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 tensorflow-hub. The embeddings for each tweet are calculated using BERT. Input samples are first loaded and processed with respect to a classification convertor of BERT to make the raw data trainable. These training and testing examples are then passed across the model.
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 cross entropy.

Training -
Batch size is set at 16; Initial Learning rate at 2e-5; Warmup Proportion = 0.2
(Note: Warmup is a period of time where the learning rate is small and gradually increases usually helps training.)

Results:

1) Overall Accuracy - 0.5518
2) Class-wise:

	Negative	Neutral	Positive
Precision	0.5111	0.5459	0.6117
Recall	0.6867	0.4398	0.5739
F - Score	0.5861	0.4871	0.5922

Code:
https://github.com/manasbedmutha98/CS613-Natural-Language-Processing/tree/master/Assignment%20III

Inspired from:
https://github.com/google-research/bert/blob/master/predicting_movie_reviews_with_bert_on_tf_hub.ipynb