GWSD: Climate Change Stance Classification in News Media

March Saper

Abstract

Opinion stance detection is a powerful tool for evaluating news media articles for misinformation or argument framing. A subset of stance detection can be found in the climate change "debate", where news media articles often publish statements which deny or mitigate the veracity of climate change. Our work leverages the Global Warming Stance Dataset (GWSD) and fine-tunes a BERT classifier to label agreement level on the veracity of climate change. Our highest-performing classifier achieves accuracy levels comparable to human annotation. These models present the opportunity for climate-change-specific misinformation labeling or argument parsing.

Introduction

Although climate change scientists have reached a consensus that climate change is real, caused by human activity, and will have tangible and adverse effects on the planet and humanity, public comments about climate change still remain controversial[1]. Within this research space, discourse analysis of opinions and quotes from news media is a developing area of interest for two reasons. First, it can facilitate accurate labeling and mitigating misinformation, such as contextualizing news articles shared on social media with factual counterpoints. Second, it can be applied to broader problems such as surveying news media outlet argumentation strategies or identifying disingenuous framings of information[2].

Background

Stance Detection is a well-researched task in Natural Language Processing and is defined as determining if an audience is in favor, neutron or against a target text. It is the foundation of many tasks, such as fake news detection, which are well-researched and published. Among fake news detection research, Convolutional Neural Networks and Recurrent Neural Networks are well-studied, and combined CNN-LSTM models comprise the architecture basis of many recent papers[3][4]. At the same time, since the introduction of Bidirectional Encoder Representations from Transformers (BERT) in 2019 there has been a growth in the use of this model for similar classification tasks[5][6].

Within the fake news research space, however, there has not been much prior effort into climate-change-specific classification. This changed, however, with the introduction of the GWSD Dataset which was presented, along with a BERT stance labeling model, in 2020[7]. This dataset facilitates research in several exciting spaces: (1) Climate change stance detection for misinformation labeling purposes, and

[1] https://aclanthology.org/2021.nlp4posimpact-1.2
[2] https://aclanthology.org/2021.nlp4posimpact-1.2
[3] https://www.sciencedirect.com/science/article/pii/S2667096820300070
[4] https://ieeexplore.ieee.org/document/9178321
[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8053013/
[6] https://ieeexplore.ieee.org/document/9666618
[7] https://aclanthology.org/2020.findings-emnlp.296/

(2) Argument mining across new outlets to uncover patterns of opinion framing. Neither this dataset, nor the BERT model associated with it have yet been widely used or studied. This presents an opportunity to validate and explore fine-tuning methods to improve upon the previously-introduced stance detection model.

Methods

GWSD Dataset

The Global Warming Stance Dataset (GWSD) consists of spans extracted from 65K news articles related to global warming published from Jan 1 2000 to April 12 2020 by 63 U.S. news sources. Leveraging a coreference pre-processing pipeline and rule-based algorithm, researchers extracted opinionated statements which make claims about climate change[8]. Eight annotators from AMT labeled 2,050 of these spans with the labels *"Agree", "Neutral",* or *"Disagree"*. These labels are used to predict the probability of each label for each stance.

The distribution of opinions skews towards the *"Agree"* and *"Neutral"* labels, with the *"Disagree"* label having roughly half of the occurrences of the previous two.

Stance	Count
Agree	776
Neutral	870
Disagree	399

DeSMOG dataset label distribution shows class imbalance for the "Disagree" label.

Stance detection is an inherently challenging task, and the probability model approach helps smooth labels to sentences which are otherwise ambiguous. In the dataset there are numerous examples of sentences which were classified as all three stances by the annotators. The nature of the dataset, as short quotes largely disambiguated from the broader article context, as well as the inherent nuances involved in conveying an opinion make this an interesting dataset for modeling with deep learning as there may be some subtext not immediately obvious to an annotator that a model can learn.

Global warming is inevitably going to be, at best, managed.											
W1	W2	W3	W4	W5	W6	W7	W8	Disagree	Agree	Neutral	Stance
d	a	a	n	a	n	d	n	0.199417	0.477698	0.322885	Agree
Climate deniers blame global warming on aliens from outer space.											

