Meral Basit, ZiJing Xu
December 8, 2024
{ meral, zijingxu } @ischool.berkeley.edu

Warming Up to NLP: Fine-Tuning Classification of Climate Change Stances and Political Messaging

1 Abstract

Climate change sentiment detection represents an underexplored challenge in natural language processing, addressed in this study through a novel methodology employing ClimateBERT and T5 models on the Global Warming Stance Detection Dataset (GWSD). By focusing on classifying sentences into agreement, disagreement, and neutral stances regarding climate change, and exploring the generation of ideologically-coded text, the research aims to provide computational insights into partisan environmental discourse. Despite acknowledging limitations such as potential data bias and class imbalance, the study advances understanding of climate communication by leveraging machine learning to parse complex, politically nuanced textual representations, with success measured through classification accuracy and the model's ability to generate and correctly classify partisan-aligned sentences.

2 Introduction

While scientists largely agree that global warming is predominantly the result of human activities, roughly one-third of Americans believe it stems from natural changes. [1] [2] As climate change is one of the most pressing yet divisive issues today, detecting sentiment in online media can be imperative for environmental campaign strategy, making tools like natural language processing (NLP) incredibly useful.

Unfortunately, climate change has received comparatively little attention in the NLP space. [3] This is a problem because climate change topics are historically harder to detect by models, and researchers often require multiple types of models to correctly detect all instances of climate-related topics. [4]

Our goal is to develop a model that will be able to classify sentiment toward climate change at the sentence level. This would be useful when analyzing trends in news, blog posts, and social media posts.

3 Background

3.1 Data

The **Global Warming Stance Dataset (GWSD)** was developed to facilitate the detection and analysis of positions in the global warming debate within media texts by a team of researchers at Stanford University (hereby referred to as the ”original research team”). [5] The dataset was created from 56,000 news articles published by 63 U.S. outlets between 2000 and 2020. A subset of 2,050 sentences, extracted using dependency parsing to identify opinion spans, which consists of our main data source, was annotated through Amazon Mechanical Turk (MTurk)[1]. Annotators rated each sentence as agreeing, disagreeing, or neutral toward the stance “Global warming is a serious concern.” The final dataset contains a set of aggregated probabilities indicating the likelihood of the sentences' opinions on agreeing, disagreeing, or neutral towards the above stance. GWSD's comprehensive annotation and careful demographic consideration make it a crucial resource for exploring opinion-framing and stance detection, which we have taken advantage of when applying political leanings and enhanced classification of climate change stances in our own research.

The dataset has a a class imbalance, where labels are: 45% *agree*, 38% *neutral*, and 17% *disagree*. The dataset also came with a designated test set, which we used in the three class approach (see below) so our results would be comparable to the original paper. For the seven class approach (see below),

[1] Amazon Mechanical Turk: https://www.mturk.com/

due to data transformations, we created random splits of test, train, and validation sets.

3.2 Problem Approach

We employed two (2) umbrella steps of enhanced classification:

- We first worked with the three existing classes from the GWSD dataset {agree, disagree, neutral} to be able to compare our findings with those of the original research group. Through hyperparameter fine-tuning, we aimed to create ClimateBERT and BERT models to compare the differences in classification results.
- We then utilized a seven-class approach which captured degrees of political leaning {very left, solidly left, somewhat left, center, somewhat right, solidly right, very right} to classify sentences into possible political leanings. We wanted to expand the number of categories in the hopes of eventually making a text generator that could give answers with more granular leanings.

We will explore our category creation methodology further in the next section.

Previous work on this dataset has been done using BERT. [5] We decided to use ClimateBERT, a language model that has been further trained on climate-specific texts and has out-competed BERT models in this domain [6]. Our working hypothesis is that ClimateBERT would have higher classification accuracy and model F1 score than previous BERT models, due to ClimateBERT's domain specificity with climate-related language.

We then employed a text generation model to create machine-generated sentences with GWSD-established political leaning categories. By generating and classifying synthetic sentences, we aimed to explore how the machine responds to certain ideological framings and languages that characterize partisan discussions of climate change, providing insight into how political attitudes might manifest in media communication.

