Question 5 Report

Hyperparameters

1 Learning Rate

The learning rate is one of the most critical hyperparameters in deep learning training. It determines the step size during optimization and significantly impacts the convergence speed and final performance of the model. Usually values between 1e-5 and 5e-5 are chosen for training models. Since the notebook provided by the authors of IndicNER were using 5e-5, and we were tuning a model for that same task.
So, I selected ***5e-5*** as my learning rate. It worked out quite well for both IndicBERT and IndicNER.

2 Batch Size

Batch size determines the number of samples processed in each forward and backward pass during training. It affects both memory usage and computational efficiency, making it an essential hyperparameter to consider. Since for training, I had access to limited colab GPU time and memory, it was a major constraint while considering the batch sizes. Too low of a batch size and the computation would take too long, or too high of a batch size and the GPU would run out of memory. So for min-maxing, I experimented with different batch sizes to optimize training performance while trying to finish all epochs in the given time limit.

A batch size of ***8*** for the validation data and ***40*** for the training data while tuning the IndicNER model.
A batch size of ***8*** for the validation data and ***50*** for the training data while tuning the IndicBERT model.

3 Weight Decay

Weight decay is a regularization technique used to prevent overfitting by penalizing large weights in the model. It adds a regularization term to the loss function, encouraging the model to learn simpler patterns and improve generalization. Initially, the weight_decay was not utilized but after introducing it to the model, the loss was slightly lower (not very significant in this case).
In our experiments, a weight decay value of ***1e-6*** was used considering our learning rate of 5e-5.

4 Number of Epochs

The number of epochs defines the number of times the entire training dataset is passed through the model during training. It determines the duration of training and plays a crucial role in model convergence and generalization. Taking a higher value of number_of_epochs would have been ideal but considering our computational limitations, it was not reasonable in this case.
A minimum of 5 epochs were mandatory in this assignment, so ***5*** epochs were done for each of the 2 models.

5 Datasize

Not really a hyper-parameter, but the data size had to be reduced to finish training in a reasonable time. Initially the training data size was of length 455248, out of which ***100000*** were randomly selected with a seed of 53.
The data sampled is the same for both of the models.

Results

For question 2, outputs for the models are as follows:

```
IndicNER

Over all the predicted words

TRAINING
MACRO-f1  : 0.937799
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC
Precision : 0.989390 : 0.956916 : 0.949416 : 0.930479 : 0.925118 : 0.927534 : 0.873502
Recall    : 0.988391 : 0.954952 : 0.958131 : 0.901225 : 0.941647 : 0.944803 : 0.888624
f1        : 0.988890 : 0.955933 : 0.953754 : 0.915618 : 0.933309 : 0.936089 : 0.880998

TESTING
MACRO-f1  : 0.794872
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC
Precision : 0.961146 : 0.922727 : 0.918216 : 0.715054 : 0.591603 : 0.880539 : 0.614035
Recall    : 0.972012 : 0.878788 : 0.903108 : 0.696335 : 0.541958 : 0.807420 : 0.744681
f1        : 0.966548 : 0.900222 : 0.910599 : 0.705570 : 0.565693 : 0.842396 : 0.673077

VALIDATION
MACRO-f1  : 0.815108
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC
Precision : 0.966025 : 0.869138 : 0.893069 : 0.762515 : 0.714050 : 0.828881 : 0.682310
Recall    : 0.966938 : 0.880177 : 0.860687 : 0.765187 : 0.663594 : 0.864983 : 0.697417
f1        : 0.966481 : 0.874623 : 0.876579 : 0.763848 : 0.687898 : 0.846547 : 0.689781
```