[8] https://github.com/yiweiluo/GWStance

W1	W2	W3	W4	W5	W6	W7	W8	Disagree	Agree	Neutral	Stance
d	a	n	a	n	n	a	n	0.051909	0.387853	0.560239	Neutral

Examples from the DeSMOG dataset which highlight the difficulty of stance detection.
(a=agree, n=neutral, d=disagree)

Evaluation Metrics

When choosing evaluation metrics it is important to optimize for use-case as different types of failures have greater detriment depending on how the model is applied. For our research, we focus on F1 scores for each of the three categories. The F1 score represents the harmonic mean between precision and recall and is a standard way to evaluate models for their overall performance.

Baseline

Stance detection is a challenging task, even for a human annotator. Towards that end, the accuracy of our models will have a relatively low baseline, compared to other stance detection datasets. As a baseline comparison, we use the majority class model and annotator accuracy.

Baseline	Accuracy
Majority Class	0.41
Anntator	0.71

Baseline model performance values.

BERT Baseline

We build our baseline BERT model using the Bert Base Cased pretrained tokenizer and BERTbase architecture, following best-practices described in the DeSMOG paper (where fine-tuning on a language modeling task with raw new data was not found to increase performance). The Baseline BERT tokenizes with a max length of 256 and leverages the pooled layer output for classification. The DeSMOG paper also examined downsampling to mitigate class-imbalance, but found that had no effect on BERT performance.

Weighted BERT

A feature of the GWSD dataset is the inclusion of Bayesian probabilities for each possible label. This provides the opportunity to update the BERT training model loss function with additional scaling by label probability. While the DeSMOG paper took the approach of training every instance with every scaled label, we instead include only the weight of the most probable class[9]. All other BERT architecture remains the same.

[9] https://www.sciencedirect.com/science/article/pii/S1389128621002711#sec3.3

BERT-CNN

Building on the weighted BERT model, we also explore the use of a Convolutional Neural Network to further fine-tune the BERT outputs. Towards this end, we explore two structures. First, a CNN built on the sequence output, with the CLS and SEP tokens removed. Second, we follow best-practices outlined upon the introduction of BERT[10] and leverage a concatenation of the final four hidden layers of the BERT output as inputs to a CNN.   Reference the illustrated bert and the paper that suggests that 4-hidden layers are best to use with cnn on top of bert

Results

Models were trained using 5-fold cross validation and evaluated against a held-out test set provided by the GWSD dataset. The 200-sample test set is stratified by label and political leaning of the source media outlet. The remaining 1850 examples were used for training.

Models were trained over a minimum of 7 epochs, and the best-performing model, as determined by validation accuracy, was then evaluated against the held out test set. Since model parameters were taken from prior art best practices, no hyper-parameter tuning was done in order to complete experimentation in a limited resource computation environment.

The highest-performing models, by both accuracy and F1-scores, were the BERT models which used the pooled output layer for classification (BERT Baseline and Weighted BERT). The addition of label weights in training showed a significant improvement in F1 score for *Disagree* and *Agree* labels, as well as a slight increase in overall accuracy.

