Since BERT is a pre-trained model, only a small amount of data is needed to train the model to tune the parameters. When doing data split, we only divide the data into two parts: training set and validation/test set.

The optimal sample size of training set is 600, which is obtained from earlier experiments. The sentences selected for training are those with the same sentiment labeled by both human analysts.

Considering the amount of data, only all sentences with tokens are labeled by human. To examine the average sentiment at the article level, we sample 1000 sentences from the full text and then label the sentences without tokens manually.