DeSMOG: Climate Change Stance Classification in News Media

March Saper

Abstract

Last thing to do…..

Introduction

- Why climate change research matters (find motivation from summary paper)
- Why stance classification matters for climate news
- Different motivations for that: research for arg parsing/ labeling climate mis-information (hint at different eval metrics)

Background

- Prior research in spam and fake news detection… reference the two cnn-lstm fake news papers
- Papers on bert maybe just the general bert being used for classification…in general
- he desmog dataset as an intro to build off of as a new and novel dataset that expands to a specific area of mis-information.

Methods

DeSMOG Dataset

- Describe the dataset and how it was made (link to github here…)
- Bayesian model to produce label and why that matters
- Class imbalance mention but also reference paper saying downsampling didn't imrove

Stance	Count
Agree	776
Neutral	870
Disagree	399

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

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.											
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

Introduce f1 as important here…
Talk about accuracy as epoc-based metric during training
Issue: probably want precision to be maximized for agree since don't want to promote mis-information…but if precision is minimzed, then maybe we want to focus on higher disagree/neutral precision for warning labels… which is what see withweighted so that is okay. ?? maybe

Feel like weighted actually made it worse bc we want to be very precise with labeling of agree.
Precision of finding disagree is something we want to maximize so that we can correctly flag mis-information…maybe

For argument parsing/language modeling we may more about overall accuracy or f1 scores improving unilaterally.

For a different task such as labeling problematic disagree phrases we care about [predicion/recall?? Of disagree?? Or of agree??]
Basically… just there are two use cases.

All around accuracy can be useful for modeling purposes…such as what was introduced in this paper For modeling tasks, focusing on f1 to balance accuracy and precision
Another approach is to label mis-information… for that we care about…..???

- Accuracy:
- Precision:
- Recall:
- F1:

Baseline

Show majority class results and human accuracy… reference above howdifficult task.
Then bert - base-case also why didn't use pre-trained bert from news articles.

Epochs, and max length taken from bert paper

Weighted BERT

Discuss use of weighted values in BERT (there should be a reference here…)
How did i feed in the weights
Discuss use of pooled output as choice for classification

BERT-CNN

Extension of what was presented in the DeSMOG paper, as in wasn't discussed there at all
Reference the illustrated bert and the paper that suggests that 4-hidden layers are best to use with cnn on top of bert

CNN-LSTM

Talk about what kind of lstm is chosen here.
Talk about word2vec embeddings and this benchmark comparison between dynamic embeddings from bert and these static embeddings
MAJOR TODO [1][2]

Results

Talk about 5 fold cross validation and held out dataset for final evaluation. Talk about training to over-fit and then saving out the best epoch.
Mention no hyperparameter tuning, parameters taken as best-practics from prior papers

Show a chart that lists the model tried, and accuracy, and f1 score by class

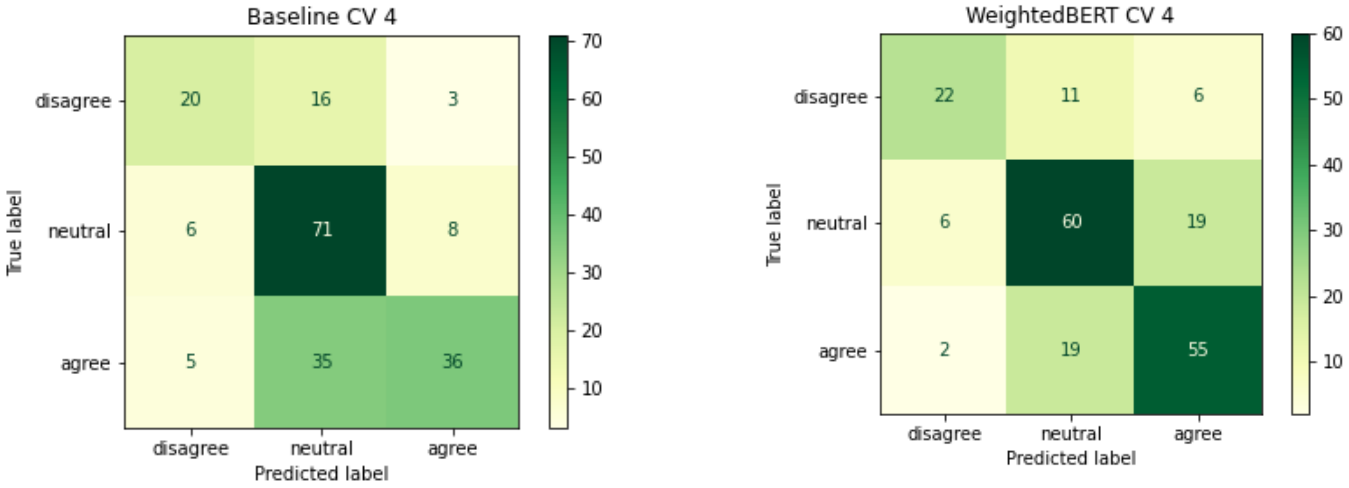
Model	Accuracy	F1-Disagree 39 Support	F1-Neutral 85 Support	F1-Agree 76 Support
BERT Baseline	0.64	0.57	0.69	0.59
Weighted BERT	0.69	0.64	0.69	0.71
BERT-CNN	0.69	0.57	0.72	0.70
BERT-4CNN	0.60	0.51	0.64	0.61
CNN-LSTM				

