BERT Classification of Paris Agreement Climate Action Plans

Tom Corringham [1] Daniel Spokoyny [2] Eric Xiao [3] Christopher Cha [3] Colin Lemarchand [3] Mandeep Syal [3]
Ethan Olson [3] Alexander Gershunov [1]

Abstract

As the volume of text-based information on climate policy increases, natural language processing (NLP) tools can distill information from text to better inform decision making on climate policy. We investigate how large pretrained transformers based on the BERT architecture classify sentences on a data set of climate action plans which countries submitted to the United Nations following the 2015 Paris Agreement. We use the document header structure to assign noisy policy-relevant labels such as mitigation, adaptation, energy, and land use to text elements. Our models provide an improvement in out-of-sample classification over simple heuristics though fall short of the consistency observed between human annotators. We hope to extend this framework to a wider class of textual climate change data such as climate legislation and corporate social responsibility filings and build tools to streamline the extraction of information from these documents for climate change researchers.

1. Introduction

The United Nations Framework Convention on Climate Change (UNFCCC) is a global framework for addressing the challenges of anthropogenic climate change (Leggett, 2020). Under the 2015 Paris Agreement each UNFCCC signatory agreed to submit a Nationally Determined Contribution (NDC) upon ratification of the agreement by the country's national government (UNFCCC, 2016). These climate action plans set objectives and timelines for reductions in greenhouse gas emissions for each country. The documents, while often aspirational in nature, provide useful information about the challenges facing each country, their stance towards climate change, and their ambitions regarding mitigation efforts (Tribett et al., 2017). Here, a deep learning model is applied to 165 of these documents to build a sentence classifier which could be used to generate policy-relevant metrics over a wide range of documents.

The volume of information contained in climate policy documents is growing rapidly. Every year large stakeholders such as governments and corporations produce text-based climate assessments and action plans to communicate and satisfy regulatory requirements. NLP and machine learning can be used to provide data for climate policy analysis, improve tools for evaluating policy, and provide new tools for policy assessment (Rolnick et al., 2019). Supervised and unsupervised NLP content analysis methods have been used to analyze political texts (Grimmer & Stewart, 2013) including climate negotiation texts (Venturini et al., 2014; Ruiz et al., 2016; Baya-Laffite & Cointet, 2016), climate adaptation analyses (Biesbroek et al., 2020), and corporate climate financial disclosures (Luccioni & Palacios, 2019).

BERT (Devlin et al., 2019) is a bidirectional transformer model that has been pretrained on a large corpus of textual data using the masked language modeling objective. BERT and other pretrained transformer models through finetuning have achieved state-of-the-art results in a variety of NLP tasks (Rogers et al., 2020) including sentence classification. This is makes them good candidates for climate change text applications where large labeled data sets are currently unavailable.

Recently, the BERT model has been used to extract information from climate-related regulatory disclosures. Varini et al. (2021) use BERT to classify sentences from U.S. Securities and Exchange Commission (SEC) filings as climate related or not climate related. Kölbel et al. (2020) apply BERT to SEC filings to distinguish between sentences that discuss physical climate risk (e.g., due to sea level rise or extreme weather events) and transition risk (due to expected changes in climate-related regulation). Using the classified text they develop metrics that they relate to credit default swap rates. Bingler et al. (2021) use BERT to classify sentences and paragraphs from corporate risk disclosure documents into four categories to assess the impact of the Task Force on Climate-related Financial Disclosures (TCFD).

[1] Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA [2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA [3] University of California San Diego, La Jolla, California, USA. Correspondence to: Tom Corringham <tcorringham@ucsd.edu>.

Tackling Climate Change with Machine Learning workshop at ICML 2021, Copyright 2021 by the authors.

Here, BERT is applied in a similar way to classify sentences in national climate action plans. Metrics derived from BERT-classified climate documents could be used to investigate the relationships between document characteristics and country characteristics such as exposure to climate risks or energy and resource endowments. Understanding these relationships could be used to evaluate, monitor, and improve global climate policy.