4 Methods

4.1 Three Class Model

We began by establishing a majority class classifier as a baseline model. Next, we worked on our BERT and ClimateBERT classifiers, each paired with a neural net on top. Both models were made fully trainable. Because we wanted to compare our final models, we put equal effort into hyperparameter tuning both models. We applied all the steps outlined below to each model and selected the optimal set of parameters for each.

Initially, we conducted a grid search on dropout rates, learning rates, and the choice between using the pooler layer or the CLS token. After extensively evaluating the effects of each hyperparameter change, we selected the two most promising sets and tested them with three different random seeds to determine the optimal configuration.[2]

After this, we experimented with the number of epochs using Keras's `Early Stopping` callback.[3] After a tunable waiting period, `Early Stopping` checks end of every epoch to determine if the loss is still decreasing. Once that is no longer true, `Early Stopping` will stop training. After using this approach, we found the number of epochs that reliably gave us the highest validation accuracy while not over fitting.

Finally, we saw that the class imbalances still seemed to be affecting the validation class accuracies. To address this, we implemented class weights, as we saw that the original research team also mentioned that their implementation made a large difference in their results. [5] After implementing class weights that were inversely

[2] See Grid Search analysis Jupyter Notebook for full analysis on validation data.

[3] K. Team, "Keras documentation: EarlyStopping," *keras.io*. https://keras.io/api/callbacks/early_stopping/

proportional to the number of examples in that class, we found that *disagree* was still being underpredicted, so we manually increased that weight.

4.2 Seven Class Model – Political Leaning Classification

4.2.1 Creation of Political Leaning Categories

In the GWSD dataset, political leanings are inferred through a sophisticated Bayesian hierarchical ordinal regression model that analyzes sentence stances. Each sentence is annotated by multiple participants who classify its stance toward ″Global warming is a serious concern″ as *agree*, *disagree*, or *neutral*, with the model accounting for annotator demographic characteristics and sentence-specific variations.

We further leverage the stances as a proxy for political leaning categories. The final stance classification is determined by the highest probability distribution, with stance labels mapped to political leanings: ″*Neutral*″ becomes ″center″, ″*Disagree*″ becomes ″right-leaning″ (i.e. the sentences disagrees with global warming's being a serious concern), and ″*Agree*″ becomes ″left-leaning″. Here, we assume that right-wing media tends to be more global warming denying while left-wing media tends to be more global warming accepting. The political leaning intensity is further nuanced by probability thresholds: below 0.50 indicates ″somewhat″ leaning, between 0.50-0.75 signifies ″solidly″ leaning, and above 0.75 represents ″very″ left- or right-leaning. This methodology enables a probabilistic approach to capturing the complex, context-dependent nature of political stance in discussions about global warming.

4.2.2 Modeling

Due to severe class imbalances, we needed to address training / validation / testing splitting such that it created somewhat proportional representations of each class. The distribution of classes is shown below:

Table 1: Political Class Distributions

Class	Size (n)
center	656
very left	611
very right	306
solidly left	157
somewhat left	149
somewhat right	89
solidly right	82

We initially leveraged a combined strategy of stratified sampling and class weights to balance out class representation. By implementing a proportional sampling strategy that constrains group sizes between 50 and 500 instances while maintaining original class distributions, the technique ensures representative data selection. Concurrently, class weights are computed inversely proportional to class frequencies to compensate for imbalanced datasets during model training [7]. Our approach hopes to preserve the underlying statistical characteristics of the data, thereby mitigating potential bias and improving the model's ability to learn from minority classes. We have also weighted Keras's `SparseCategoricalCrossentropy` to take class weights into consideration within the loss function. [8]

We built a baseline model using classic BERT[4] to generate baseline classification accuracy and F1 score, along with a majority class baseline model. We then employed fine-tuning via addressing more complex concerns of class imbalance using 5 differing classification models:

- ClimateBERT, with stratified sampling only
- ClimateBERT, with stratified sampling and focal loss [9]
- ClimateBERT, with stratified sampling and Synthetic Minority Oversampling Technique

[4] `bert-base-closed|`

(SMOTE) [10]

- ClimateBERT, with stratified sampling and early stopping of epochs [11], and learning rate reduction [12]
- ClimateBERT, with stratified sampling and early stopping of epochs and learning rate reduction and SMOTE

We then employed a T5 model[5] to machine generate texts using the original GWSD with our labeled political leaning categories as training data. We then computed the pairwise cosine similarities of each political leaning class between the generated dataset and the GWSD dataset.