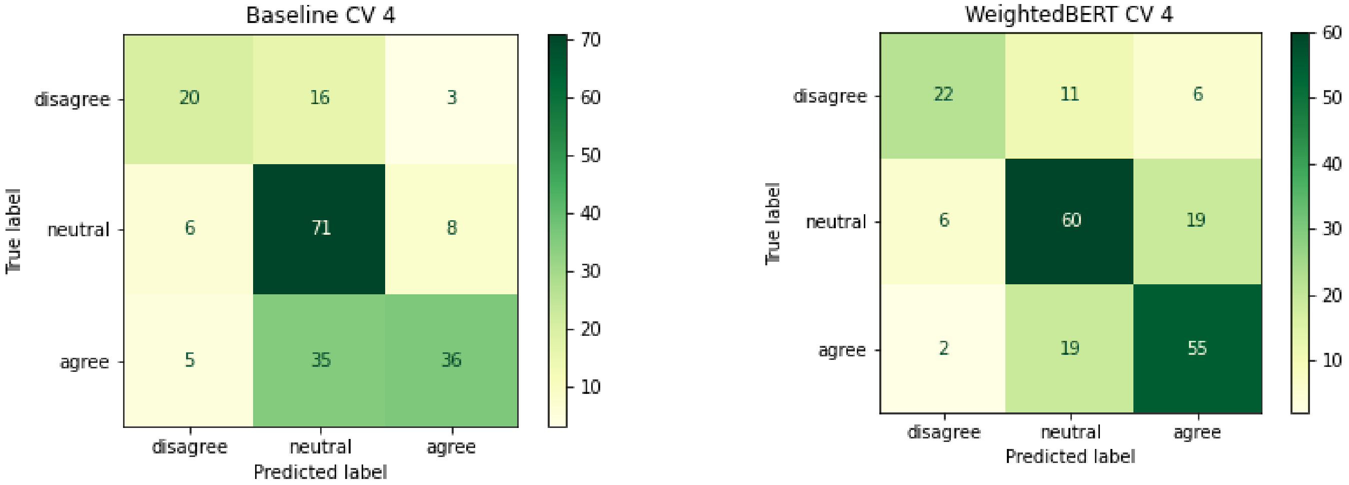
Model	**Accuracy**	**F1-Disagree** **39 Support**	**F1-Neutral** **85 Support**	**F1-Agree** **76 Support**
BERT Baseline	0.64	0.57	0.69	0.59
Weighted BERT	0.69	0.64	0.69	0.71
BERT-CNN	0.69	0.57	0.72	0.70
BERT-4CNN	0.60	0.51	0.64	0.61

Test-set performance for highest-scoring cross-fold validation.

Discussion

A comparison between the Baseline BERT model and the Weighted BERT model shows a significant increase in the number of correctly predicted *"Agree"* labels, corresponding to an increase in recall for that label.

[10] https://aclanthology.org/N19-1423/



Confusion matrices for two highest-scoring by accuracy BERT models evaluated on the held-out test set.

While maximizing F1 scores and recall is beneficial for news media stance and argument modeling, we also propose that certain types of errors are more serious than others. Namely, the classification of a *"Disagree"* stance as *"Agree"*. This mistake impacts not only news media argument/stance modeling but also (and perhaps more importantly) labeling of climate change mis-information. To better understand the mistakes the Weighted BERT model was making, we examined the held out test dataset for these errors. Among the 39 *"Disagree"* support sentences, 6 were mis-classified by Weighted BERT. We classify these errors into two categories: mis-classifications made by both the Baseline BERT and Weighted BERT, and mis-classifications made only by the Weighted BERT.

Examining these mis-classifications, a couple of themes stand out:

- *Use of words such as "evidence", "theory", and "taught":* It seems likely that the Weighted BERT model is learning to associate these words with an increased agreement with climate change, and losing the contrary context.
- *Use of extreme words such as "dire", "threat", and "finally":* It also seems plausible that the model is associating extreme words with agreement with climate change.

It should also be noted that all of these mis-classifications were labeled as all three of *"Disagree", "Neutral",* and *"Agree"* by the annotators, further highlighting the challenge of this task even for human readers.

Sentence	W1	W2	W3	W4	W5	W6	W7	W8	Truth	Baseline	Weighted
I am hoping that the scientists and politicians who have been blindly demonizing carbon dioxide for 37 years will one day open their eyes and look at the evidence.	d	d	a	d	d	n	d	d	disagree	disagree	agree

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.	n	d	a	a	d	n	n	d	disagree	disagree	agree
The world is finally on a path toward controlling global warming.	n	d	d	d	a	n	d	n	disagree	agree	agree
Global Warming is a theory and should be taught as such in our schools.	d	d	n	n	d	d	d	d	disagree	neutral	agree
When liberals talk about the dire threat of global warming, liberalsâ™re actually seizing opportunistically on the issue.	a	d	a	a	n	a	d	n	disagree	disagree	agree
But the erosion of science reaches well beyond the environment and climate.	d	d	d	n	n	n	d	n	disagree	neutral	agree