2. Data and Labeling

165 English-language NDCs and Intended NDCs (UNFCCC, 2021) were obtained in HTML format from Climate Watch (Climate Watch, 2019). Paragraph, list, and table elements were extracted from these HTML documents. The text elements typically contain multiple sentences, sentence fragments, and in the case of tables, numeric data. Numeric data were removed from the tables, and text elements were sentencized (Honnibal et al., 2020). Sentences under 10 words in length were discarded to remove less useful sentence fragments. This process generated 25,500 unique sentences. Document length ranged from 18 to 482 sentences with a mean of 154 and a median of 130 sentences. Manual classification of sentences into topic groups was not feasible. Instead, "weak" labels were generated for each sentence by exploiting the nested headers, subheaders, and table structures within the HTML documents.

The full set of lemmatized words found in the HTML headers and table row names were manually divided into 11 topic areas by human climate policy experts. For example, terms such as "deforestation" or "LULUCF" (land use, land-use change, and forestry) were assigned to the Land Use topic. The topic identified with the most deeply nested header was assigned to all sentences within that text element. In cases where multiple topic words appeared in a given header, the less frequent topic was assigned (e.g., if both Land Use and Mitigation keywords were present then the less frequent Land Use label was applied). In some cases, no topic was assigned in which case the sentence was labeled as "No Label." The distribution of topics was not uniform. Some topics, such as Mitigation or Adaptation appear more frequently than others such as Industry or Environment (Table 1). These reference labels are referred to as weak labels to emphasize that they are noisy and often do not correspond to topic labels that would be assigned by human annotators.

Table 1. Frequency of weak labels over NDC sentences.

Weak Label	Frequency (%)
No Label	16.3
Adaptation	15.0
Agriculture	4.7
Economic	4.5
Energy	5.0
Environment	3.0
Equity	7.1
Industry	2.0
Land Use	3.4
Mitigation	16.0
Strategy	21.7
Waste	1.2

3. Model Framework

The weakly labeled sentences were split into training, validation, and test sets comprising 80, 10, and 10 percent of the sentences, respectively. The transformer models were iteratively optimized on the training data. At each epoch of the training process, model loss was calculated using the validation data. If the validation loss increased for three epochs in a row the training process was halted and the model with the lowest validation loss was chosen. The final model was then applied to the hold-out test set of sentences to evaluate model skill.

Two uncased transformer models were trained and tested against the data: BERT$_{\text{BASE}}$ and SciBERT (Beltagy et al., 2019). SciBERT is a BERT model pretrained on a large corpus of scientific publications which has been shown to provide improvements on standard NLP tasks on data sets from scientific domains. Examples of the training, validation, and test set skill scores are shown for the BERT model in Table 2.

Table 2. BERT training skill scores.

Data	Accuracy	Precision	Recall	F_1
Train	0.907	0.692	0.673	0.669
Validate	0.839	0.450	0.436	0.429
Test	0.847	0.417	0.406	0.397

The BERT and SciBERT model performances were compared to three benchmark classifiers. The null Random classifier assigned labels randomly with equal frequencies. The Majority classifier assigned the most common topic, Strategy, to all sentences. The Contains classifier applied a simple heuristic: if any of the topic words associated with a topic label appeared in a sentence then the sentence received that topic label. As with the weak reference labels, if multiple topic words appeared within the same sentence the lowest frequency topic label was assigned. The reasoning is that lower frequency labels have greater specificity and are likely to capture more salient content.

Finally, a balanced 600-sentence subset of the test set (50 sentences with each weak label) was manually labeled by two student annotators. The human labels were evaluated relative to the weak labels to provide an upper bound to the NLP classification skill. The two sets of human labels were compared to quantify inter-annotator agreement.

4. Results

4.1. Model Evaluation

Table 3 presents weighted skill metrics for each of the classifiers. The Random and Majority classifiers perform poorly with weighted F_1 scores of 0.09 and 0.07, respectively. The Contains heuristic shows some improvement over the null classifiers with F_1 of 0.17. BERT outperforms these classifiers with F_1 of 0.40. SciBERT is marginally less accurate than BERT and has a lower F_1 score, perhaps indicating that the policy documents are more similar to general text corpora than to collections of scientific documents.

Table 3. Model performance.