5 Results and Discussion

5.1 Three Class Model

Table 2: BERT Classification Report

Class	**Precision**	**Recall**	**F1-Score**
Agree	0.81	0.38	0.52
Disagree	0.47	0.87	0.61
Neutral	0.66	0.72	0.69
Accuracy: 0.62			
Macro avg	0.65	0.66	0.61
Weighted avg	0.68	0.62	0.61

Table 3: ClimateBERT Classification Report

Class	**Precision**	**Recall**	**F1-Score**
Agree	0.82	0.67	0.74
Disagree	0.71	0.62	0.66
Neutral	0.66	0.81	0.73
Accuracy: 0.72			
Macro avg	0.73	0.70	0.71
Weighted avg	0.73	0.72	0.72

Both ClimateBERT and BERT performed much better than the Baseline majority class classifier, which had a 0.32 accuracy, 0.18 macro F1 score, and 0.21 weighted F1 score.

ClimateBERT had a 10% higher accuracy than BERT. ClimateBERT also has smaller differences between macro and weighted average values for precision and recall than BERT, suggesting that ClimateBERT is more consistent across classes.

The original research team was able to get their BERT model to 75% accuracy. [5] The improvement we observed between BERT and ClimateBERT suggests that using ClimateBERT could have further boosted their accuracy, even though we did not surpass it.

Interestingly, BERT had a higher recall than ClimateBERT for the minority class, *disagree*. However, BERT also had a much lower precision than ClimateBERT, which suggests that BERT is over predicting the minority class. This may be due to the difference in class weights between the models, but the difference in weights was fairly minor.[6]

Of the 9 *disagree* sentences correctly classified by BERT but not by ClimateBERT, ClimateBERT misclassified 8 as *neutral* and 1 as *agree*. We took a closer look at the *agree* instance, and found that the sentence was using sarcasm.[7] BERT correctly labeling that sentence could have been due to it detecting sarcasm better, or due to it simply overlabeling *disagree*. However, ClimateBERT was further pretrained on ”common news, research articles, and climate reporting of companies” [6] and so it may not be as adept at classifying common language sentiment. We will revisit this in the conclusion.

Notably, when the models agreed on the wrong answer, we found that they most often predicted *agree* sentences as *neutral*. This may be due to a class weight overcorrection. However, we also observed that these sentences were shorter than average (94.5 versus 115.6 characters) and often stating facts as opposed to opinions. When examining these sen-

[5] `t5-base|`

[6] BERT: {Neutral: 0.90, Disagree: 2.01, Agree: 0.73} ClimateBERT: {Neutral: 0.90, Disagree: 2.05, Agree: 0.73}

[7] Sentence: ”The fifteen-year long “ global warming ” campaign all along meant “ climate change ” and that this in turn means that places supposed to get hotter get hotter and that places that are supposed to get colder — under global warming, er, climate change — get colder.”

tences by eye, it was somewhat difficult to discern if they were *neutral* or *agree*, which suggests that imperfect labeling may also be contributing to these errors.[8]

5.2 Seven Class Model

The majority class baseline has an accuracy of 0.06, which vastly underperformed all machine learning-based classification.

While political leaning classification could not get near the original research team's BERT model accuracy of 0.75, we found that our results outperformed random classification: in 4 out of 5 ClimateBERT models, accuracy was above 0.50. However, we note that our Focal Loss model performed exceptionally poorly compared to others. One theory that we have is that stratified sampling actually largely reduced the issue of the severely imbalanced class proportions, which focal loss best performs in. [9] Additionally, upon further research, we have learned that multiclass classification problems are better solved with cross-entropy loss, as focal loss is better tuned for object detection [13], which was out of scope for this study.

