Preprocessing steps:

I split the data into two sets based on the ‘identification’ attribute. Then I removed the duplicated data. After that, I noticed the size of training data was quite large. I sampled part of the data to perform the following steps to avoid wasting too much time training the model. Also, I used the label encoder to transformed the emotion into integers.

Feature Engineering

I used the TF-IDF transformation t convert raw text data into a sparse matrix. I also filtered out some stop words since the stop words were usually consider useless in analyzing the text.

Model

I tried using random forest, XGBClassifier to train the data. I also tried using decision tree to filter the important term in TF-IDF matrix. But the result was not that good. I guessed it was because the distribution of the labels was not fair. Maybe the terms that decision tree filtered were affected by the more frequent label. In the end, I used only XGBClassifer to train the dataset and predicted the final answer.