Classifier	Accuracy	Precision	Recall	F_1
Random	0.813	0.117	0.081	0.089
Majority	0.757	0.041	0.203	0.069
Contains	0.829	0.229	0.170	0.171
SciBERT	0.843	0.398	0.379	0.362
BERT	0.847	0.417	0.406	0.397
Human*	0.867	0.281	0.250	0.251

* The Human metrics are calculated on a 600-sentence subset of the hold-out test set.

To put the BERT F_1 score in context, the Contains and BERT predicted labels were tested against the human labels (Table 4) on the balanced 600-sentence subset of the test set. One of the human annotators, referred to here as Student, was a student researcher with no knowledge of climate policy who was simply directed to label sentences using their best judgment. The other human annotator, the Expert, was a student with climate policy research experience and familiarity with the NDC documents. In these results "Human" skill scores are averages over both annotator scores.

Table 4. Comparison of Contains and BERT to human annotators.

Weighted F_1		Reference Label		
		Human	Student	Expert
Classifier	Contains	0.350	0.399	0.302
	BERT	0.301	0.284	0.317
	Student			0.472

On average, the simple Contains heuristic shows better agreement with the human annotators' labels than the BERT classifier. This is not surprising, given that BERT was optimized to predict the weak labels which provide a very noisy representation of semantic content. Ideally BERT would be trained on human-annotated data, but in many applications such data sets are expensive to generate.

Interestingly, the Contains heuristic only outperforms BERT trained on weak labels when compared with the Student labels. When compared to the Expert labels BERT slightly outperforms Contains although the difference in F_1 scores is not significant (using bootstrapped 95% confidence intervals). It may be that the Contains heuristic more closely mimics an untrained annotator while BERT is better able to emulate expert-level context-sensitive classification. More annotated data would be required to explore this possibility.

4.2. Error Analysis

An illustrative set of test sentences (edited here for concision) are presented in Table 5 with their weak reference labels and the Contains, BERT, Student and Expert predicted labels. In the first sentence the classifiers agree: the keywords "emission" and "mitigation" both indicate that the sentence concerns Mitigation. The second sentence is correctly labeled Strategy by BERT and the human annotators, but not by Contains, i.e., BERT outperforms Contains. The third sentence contains keywords from different topics ("sequestration" indicates the Mitigation topic; "afforestation" indicates Land Use). Here Contains matches the weak label. BERT predicts Agriculture which is semantically similar to Land Use suggesting potential improvements to the classification algorithm. The Student seizes upon the first keyword "mitigation" as a label, demonstrating a potential weakness of manual annotation. The fourth sentence predictions are confused although BERT has matched the Expert annotator's label. Finally, the last sentence is not related to climate change but instead provides background information on a country's recent history. Relative to manual annotation the weak reference label seems inappropriate; in this case the sentence has fallen under an HTML section header that indicates the Environment topic. It is not surprising that none of the classifiers match the weak label.

5. Discussion and Future Work

Using weak topic labels derived from the document header structure as reference labels is clearly inferior to a system in which a large number of training, validation, and test sentences are manually annotated by climate policy experts. However, manual annotation is often infeasible for large corpora. While BERT outperforms simpler methods and even the human annotators in predicting the weak reference labels, the simpler Contains classifier provides better agreement with the human annotators. The difference is not as pronounced when the annotator with more domain-specific expertise is used as the reference, but the results underscore the importance of clean reference data for training deep learning models.

Table 5. Error analysis.

SENTENCE	LABEL	CONTAINS	BERT	STUDENT	EXPERT
It is envisaged that **emission** reduction will be achieved through the **mitigation** actions in the sectors.	MITIGATION	MITIGATION	MITIGATION	MITIGATION	MITIGATION
The Steering Committee is the supreme body for **decision making** and sectoral **implementation**.	STRATEGY	MITIGATION	STRATEGY	STRATEGY	STRATEGY
The **mitigation** actions that enhance **afforestation** are projected to result in the **sequestration** of 1 mtCO2e annually.	LAND USE	LAND USE	AGRICULTURE	MITIGATION	LAND USE
In the absence of project activity, **fossil fuels** could be burned in **power plants** that are connected to the grid.	STRATEGY	EQUITY	ENERGY	INDUSTRY	ENERGY
Due to the outbreak of the Ebola Virus the development gains made after a 10-year civil war were rudely reversed.	ENVIRONMENT	NO LABEL	MITIGATION	NO LABEL	NO LABEL

The development of this framework suggests several areas for improvement. First, sentences could be classified with multiple labels. For example, the phrase “The mitigation actions that enhance afforestation are projected to result in the sequestration of 1 mtCO2e annually” could be classified as both a Mitigation sentence and a Land Use sentence.