We found that a ClimateBERT model with early stopping, learning rate reduction, and non-synthetic minority class oversampling performed the best. Interestingly, the baseline ClimateBERT model performed similarly well with a slightly lower macro average F1 score. This could suggest that stratified sampling was the key to resolving class representation issues, and further fine-tuning of model layers is much more likely to produce more marked differences in performance metrics.

Referring to detailed charts in Appendix B, it is interesting to see that, throughout all the models, some classes are not predicted at all. Upon examining both our models and the GWSD dataset itself, we theorize that semantic differences in languages between moderate umbrella political leaning classes (e.g. somewhat left vs. somewhat right) are not large. Further, our models are based on mutually exclusive political classes that may not be the best representation of the political leaning spectrum.

Table 4: Political Classification Model Performance Comparison

Model	Accuracy	W. Avg. F1	Macro Avg. F1
Trad. BERT	0.33	0.24	0.11
ClimateBERT	0.59	0.51	0.28
CB Focal Loss	0.08	0.05	0.07
CB SMOTE	0.57	0.49	0.28
CB ES + LR	0.59	0.52	0.31
CB ES + LR + SMOTE	0.57	0.50	0.29
Best Performance	ClimateBERT ES + LR		

Notes: CB = ClimateBERT, W. Avg. = Weighted Average
ES + LR: Early Stopping + Learning Rate Reduction

Finally, the text generation pipeline leverages a T5 model to create politically-nuanced statements about climate change through a T5 tokenization preprocessing and generation approach. As we empirically realized that text generation is much more unpredictable and random than classification, we employed more detailed sampling techniques, including temperature control, [8] top-k and top-p filtering, [7] and penalties to reduce repetition, which help produce diverse and contextually relevant text. Each generated statement is labeled with its corresponding political leaning, creating a synthetic dataset that mirrors the distribution and characteristics of the original corpus.

We generated 100 sentences from each political leaning class and performed a pairwise cosine similarity analysis between sentences in the generated text and sentences in GWSD within their corresponding classes.

As all results are similar, the similarity of each class is likely due to randomness. Within each class's generated sentences, we noticed that sentences are not grammatically sound or are even completely nonsensical, producing chunks of phrases using

[8] see examples in Appendix D.

Table 5: Semantic Similarity Across Political Leaning Classes

Political Leaning	Avg. Similarity
Solidly Left	0.3166
Center	0.2800
Very Right	0.3817
Solidly Right	0.3627
Very Left	0.3368
Somewhat Left	0.3052
Somewhat Right	0.3342

words from the GWSD sentences. Since our T5 model has only 222,903,552 parameters, compared to ChatGPT-3's 175 billion parameters, our text generation model is far from generating well-structured sentences. We further theorize that the "Center" class having the lowest similarity might indicate greater difficulty in generating neutral-sounding statements, and extreme political leanings ("Very Right", "Solidly Right") seem to have slightly higher semantic similarity, potentially due to more distinct linguistic patterns.

6 Conclusion

ClimateBERT did have a marked increase in accuracy and F1 score, compared to both BERT and our Baseline models. Our BERT model tended to overlabel *disagree*, causing it to have a higher recall but lower precision for that class. Next steps would include gathering more labeled sentences, readjusting the class weights, and seeing if this problem persists. Interestingly, BERT did perform better than ClimateBERT on one sentence that used sarcasm. An interesting next step would be to collect data to see how BERT and ClimateBERT perform on climate change related sentences that use sarcasm and other language constructs that are more common in everyday speech.

In order to see more meaningful modeling performance for classifying political leanings, as next steps, we propose investigating a continuous scale of political leaning that may transform our classification problem into a logistic regression problem. To do so, we further propose the need to have more non-discrete labeling (i.e. rather than labeling sentences into given *agree, disagree, neutral* categories) to produce the appropriate training data, which may require a more examination of manual labelers' backgrounds for more diverse, proportional, and accuracy representations.