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

[1] https://keras.io/examples/nlp/bidirectional_lstm_imdb/
[2] https://keras.io/examples/nlp/text_classification_from_scratch/

Discussion

If need space: insert images of training loss/acc/val stuff… then talk about small dataset and overfitting??



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

Discussion from a f1 perspective

Discussion from a misinformation labeling perspective…focusing on egregious errors…

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)

Discussion from a SpaCy perspective

- Baseline bert performed better on disagree/neutral. The weighted average bert tended to mis-classify those as agree… which is not great bc don't want to spread false information…

lk about error analysis here… which you need to do… just do it for the highest performing model… so the weighted BERT… bc don't think will have cnn lstm done in time and should probably focus on just getting some good discussion in here instead of the

However…some nuance here. We decided to model for (some metric that makes the weighed bert look better)...however a different application might favor the baseline bert because of it's precision at identifying disagree/neutral and not mis-labeling as agree…which the weighted bert does do…

So… for the task of flagging mis-information weighted bert probably not better because incorrectly assigning "agree" label to disagree/neutral stances… some examples:

Conclusion

Summarize what did

- Worked w new dataset
- Expanded on models applied to the dataset
- Achieved results roughly commiserate with human performance
- Investigated model failures and buzzwords

Future model work:

- Bert trained on fake news stance classification maybe (alternate word embeddings)
- Feed in batches of sentences instead of 1 sentence at a time
- Hyper parameter tuning of bert cnn and cnn-lstm

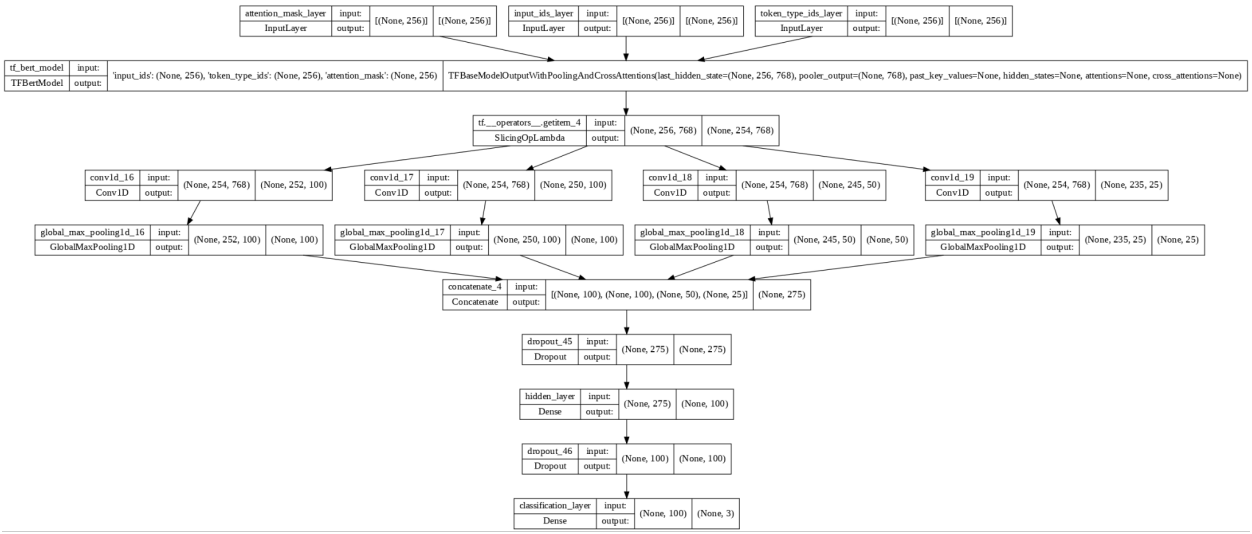
Leveraging argument parsing to see if certain types of argument presentations are more frequently mis-classified