To generate predicted topic labels, BERT takes the argmax of a vector of weights derived from the final hidden layer of the neural network. These weights can be normalized to provide sentence topic probabilities. Such soft labels may generate more meaningful document-level climate policy metrics than hard single-topic sentence labels. Similarly, simple bag-of-words classifiers often yield multiple predictions when conflicting keywords appear in the same sentence. In this study, the two human annotators were instructed to list all relevant topics and then to choose the topic they felt was most relevant. The next step in this research will be to evaluate these multi-label classifiers against weak multi-labels and against each other.

A more complex multi-labeling approach could account for hierarchies within the set of topics. For example, there are energy strategies that fall under the framework of mitigation (e.g., transition to renewables) and energy strategies that fall under the framework of adaptation (e.g., protection of nuclear power plants from sea level rise). Classification with hierarchical topic labels could further improve metrics for policy analysis.

The selection of topics and of topic words using text from the HTML headers was performed by two trained climate researchers. Manual classification is inherently subjective, and compromises were required between the experts to obtain a reasonable classification scheme. Automated approaches such as those discussed in Lucioni & Palacios (2019) or Kölbel et al. (2020) may offer some improvement. Furthermore, the number of topics selected in this analysis was limited to 11 plus the null label for ease of computation and to avoid problems arising from sparse labels. Initially over 25 topics were proposed on the basis of the header words. If possible, it would be interesting to extend the analysis to consider a much wider range of topics, including specialized topics such as indigenous community involvement (David-Chavez & Gavin, 2018) or the impacts of climate change on coastal communities and marine ecosystems (Gallo et al., 2017).

6. Conclusion

Under the Paris Agreement, signatories are expected to submit updated NDCs every five years. As of May 2021, eight countries have submitted their second NDC (UNFCCC, 2021) though more plans are expected once the COVID-19 pandemic is brought under control. In the U.S., 33 states have released climate action plans (C2ES, 2020). Globally, 28 cities in the C40 Cities Climate Leadership Group have published Paris Agreement compatible climate action plans (C40, 2021). The continued development of state-of-the-art NLP tools tailored to climate policy will allow climate researchers and policy makers to extract meaningful information from this growing body of text, to monitor trends over time and administrative units, and to identify potential policy improvements.

References

Baya-Laffite, N. and Cointet, J.-P. Mapping Topics in International Climate Negotiations: A Computer-Assisted Semantic Network Approach. In Kubitschko, S. and Kaun, A. (eds.), *Innovative Methods in Media and Communication Research*, pp. 273–291. Springer International Publishing, Cham, 2016. ISBN 978-3-319-40699-2 978-3-319-40700-5. doi: 10.1007/978-3-319-40700-5_14. URL http://link.springer.com/10.1007/978-3-319-40700-5_14.

Beltagy, I., Lo, K., and Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv:1903.10676 [cs]*, September 2019. URL http://arxiv.org/abs/1903.10676. arXiv: 1903.10676.

Biesbroek, R., Badloe, S., and Athanasiadis, I. N. Machine learning for research on climate change adaptation policy integration: an exploratory UK case study. *Regional Environmental Change*, 20 (3):85, September 2020. ISSN 1436-3798, 1436-378X. doi: 10.1007/s10113-020-01677-8. URL http://link.springer.com/10.1007/s10113-020-01677-8.

Bingler, J. A., Kraus, M., and Leippold, M. Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures. *SSRN Electronic Journal*, 2021. ISSN 1556-5068. doi: 10.2139/ssrn.3796152. URL https://www.ssrn.com/abstract=3796152.