Appendix A

References

[1] Intergovernmental Panel on Climate Change. *AR6 Synthesis Report: Summary for Policymakers Headline Statements*. www.ipcc.ch, 2023. URL: https://www.ipcc.ch/report/ar6/syr/resources/spm-headline-statements/.

[2] Jennifer Marlon et al. *Yale Climate Opinion Maps 2019 - Yale Program on Climate Change Communication*. Yale Program on Climate Change Communication, Feb. 2022. URL: https://climatecommunication.yale.edu/visualizations-data/ycom-us/.

[3] Manfred Stede and Ronny Patz. “The Climate Change Debate and Natural Language Processing”. In: *Association for Computational Linguistics* (Aug. 2021), pp. 8–18. URL: https://aclanthology.org/2021.nlp4posimpact-1.2.pdf (visited on 12/07/2024).

[4] Markus Leippold and Francesco Saverio Varini. “ClimaText: A Dataset for Climate Change Topic Detection”. In: (Dec. 2020).

[5] Yiwei Luo, Dallas Card, and Dan Jurafsky. “Detecting Stance in Media on Global Warming”. In: *arXiv (Cornell University)* (Jan. 2020). DOI: 10.48550/arxiv.2010.15149. (Visited on 12/07/2024).

[6] Nicolas Webersinke et al. “CLIMATEBERT: A Pretrained Language Model for Climate-Related Text”. In: *SSRN Electronic Journal* (2022). DOI: 10.2139/ssrn.4229146.

[7] Kamaldeep Singh. *How to Improve Class Imbalance using Class Weights in Machine Learning*. Analytics Vidhya, Oct. 2020. URL: https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/.

[8] Siladittya Manna. *Weighted Categorical Cross-Entropy Loss in Keras - The Owl - Medium*. Medium, Aug. 2023. URL: https://medium.com/the-owl/weighted-categorical-cross-entropy-loss-in-keras-edaee1df44ee (visited on 12/09/2024).

[9] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), pp. 1–1. DOI: 10.1109/tpami.2018.2858826.

[10] Jason Brownlee. *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery, Jan. 2020. URL: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[11] Jason Brownlee. *Use Early Stopping to Halt the Training of Neural Networks At the Right Time*. Machine Learning Mastery, Aug. 2020. URL: https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/.

[12] Vrunda Bhattbhatt. *Learning Rate and Its Strategies in Neural Network Training*. Medium, Jan. 2024. URL: https://medium.com/thedeephub/learning-rate-and-its-strategies-in-neural-network-training-270a91ea0e5c.

[13] Guancheng Chen and Huabiao Qin. “Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving”. In: *The Visual Computer* (Jan. 2021). DOI: 10.1007/s00371-021-02067-9. (Visited on 03/09/2021).

Appendix B: Seven Class Model Results

Class	**Precision**	**Recall**	**F1-Score**
center	0.33	0.73	0.45
somewhat left	0	0	0
solidly left	0	0	0
very left	0.33	0.33	0.33
somewhat right	0	0	0
solidly right	0	0	0
very right	0	0	0
Macro avg	0.09	0.15	0.11
Weighted avg	0.20	0.33	0.24

Table 6: Traditional BERT Classification Report

Class	**Precision**	**Recall**	**F1-Score**
center	0.52	0.88	0.65
somewhat left	0	0	0
solidly left	0	0	0
very left	0.81	0.69	0.74
somewhat right	0	0	0
solidly right	0	0	0
very right	0.50	0.71	0.59
Macro avg	0.26	0.33	0.28
Weighted avg	0.48	0.59	0.51

Table 7: ClimateBERT Classification Report

Class	**Precision**	**Recall**	**F1-Score**
center	0	0	0
somewhat left	0	0	0
solidly left	0.11	0.07	0.00
very left	0	0	0
somewhat right	0.04	0.60	0.08
solidly right	0	0	0
very right	0.28	0.32	0.30
Macro avg	0.06	0.14	0.07
Weighted avg	0.05	0.08	0.05

