Assignment III
Code Mixed Sentiment Analysis

Manas Satish Bedmutha
16110031

Approach:

Data - Upon primary inspection, the tokens under lang_id = O are found to be majorly punctuations. For this very specific task, we choose to ignore them.

Model - Multilingual BERT + MLP
We use multilingual BERT with weights from tensrflow-hub. The embeddings for each tweet are calculated sing BERT. For the data preparation for BERT, in the (Input + A - B) section we set B to be empty to generate the training and testing examples.

This setup is then fine tuned to acclimatize to our dataset. Finally the 768-dimensional output embeddings are passed through an MLP with 512 and 256 nodes. This is then sent across an output layer that classifies our data. Since this is all within the bert tensorflow framework all training and optimization is done internally. Loss function used is categorical crossentropy.

Results:

Accuracy - 0.53044873
Precision - 0.8376
Recall - 0.7819268
F-score - 0.808806

Inspired from:
https://github.com/google-research/bert/blob/master/predicting_movie_reviews_with_bert_on_tf_hub.ipynb (Courtesy of Pratik Kayal)