- Perhaps can use combined model methods to identify confusing argument structures/confusing words
- Spacy to identify potentially confusing statements?
    - Maybe there is a link between these word usage and confusing the readers? (COOL IF FIND LINK FOR THIS)

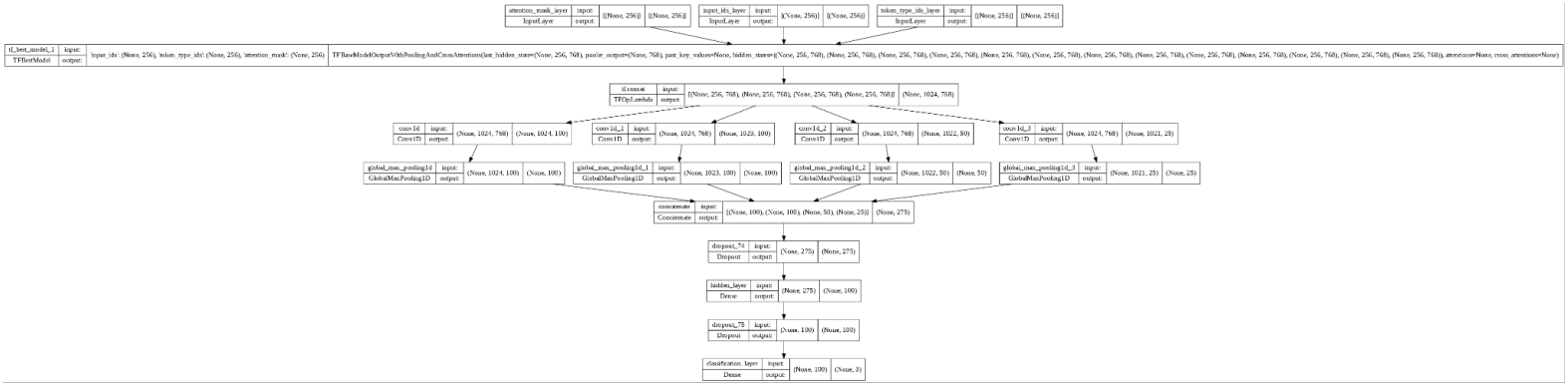
- Apply trained model to more data..like in the paper to look for argument styles across news types….