Weighted BERT model failures which mis-classify anti-climate change stances as “agree” and compared to baseline BERT. (a=agree, n=neutral, d=disagree)

Conclusion

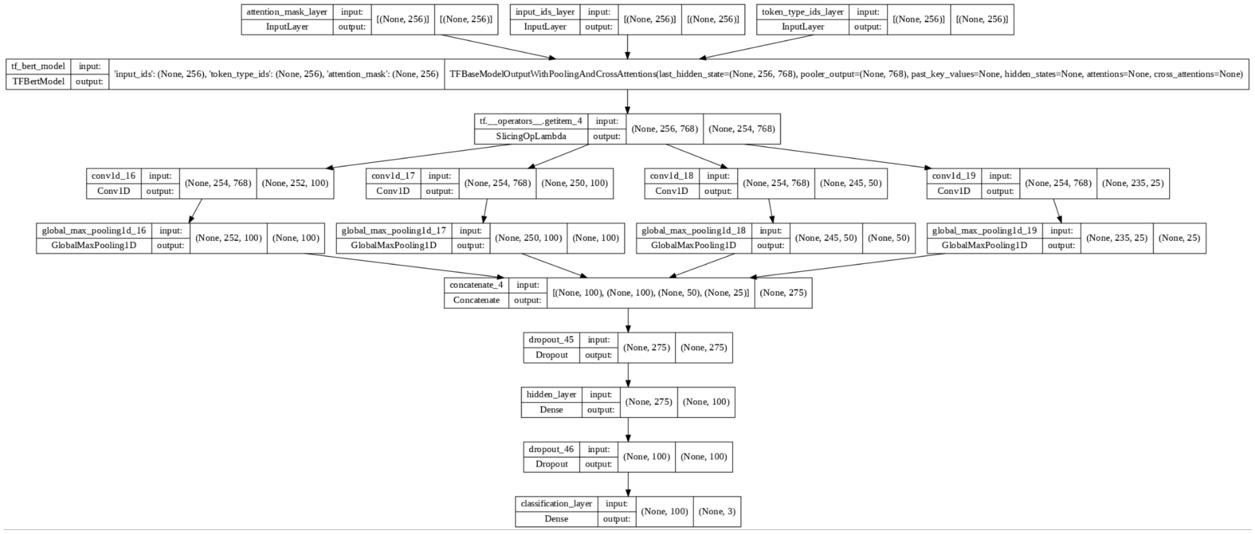
In this paper, we presented validation of deep learning approaches to classify stance on global warming in the GWSD dataset, a relatively new dataset to the stance-classification research. Our approaches included the use of a BERT-CNN for the first time. Our models produced accuracy results comparable to human classification and demonstrated the complexity of stance classification, even for a native English speaker. This challenge was highlighted in our discussion of our highest-performing model’s failures.

As future work, we propose continued examination of stance - specifically stance labels which were challenging for annotators to reach a consensus on to learn the impacts of how arguments are presented, and how readers interpret the sentence, especially at first glance. Efforts towards this research already exist, and could be further supported by use of language modeling tools to mine sentence structure or named entities. [11]

