Hi Prof. My name is WangYifan . My work in our group is to use the Google Ber to classify the fake or real news. Here is my total pipeline. First, we got the raw text with title, text and label. Then we use some methods of data processing to operate the text. After the data processing, we put them into the Bert model to train the data, which includes the Bert itself and the Classifier, here I used the feed-forward neural network and add a softmax layer to normalize the output. In the end, we got the predication and other details.

AS for the first part, I use two methods: drop non-sentence and some EDA methods. I read some text within the fake\_or\_real news and I find that it contains various type of non-sentence, so I use the regular expression to drop them, here u can see. And then, I use random insertion and synonym replacement to augment the text. Here u can see.

AS for the second part, we put the text which we got from the first part into the bert model. The Bert model uses 12 encode layers and finally classifier to get the output. U can see from here.

In the end, we combine different methods of data processing and u can see the f1 score from the chart. We can the best f1 score from Cased text+ drop sentence.

Last but not least, Here are my references. I learn the EDA from the two web site and through two articles, I learn that we shouldn’t remove Stopwords which otherwise will destroy the context of sentence. The end is implementation of BERT with Pytorch. Sorry for no time to explain more. Thanks, Prof