The dataset imbalance problem becomes worse once the unrelated data is excluded. The ratio of discuss/agree/disagree/ is roughly 11:5:1 so there is huge discrepancy here. As such, weighted sampling is going to be essential to achieving good results. Again, accuracy is not suitable as classifying all samples as discuss will give a high accuracy but will be useless.

Following on from 2.a.ii I also experiment with multiple possible models. However, this time I also experiment with different transformers to evaluate the effect of these. Initially, with BERT I train a CNN with a GRU, an MLP head, a LSTM, and a CNN. Without any changes to the sampling weights the results are:

[TABLE]

With log inverse weight sampling, the results become:

[TABLE]

By switching to Longformer, in the hope of dealing with the extended length of the article bodies better, I find, with inverse weight sampling:

[TABLE]