our analysis. Similarly, we are thankful to Climate Watch (CW) for their data provision, which played a foundational role in the development of our dataset and the subsequent insights derived from our study. The contributions from both IKI and CW have been instrumental in advancing our understanding of climate policy implementation and evaluation, and we are profoundly appreciative of their support and collaboration.

References

Riyong Kim Bakkegaard, Skylar Bee, Prakriti Naswa, Todd Ngara, Anne Olhoff, Sudhir Sharma, and Denis DR Desgain. 2015. Developing INDCs: a guidance note. Report, UNEP DTU Partnership, Copenhagen. Publication Title: Developing INDCs: a guidance note.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Center for Climate and Energy Solutions C2ES. 2017. Legal Issues Related to the Paris Agreement.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH GIZ. 2024. CPU-Paper - a GIZ Collection.

Fernando Gonzalez, Zhijing Jin, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2023. Beyond Good Intentions: Reporting the Research Landscape of NLP for Social Good. ArXiv:2305.05471 [cs].

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297. Publisher: Cambridge University Press.

Matyas Juhasz, Tina Marchand, Roshan Melwani, Kalyan Dutia, Sarah Goodenough, Harrison Pim, and Henry Franks. 2024. Identifying Climate Targets in National Laws and Policies using Machine Learning. ArXiv:2404.02822 [cs] version: 2.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36(4):1234–1240.

Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Íñigo Puente, Jorge Córdova, and Gonzalo Córdova. 2023. Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs. In *Document Analysis and Recognition – ICDAR 2023 Workshops: San José, CA, USA, August 24–26, 2023, Proceedings, Part I*, pages 20–33, Berlin, Heidelberg. Springer-Verlag.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023a. ClimateBERT-NetZero: Detecting and Assessing Net Zero and Reduction Targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756, Singapore. Association for Computational Linguistics.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023b. Bridging the Gap in ESG Measurement: Using NLP to Quantify Environmental, Social, and Governance Communication.

Victor Schmidt, Thomas Bouvier, Marion Coutarel-Huez, and Khalil Chaouali. 2024. CodeCarbon.io.

Anne J. Sietsma, James D. Ford, and Jan C. Minx. 2024. The next generation of machine learning for tracking adaptation texts. *Nature Climate Change*, 14(1):31–39. Number: 1 Publisher: Nature Publishing Group.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change. ArXiv:2401.09646 [cs].

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. ArXiv:2209.11055 [cs].

UNEP. 2018. Pocket Guide To NDCs under the UNFCCC.

UNFCCC. 2023. COP28 UAE - United Nations Climate Change Conference.

Secretariat UNFCCC. 2016. Report of the Conference of the Parties on its twenty-first session, held in Paris from 30 November to 11 December 2015. Addendum. Part two: Action taken by the Conference of the Parties at its twenty-first session.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. ArXiv:2309.07597 [cs].

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.