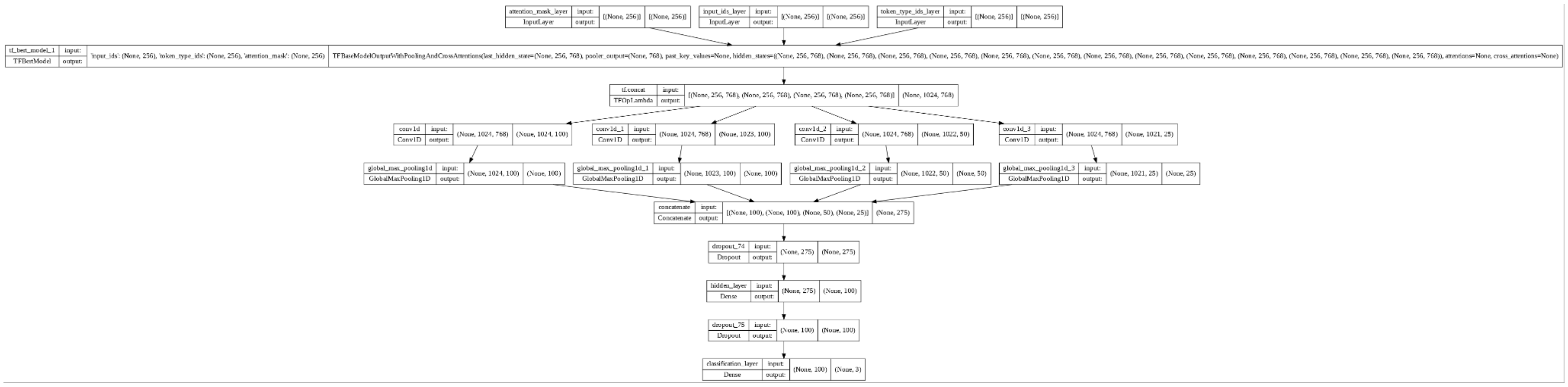
[11] https://journals.sagepub.com/doi/full/10.1177/0002764219878224