Results for IndicNER (with class-wise precision, recall, f1 too)

```
IndicBERT

Over all the predicted words

TRAINING
MACRO-f1  : 0.851864
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC
Precision : 0.970742 : 0.893559 : 0.910150 : 0.843817 : 0.844869 : 0.860210 : 0.776566
Recall    : 0.977942 : 0.898301 : 0.883055 : 0.800507 : 0.744588 : 0.892464 : 0.649215
f1        : 0.974329 : 0.895924 : 0.896398 : 0.821592 : 0.791565 : 0.876040 : 0.707203

TESTING
MACRO-f1  : 0.787586
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC
Precision : 0.956119 : 0.876106 : 0.916045 : 0.785075 : 0.637681 : 0.862934 : 0.614583
Recall    : 0.972695 : 0.853448 : 0.894353 : 0.674359 : 0.584718 : 0.785589 : 0.627660
f1        : 0.964336 : 0.864629 : 0.905069 : 0.725517 : 0.610052 : 0.822447 : 0.621053

VALIDATION
MACRO-f1  : 0.763666
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC
Precision : 0.956037 : 0.843084 : 0.872354 : 0.705747 : 0.641541 : 0.788129 : 0.608163
Recall    : 0.960516 : 0.849119 : 0.836398 : 0.707373 : 0.565731 : 0.825561 : 0.541818
f1        : 0.958271 : 0.846091 : 0.853998 : 0.706559 : 0.601256 : 0.806411 : 0.573077
```

Results for IndicBERT (with class-wise precision, recall, f1 too)

We can see that the Macro-F1 scores for the testing data are very close for both the models.
Macro-F1 score for testing, training and validation(IndicBERT) : 0.788, 0.852, 0.764
Macro-F1 score for testing, training and validation(IndicNER) : 0.795, 0.938, 0.815

Note that for question 2, only seven labels are used (i.e. excluding MISC) as there is no MISC label in the training data.
But for question 4, all nine labels are used to calculate the metrics

For question 4, the results are as follows:

```
BERT Model
MACRO-f1  : 0.505228
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC  :  B-MISC  :  I-MISC
Precision : 0.811083 : 0.600000 : 0.750000 : 0.684211 : 0.652174 : 0.500000 : 0.250000 : 0.000000 : 0.000000
Recall    : 0.952663 : 0.900000 : 0.750000 : 0.565217 : 0.714286 : 0.500000 : 1.000000 : 0.000000 : 0.000000
f1        : 0.876190 : 0.720000 : 0.750000 : 0.619048 : 0.681818 : 0.500000 : 0.400000 : 0.000000 : 0.000000
Prec_all  : 0.781513
Recall_all: 0.781513
f1_all    : 0.781513
```

Manual vs IndicBERT

Overall precision, recall and macro-f1 are all 0.781 when calculated over all sequences.

```
GPT
MACRO-f1  : 0.205683
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC  :  B-MISC  :  I-MISC
Precision : 0.755556 : 0.571429 : 0.500000 : 0.000000 : 0.000000 : 0.000000 : 0.000000 : 1.000000 : 1.000000
Recall    : 0.984211 : 0.400000 : 0.333333 : 0.000000 : 0.000000 : 0.000000 : 0.000000 : 0.040000 : 0.025000
f1        : 0.854857 : 0.470588 : 0.400000 : 0.000000 : 0.000000 : 0.000000 : 0.000000 : 0.076923 : 0.048780
Prec_all  : 0.741313
Recall_all: 0.741313
f1_all    : 0.741313
```

Manual vs ChatGPT

Overall precision, recall and macro-f1 are all 0.741 when calculated over all sequences.

```
NER Model
MACRO-f1  : 0.591404
Metric    :     O    :   B-PER  :   I-PER  :   B-ORG  :   I-ORG  :   B-LOC  :   I-LOC  :  B-MISC  :  I-MISC
Precision : 0.825991 : 0.750000 : 1.000000 : 0.857143 : 0.611111 : 0.714286 : 0.500000 : 0.000000 : 0.000000
Recall    : 0.986842 : 0.900000 : 0.916667 : 0.521739 : 0.523810 : 0.833333 : 1.000000 : 0.000000 : 0.000000
f1        : 0.899281 : 0.818182 : 0.956522 : 0.648649 : 0.564103 : 0.769231 : 0.666667 : 0.000000 : 0.000000
Prec_all  : 0.818533
Recall_all: 0.818533
f1_all    : 0.818533
```

Manual vs IndicNER

Overall precision, recall and macro-f1 are all 0.818 when calculated over all sequences.