Report on Emotion Detection Task using Encoder only models

1. *Experiment 1:*

- **WandB Project Link** : RobertaProject
- **Model used**: RoBERTa (A Robustly Optimized BERT Pretraining Approach)
- **Highest F1-score achieved**: 0.5746
- **Number of Epochs**:2
- **Learning rate**:2e-5
- **Weight decay**: 0.1
- **Differences between BERT**:
  a. RoBERTa model is a variant of BERT designed specifically for sequence classification tasks, hence it is a very good choice for our experiment here.
  b. RoBERTa is a modification of BERT that aims to improve performance by optimizing the pretraining process.
  c. Similar to BERT, RoBERTa is based on the Transformer architecture with self-attention mechanisms. It includes an encoder with 12 layers and a classifier head for the classification output.
  d. RoBERTa omits the next sentence prediction objective, so it has no token type embeddings for differentiating sentence pairs, simplifying its pretraining tasks.
  e. It uses dynamic masking instead of static masking. In BERT, the same tokens are masked in every epoch, but in RoBERTa, different tokens are masked in each training step, leading to more varied training data and richer pretraining.
  f. RoBERTa is pretrained on much larger datasets than BERT and for longer period and can handle longer sequences, making it better suited for NLP tasks.
  g. Overall, RoBERTa enhances BERT to handle more complex language patterns, thereby providing stronger performance without altering the core architecture of BERT.
- **Potential improvements that can be done in future**: By looking at the training results, we should try to experiment with different weight decays and learning rates to obtain a more stable model. Because after step 400, the validation error increases which is not desirable. But since the experiment was aimed at focusing on f1-macro metric the checkpoint model is considered to be the best model.

2. ***Experiment 2:***

- **WandB Project Link** : DistilBERTProject
- **Model used**: **DistilBERT**
- **Highest F1-score achieved**: 0.5659
- **Number of Epochs**:2
- **Learning rate**:3e-5
- **Weight decay**: 0.01
- **Differences between BERT**:
    a. DistilBERT is a smaller and faster version of BERT created using a technique called knowledge distillation.
    b. DistilBERT uses fewer parameters than BERT by reducing the number of layers. Standard BERT-base has 12 layers (transformer blocks), whereas DistilBERT has only 6. This reduction makes DistilBERT faster to train and more efficient.
    c. In DistilBERT's case, BERT is the teacher, and DistilBERT is the student mimicking the teacher model(larger model).
    d. Despite its smaller size, DistilBERT retains most of BERT's language understanding capabilities while being about 60% faster and using significantly less memory. This performance retention is achieved by carefully optimizing the training process.
- **Differences between RoBERTa**:
    a. DistilBERT has only 6 layers while Roberta has 12 base layers same as BERT.
    b. DistilBERT is faster and lighter than RoBERTa.
    c. RoBERTa performs better and understands more nuanced contexts when compared to DistilBERT
    d. RoBERTa-base has 125 million hyperparameters where as DistilBERT has only 66 million.
    e. RoBERTa is resource intensive and hence slower than DistilBERT.
    f. DistilBERT is preferred when there are limited memory and other resource constrains where as RoBERTa is prefreed for high accuracy tasks.
    g. RoBERTa uses dynamic masking where as DistilBERT uses knowledge distillation for training.
- By looking at the output we can see a steady decrease in training and validation loss across epochs which is a good sign, And as the differences between RoBERTa suggests, the performance in terms of metric is slightly lower in DistilBERT than RoBERTa.

3. ***Experiment 3:***

- **WandB Project Link** : DistilRoBERTaProject
- **Model used**: **DistilRoBERTa-base**
- **Highest F1-score achieved**: 0.5492
- **Number of Epochs**:3
- **Learning rate**:2e-5
- **Weight decay**: 0.01
- **Differences from the previous models**:
    a. DistilRoBERTa-base is a smaller, lighter and faster version of RoBERTa, combining the efficiency gains of DistilBERT .
    b. It has 6 Transformer layer**s** instead of the 12 layers in RoBERTa-base, similar to DistilBERT.
    c. It has slightly higher hyperparameters(88million) compared to DistilBERT and lower compared to RoBERTa.
    d. DistilRoBERTa retains almost 95% of RoBERTa's performance, while being 60% faster than RoBERTa.
    e. DistilRoBERTa gives similar speed to DistilBERT but higher accuracy than DistilBERT.
    f. It is used where both accuracy and resource constraints are equally important.
- **Potential Improvements**: In theory, DistilRoberta is supposed to give a higher accuracy when compared to DistilBERT. But in this case we can see that the evaluation metric f1 is lower compared to DistilBERT. It can be because of slight differences in the hyperparameters being used. While though the f1 score is lower, DistilRoberta_base has provided a much stable model and has run 3 epochs almost with the same time as DistilBERT.

- **Conclusion**: By looking at the metrics, we can say that RoBERTa model performs the best, because its closer in architecture to BERT and hence has in depth understanding of the language and contexts. Further experiments can be done to tune the model with different parameters, take steps to clean the data before tokenization and see if the performance can be increased.