Appendix

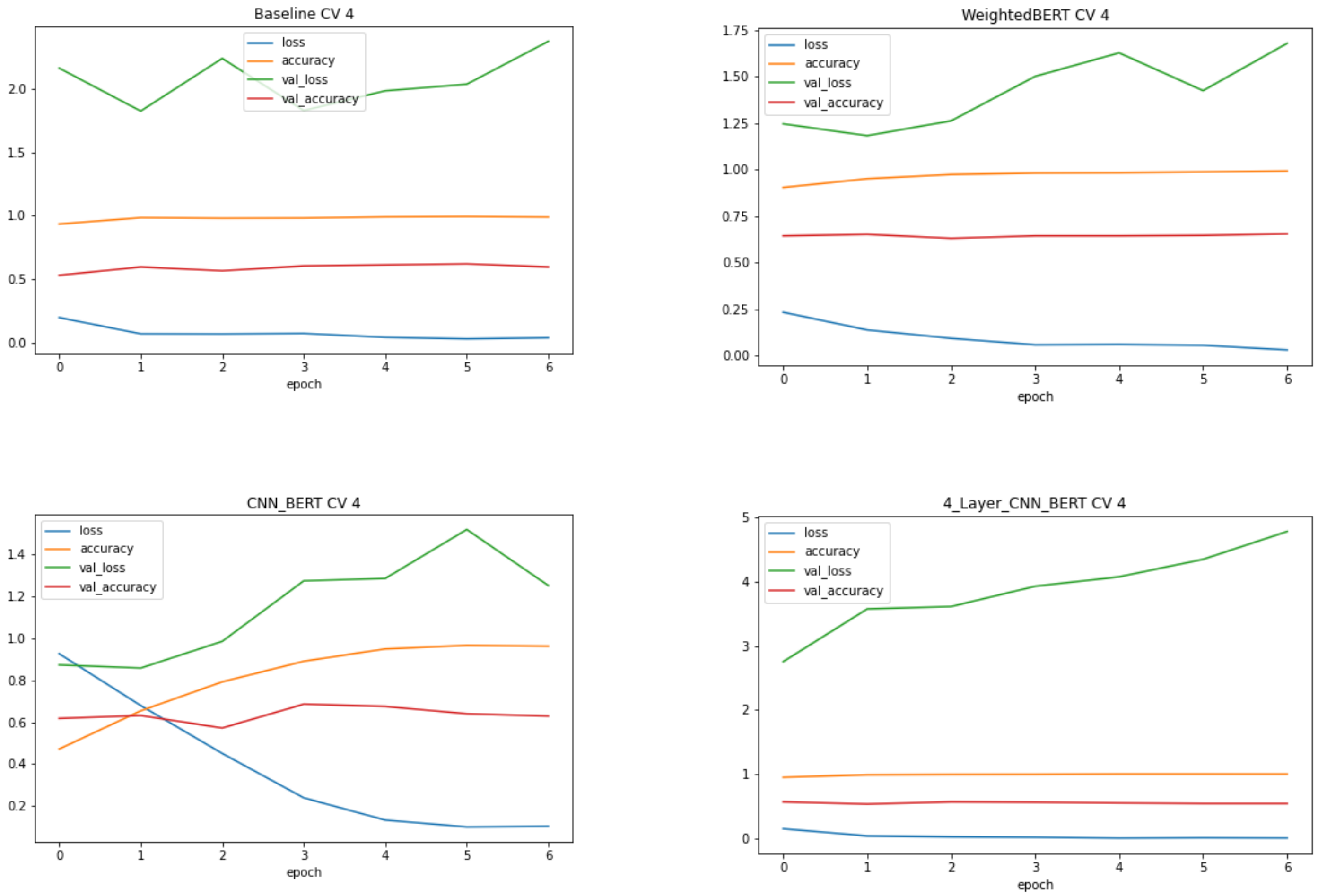
BERT-CNN Sequence Output Model


attention_mask_layer | input: | [(None, 256)] | [(None, 256)]
InputLayer | output:
input_ids_layer | input: | [(None, 256)] | [(None, 256)]
InputLayer | output:
token_type_ids_layer | input: | [(None, 256)] | [(None, 256)]
InputLayer | output:
tf_bert_model | input: | 'input_ids': (None, 256), 'token_type_ids': (None, 256), 'attention_mask': (None, 256) | TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 256, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)
TFBertModel | output:
tf.__operators__.getitem_4 | input: | (None, 256, 768) | (None, 254, 768)
SlicingOpLambda | output:
conv1d_16 | input: | (None, 254, 768) | (None, 252, 100)
Conv1D | output:
conv1d_17 | input: | (None, 254, 768) | (None, 250, 100)
Conv1D | output:
conv1d_18 | input: | (None, 254, 768) | (None, 245, 50)
Conv1D | output:
conv1d_19 | input: | (None, 254, 768) | (None, 235, 25)
Conv1D | output:
global_max_pooling1d_16 | input: | (None, 252, 100) | (None, 100)
GlobalMaxPooling1D | output:
global_max_pooling1d_17 | input: | (None, 250, 100) | (None, 100)
GlobalMaxPooling1D | output:
global_max_pooling1d_18 | input: | (None, 245, 50) | (None, 50)
GlobalMaxPooling1D | output:
global_max_pooling1d_19 | input: | (None, 235, 25) | (None, 25)
GlobalMaxPooling1D | output:
concatenate_4 | input: | [(None, 100), (None, 100), (None, 50), (None, 25)] | (None, 275)
Concatenate | output:
dropout_45 | input: | (None, 275) | (None, 275)
Dropout | output:
hidden_layer | input: | (None, 275) | (None, 100)
Dense | output:
dropout_46 | input: | (None, 100) | (None, 100)
Dropout | output:
classification_layer | input: | (None, 100) | (None, 3)
Dense | output:

BERT-CNN Hidden Layer Model

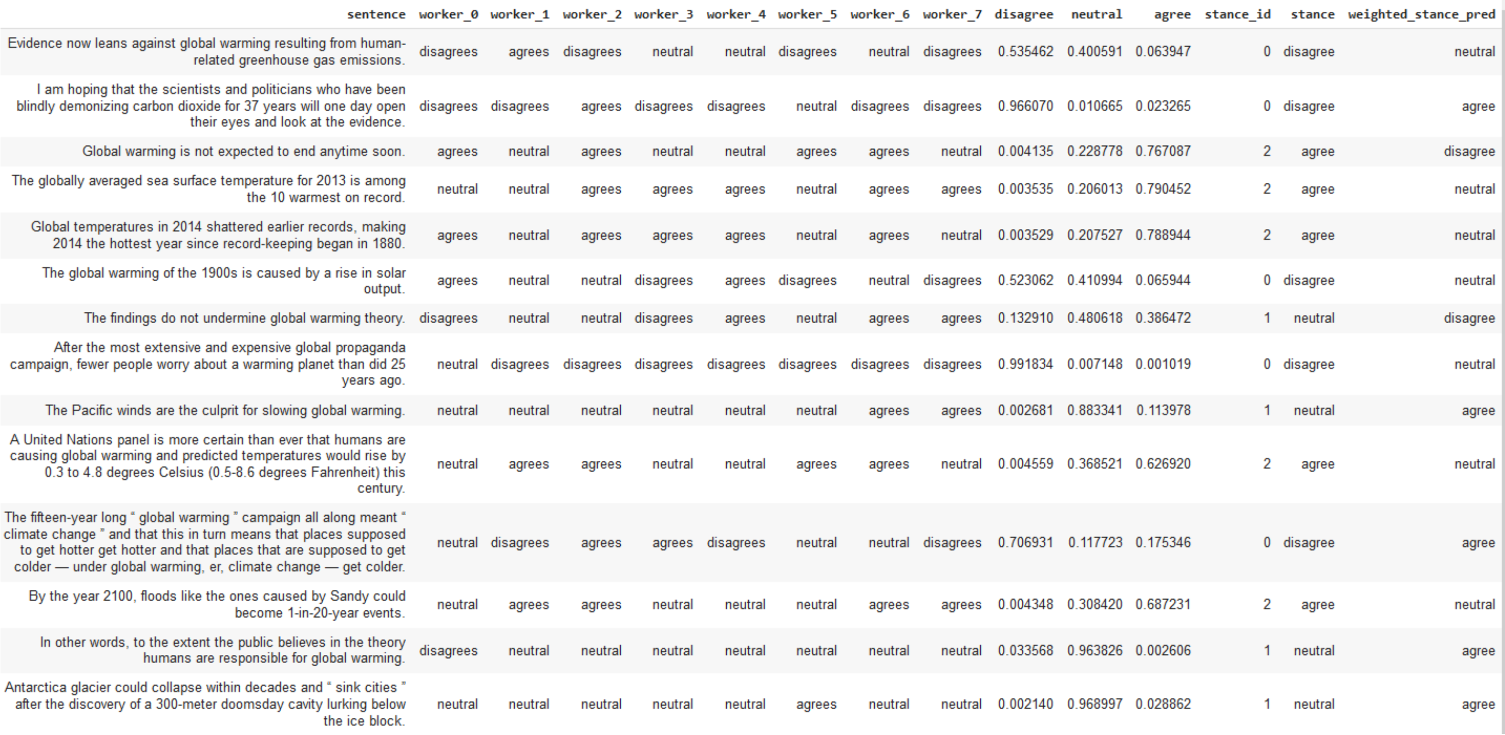

attention_mask_layer | input: | [(None, 256)] | [(None, 256)]
InputLayer | output:
input_ids_layer | input: | [(None, 256)] | [(None, 256)]
InputLayer | output:
token_type_ids_layer | input: | [(None, 256)] | [(None, 256)]
InputLayer | output:
tf_bert_model_1 | input: | 'input_ids': (None, 256), 'token_type_ids': (None, 256), 'attention_mask': (None, 256) | TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 256, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=((None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768)), attentions=None, cross_attentions=None)
TFBertModel | output:
tf.concat | input: | [(None, 256, 768), (None, 256, 768), (None, 256, 768), (None, 256, 768)] | (None, 1024, 768)
TFOpLambda | output:
conv1d | input: | (None, 1024, 768) | (None, 1024, 100)
Conv1D | output:
conv1d_1 | input: | (None, 1024, 768) | (None, 1023, 100)
Conv1D | output:
conv1d_2 | input: | (None, 1024, 768) | (None, 1022, 50)
Conv1D | output:
conv1d_3 | input: | (None, 1024, 768) | (None, 1021, 25)
Conv1D | output:
global_max_pooling1d | input: | (None, 1024, 100) | (None, 100)
GlobalMaxPooling1D | output:
global_max_pooling1d_1 | input: | (None, 1023, 100) | (None, 100)
GlobalMaxPooling1D | output:
global_max_pooling1d_2 | input: | (None, 1022, 50) | (None, 50)
GlobalMaxPooling1D | output:
global_max_pooling1d_3 | input: | (None, 1021, 25) | (None, 25)
GlobalMaxPooling1D | output:
concatenate | input: | [(None, 100), (None, 100), (None, 50), (None, 25)] | (None, 275)
Concatenate | output:
dropout_74 | input: | (None, 275) | (None, 275)
Dropout | output:
hidden_layer | input: | (None, 275) | (None, 100)
Dense | output:
dropout_75 | input: | (None, 100) | (None, 100)
Dropout | output:
classification_layer | input: | (None, 100) | (None, 3)
Dense | output:

BERT CV Experiments Epoch Training



Epoch training results from best-performing BERT CV experiments.