Table 8: ClimateBERT Focal Loss Classification Report

Class	**Precision**	**Recall**	**F1-Score**
center	0.59	0.73	0.65
somewhat left	0	0	0
solidly left	0	0	0
very left	0.56	0.79	0.65
somewhat right	0	0	0
solidly right	0	0	0
very right	0.54	0.75	0.63
Macro avg	0.24	0.32	0.28
Weighted avg	0.43	0.57	0.49

Table 9: ClimateBERT SMOTE Classification Report

Class	**Precision**	**Recall**	**F1-Score**
center	0.59	0.76	0.66
somewhat left	0	0	0
solidly left	0.40	0.13	0.20
very left	0.66	0.82	0.73
somewhat right	0	0	0
solidly right	0	0	0
very right	0.50	0.68	0.58
Macro avg	0.31	0.34	0.31
Weighted avg	0.48	0.59	0.52

Table 10: ClimateBERT ES + LR Classification Report

Class	**Precision**	**Recall**	**F1-Score**
center	0.57	0.76	0.66
somewhat left	0.25	0.03	0.05
solidly left	0	0	0
very left	0.62	0.76	0.68
somewhat right	0	0	0
solidly right	0	0	0
very right	0.54	0.71	0.62
Macro avg	0.28	0.32	0.29
Weighted avg	0.46	0.57	0.50

Table 11: ClimateBERT ES + LR + SMOTE Classification Report

Appendix C: Samples of Machine Generated Text

Center: a center political statement about climate change: Use the Center Political Statement Generator to create & . an official climate statement. — Climate change. climate policy. center statement: Generate and Climate Change.: change climate.::: Create climate:: climate action: Create an environmental statement of climate changing:: : create, edit and share – (Climate change) Gene

Solidly Left: a solid left political statement about climate change: the power of your left. Generate p: Create logical political argument about global warming: Generated by rsa. Create an uncompromising political claim about the climate: Support pacts with climate scientists: Help them make measurable progress on climate and economic change? Explain: Climate change is affecting the planet by 2030. Support: A strong, solid link political declaration about how climate changes affect our planet: - Support climate policies:

Solidly Right: a solid political statement about climate change: Generate .... Generated by the Green Party:.:: Create & Support —:: Get the message:: (and get it) – Get delegates from the White House: Get them involved: • Get people to vote on climate issues; • Bring in climatologists; and • Provide information and support for climate policy.

Somewhat Left: a somewhat left political statement about climate change. Generate te:: Generated essentially - i.e., an attack on climate action. (d) Genere climatic change: (c) Generate neoconservative statements about global warming:(d).

Somewhat Right: Generate a somewhat right political statement about climate change: - . ... ?.: Generated if you want to start generating : very good political reaction: Climate change.: (Re)enable ::: statement: :: About Climate Change:: :, an interesting political position: A bit right climate policy::

Very Left: A very left political statement about climate change: a very right political speech about it. Generate ... Very left-wing political statements about Climate Change:. extreme left: Genere an very, very small right-left political comment about this:: statement:::: #shutters:::::: Very Left about . This is the first time we have completely:.

Very Right: a very right political statement about climate change:! Generate ONE very good political declaration about Climate Change:. Generated & A very correct political position about the climate crisis:. The climate is warming.:... : and climate? Use it!

Appendix D: Example Agree Sentences Incorrectly Labeled as Neutral

- The globally averaged sea surface temperature for 2013 is among the 10 warmest on record.
- Global temperatures in 2014 shattered earlier records, making 2014 the hottest year since record-keeping began in 1880.
- A majority of respondents believe global warming has already begun.
- We can expect the Arctic to be ice-free in summer within 20 years.
- The Trump administration simply discarded prior factual findings related to climate change to support its course reversal.
- You cannot go too far on the issue of climate change.
- Protecting still-undamaged forests could have a strong climate benefit as well.
- Coal would have to be phased out even before the Paris Agreement to combat climate change.
- Harvard climate action plan explicitly recognizes what the science has made clear.
- The carbon impact increased sixfold over the period.