C2ES. U.S. State Climate Action Plans, Center for Climate and Energy Solutions, 2020. URL https://www.c2es.org/document/climate-action-plans/.

C40. C40 Climate Action Planning Resource Centre, 2021. URL https://resourcecentre.c40.org/.

Climate Watch. Climate Watch - Data for Climate Action, 2019. URL https://www.climatewatchdata.org/.

David-Chavez, D. M. and Gavin, M. C. A global assessment of Indigenous community engagement in climate research. *Environmental Research Letters*, 13(12):123005, December 2018. ISSN 1748-9326. doi: 10.1088/1748-9326/aaf300. URL https://iopscience.iop.org/article/10.1088/1748-9326/aaf300.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

Gallo, N. D., Victor, D. G., and Levin, L. A. Ocean commitments under the Paris Agreement. *Nature Climate Change*, 7(11):833–838, November 2017. ISSN 1758-678X, 1758-6798. doi: 10.1038/nclimate3422. URL http://www.nature.com/articles/nclimate3422.

Grimmer, J. and Stewart, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mps028. URL https://www.cambridge.org/core/product/identifier/S1047198700013401/type/journal_article.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303.

Kölbel, J., Leippold, M., Rillaerts, J., and Wang, Q. Does the CDS Market Reflect Regulatory Climate Risk Disclosures? *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3616324. URL https://www.ssrn.com/abstract=3616324.

Leggett, J. A. The United Nations Framework Convention on Climate Change, the Kyoto Protocol, and the Paris Agreement: A Summary. Technical Report R46204, Congressional Research Service, 2020. URL https://fas.org/sgp/crs/misc/R46204.pdf.

Luccioni, A. and Palacios, H. Using Natural Language Processing to Analyze Financial Climate Disclosures, 2019. URL https://www.climatechange.ai/papers/icml2019/34/paper.pdf.

Rogers, A., Kovaleva, O., and Rumshisky, A. A Primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs]*, November 2020. URL http://arxiv.org/abs/2002.12327. arXiv: 2002.12327.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y. Tackling Climate Change with Machine Learning. *arXiv:1906.05433 [cs, stat]*, November 2019. URL http://arxiv.org/abs/1906.05433. arXiv: 1906.05433.

Ruiz, P., Plancq, C., and Poibeau, T. Climate Negotiation Analysis. Technical Report hal-01423299, Jagiellonian University and Pedagogical University, Cracovie, Poland, 2016. URL https://hal.archives-ouvertes.fr/hal-01423299.

Tribett, W. R., Salawitch, R. J., Hope, A. P., Canty, T. P., and Bennett, B. F. Paris INDCs. In *Paris Climate Agreement: Beacon of Hope*, pp. 115–146. Springer International Publishing, Cham, 2017. ISBN 978-3-319-46938-6 978-3-319-46939-3. doi: 10.1007/978-3-319-46939-3_3. URL http://link.springer.com/10.1007/978-3-319-46939-3_3. Series Title: Springer Climate.

UNFCCC. Synthesis report on the aggregate effect of the intended nationally determined contributions. Technical Report FCCC/CP/2016/2, UNFCCC, 2016. URL https://unfccc.int/resource/docs/2016/cop22/eng/02.pdf.

UNFCCC. NDC Registry (interim), 2021. URL https://www4.unfccc.int/sites/NDCStaging/Pages/All.aspx.

Varini, F. S., Boyd-Graber, J., Ciaramita, M., and Leippold, M. ClimaText: A Dataset for Climate Change Topic Detection. *arXiv:2012.00483 [cs]*, January 2021. URL http://arxiv.org/abs/2012.00483. arXiv: 2012.00483.

Venturini, T., Baya Laffite, N., Cointet, J.-P., Gray, I., Zabban, V., and De Pryck, K. Three maps and three misunderstandings: A digital mapping of climate diplomacy. *Big Data & Society*, 1(2): 205395171454380, July 2014. ISSN 2053-9517, 2053-9517. doi: 10.1177/2053951714543804. URL http://journals.sagepub.com/doi/10.1177/2053951714543804.