Further Weighted BERT Test Result Failures



sentence	worker_0	worker_1	worker_2	worker_3	worker_4	worker_5	worker_6	worker_7	disagree	neutral	agree	stance_id	stance	weighted_stance_pred
Evidence now leans against global warming resulting from human-related greenhouse gas emissions.	disagrees	agrees	disagrees	neutral	neutral	disagrees	neutral	disagrees	0.535462	0.400591	0.063947	0	disagree	neutral
I am hoping that the scientists and politicians who have been blindly demonizing carbon dioxide for 37 years will one day open their eyes and look at the evidence.	disagrees	disagrees	agrees	disagrees	disagrees	neutral	disagrees	disagrees	0.966070	0.010665	0.023265	0	disagree	agree
Global warming is not expected to end anytime soon.	agrees	neutral	agrees	neutral	neutral	agrees	agrees	neutral	0.004135	0.228778	0.767087	2	agree	disagree
The globally averaged sea surface temperature for 2013 is among the 10 warmest on record.	neutral	neutral	agrees	agrees	agrees	neutral	agrees	agrees	0.003535	0.206013	0.790452	2	agree	neutral
Global temperatures in 2014 shattered earlier records, making 2014 the hottest year since record-keeping began in 1880.	agrees	neutral	agrees	agrees	agrees	neutral	agrees	neutral	0.003529	0.207527	0.788944	2	agree	neutral
The global warming of the 1900s is caused by a rise in solar output.	agrees	neutral	neutral	disagrees	agrees	disagrees	neutral	disagrees	0.523062	0.410994	0.065944	0	disagree	neutral
The findings do not undermine global warming theory.	disagrees	neutral	neutral	disagrees	agrees	neutral	agrees	agrees	0.132910	0.480618	0.386472	1	neutral	disagree
After the most extensive and expensive global propaganda campaign, fewer people worry about a warming planet than did 25 years ago.	neutral	disagrees	disagrees	disagrees	disagrees	disagrees	disagrees	disagrees	0.991834	0.007148	0.001019	0	disagree	neutral
The Pacific winds are the culprit for slowing global warming.	neutral	neutral	neutral	neutral	neutral	neutral	agrees	agrees	0.002681	0.883341	0.113978	1	neutral	agree
A United Nations panel is more certain than ever that humans are causing global warming and predicted temperatures would rise by 0.3 to 4.8 degrees Celsius (0.5-8.6 degrees Fahrenheit) this century.	neutral	agrees	agrees	neutral	neutral	agrees	agrees	neutral	0.004559	0.368521	0.626920	2	agree	neutral
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.	neutral	disagrees	agrees	agrees	disagrees	neutral	neutral	disagrees	0.706931	0.117723	0.175346	0	disagree	agree
By the year 2100, floods like the ones caused by Sandy could become 1-in-20-year events.	neutral	agrees	agrees	neutral	neutral	neutral	agrees	agrees	0.004348	0.308420	0.687231	2	agree	neutral
In other words, to the extent the public believes in the theory humans are responsible for global warming.	disagrees	neutral	neutral	neutral	neutral	neutral	neutral	neutral	0.033568	0.963826	0.002606	1	neutral	agree
Antarctica glacier could collapse within decades and " sink cities " after the discovery of a 300-meter doomsday cavity lurking below the ice block.	neutral	neutral	neutral	neutral	neutral	agrees	neutral	neutral	0.002140	0.968997	0.028862	1	neutral	agree

More weighted BERT sub-category failures.