Appendix

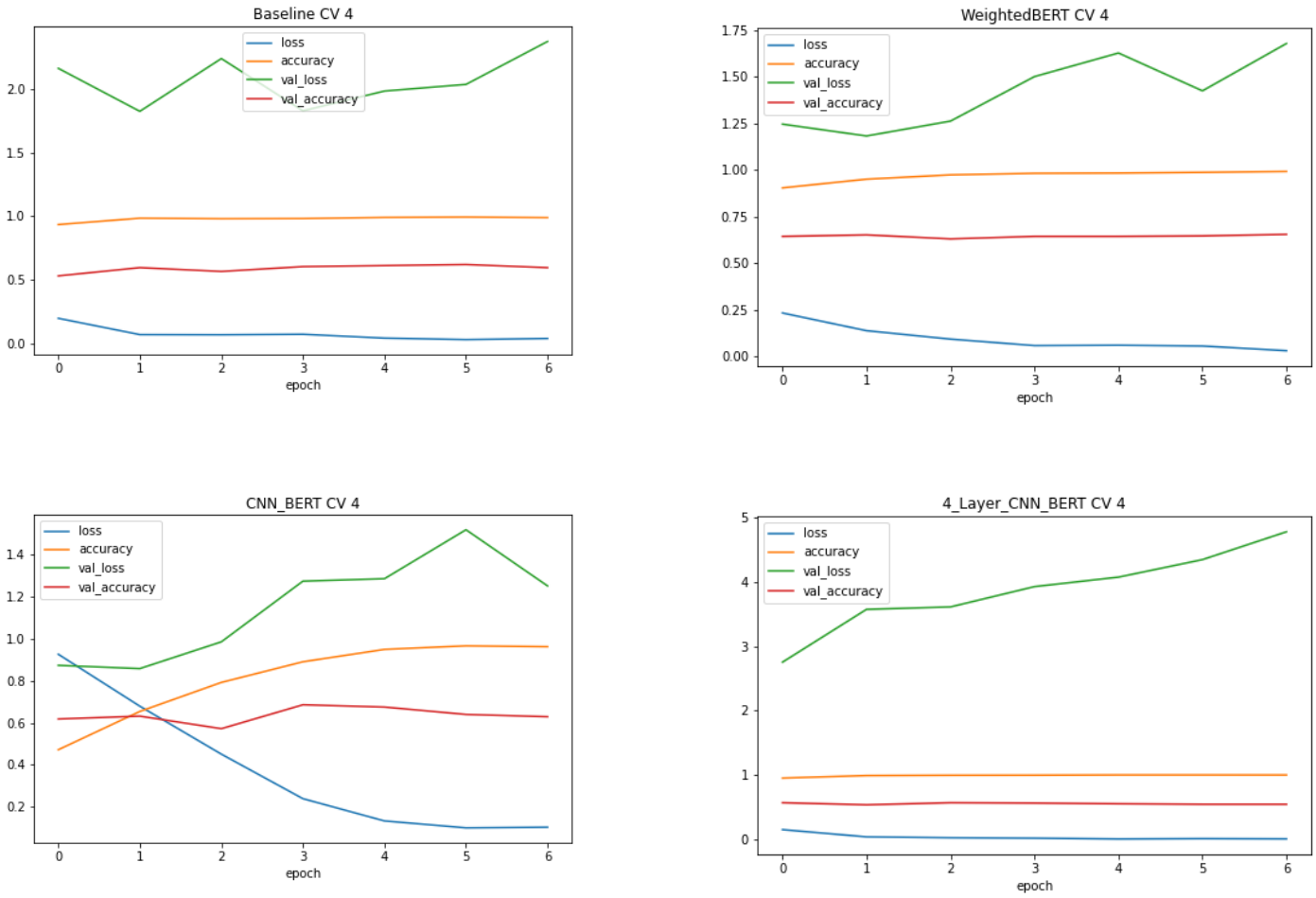
BERT-CNN Sequence Output Model



BERT-CNN Hidden Layer Model



BERT CV Experiments Epoch Training



Epoch training results from best-performing BERT CV experiments.

Weighted BERT Test Result Failures: Predicting "Agree" when stance ground truth is "Disagree"

For the weighted BERT model, 6 test cases (out of a support of 39) were predicted to be "agree" when the ground truth label was "disagree". For climate-change news misinformation labeling, these instances should be minimized.

	sentence	worker_0	worker_1	worker_2	worker_3	worker_4	worker_5	worker_6	worker_7	disagree	neutral	agree	stance_id	stance	base_stance_pred	weighted_stance_pred
7	I am hoping that the scientists and politicians who have been blindly demonizing carbon dioxide for\n37 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	disagree	agree
30	The fifteen-year long " global warming " campaign all along meant " climate change " and that this\nin turn means that places supposed to get hotter get hotter and that places that are supposed to get\ncolder — under global warming, er, climate change — get colder.	neutral	disagrees	agrees	agrees	disagrees	neutral	neutral	disagrees	0.706931	0.117723	0.175346	0	disagree	disagree	agree
73	The world is finally on a path toward controlling global warming.	neutral	disagrees	disagrees	disagrees	agrees	neutral	disagrees	neutral	0.692492	0.249896	0.057611	0	disagree	agree	agree
86	Global Warming is a theory and should be taught as such in our schools.	disagrees	disagrees	neutral	neutral	disagrees	disagrees	disagrees	disagrees	0.945406	0.051871	0.002723	0	disagree	neutral	agree
116	When liberals talk about the dire threat of global warming, liberalsâre actually seizing\nopportunistically on the issue.	agrees	disagrees	agrees	agrees	neutral	agrees	disagrees	neutral	0.439787	0.431305	0.128908	0	disagree	disagree	agree
154	But the erosion of science reaches well beyond the environment and climate.	disagrees	disagrees	disagrees	neutral	neutral	neutral	disagrees	neutral	0.665829	0.330386	0.003786	0	disagree	neutral	agree

Weighted BERT sub-category failures.