NATURAL LANGUAGE PROCESSING

Dataset Prep:

Hindi HASOC(Hate speech and offensive content) dataset was downloaded
It includes monolingual text data.
The text was preprocessed by cleaning the text ,removing unnecessary character.

Code-Mix creation:

Initially I thought of using Google translate API but due to its limit requests I had to switch to Microsoft **Mtranslate**.
The code-mix sentences were created based on CMI indexed which means the proportion of code-mixed words in a sentence.
Iterate through the text and replace the translated word with the code-mixed version.
I have used CMI indexes as (0.1,0.3,0.5,0.9).
Also I have used a very compressed dataset just to show the working of model and its architecture.Also due to limited compute power of Google colab,I found it difficult to train on the entire dataset.

Fine-tuning Pre trained LM:

For training the labels were one-hot encoded and fed into the Language Models such as BeRT. to fine-tune on the code-mixed dataset.
Tokenize the code-mixed sentences using the tokenizer associated with the LM to convert them into tokens.
Then converting the tokens into input features compatible with the LM including attention masks.
All this is down using the Custom Dataset function.(separately for each code-mixed data)
The LM was fine tuned on the code-mixed dataset.
The hyperparameters were monitored.
The performance was evaluated.Although the values of evaluations weren't very appealing given the case was model was trained on very few instances and perhaps there is some scope of architecture improvement which I am not aware of.

I tried implementing the multi-lingual BeRT but faced some errors hence wasn't able to run it although the code remains same for it's just the model instantiation is different.

Note: I have referred to hugging face for some part of the code.