CS689A

Computational Linguistics for Indian Languages

Detailed report

Neelu Lalchandani
231110031

1 Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP), aiming to identify and classify named entities such as persons, locations, organizations, and miscellaneous entities within text data. In this report, I have presented the results of fine-tuning two pre-trained models, IndicNER and IndicBERT, on the Naamapadam corpus—a dataset in my mother tongue-HINDI. The objective is to evaluate their performance in NER tasks specific Hindi language.

2 Output of both models

[4929/4929 1:05:50, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Loc Precision	Loc Recall	Loc F1	Loc Number	Org Precision	Org Recall	Org F1	Org Number	Per Precision	Per Recall	Per F1	Per Number	Overall Precision	Overall Recall	Overall F1	Overall Accuracy
1	0.318200	0.308777	0.727424	0.598649	0.656784	10213	0.478011	0.518700	0.497525	9786	0.648397	0.658213	0.653268	10568	0.609909	0.593647	0.601668	0.905673
2	0.264900	0.283320	0.720779	0.670812	0.694898	10213	0.524003	0.539853	0.531810	9786	0.660730	0.698618	0.679146	10568	0.634472	0.638499	0.636479	0.912282
3	0.230600	0.277866	0.702610	0.703711	0.703160	10213	0.553387	0.524320	0.538462	9786	0.691757	0.685276	0.688501	10568	0.652653	0.639906	0.646216	0.915135

Figure 1: indicBERT evaluation metrics

```
Macro-F1 score for training and validation dataset, Epoch 1 = 0.601668
Macro-F1 score for training and validation dataset, Epoch 2 = 0.636479
Macro-F1 score for training and validation dataset, Epoch 3 = 0.646216
```

[5031/5031 1:11:21, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Loc Precision	Loc Recall	Loc F1	Loc Number	Org Precision	Org Recall	Org F1	Org Number	Per Precision	Per Recall	Per F1	Per Number	Overall Precision	Overall Recall	Overall F1	Overall Accuracy
1	0.151700	0.174329	0.811946	0.861157	0.835828	10213	0.686357	0.698140	0.692199	9786	0.801242	0.842638	0.821419	10568	0.769028	0.802565	0.785439	0.948370
2	0.118400	0.179417	0.802678	0.862724	0.831619	10213	0.672565	0.699162	0.685605	9786	0.812045	0.833176	0.822474	10568	0.764480	0.800144	0.781905	0.947292
3	0.090700	0.195268	0.811352	0.852345	0.831344	10213	0.677690	0.693133	0.685325	9786	0.807854	0.835068	0.821236	10568	0.767892	0.795400	0.781404	0.947059

Figure 2: indicNER evaluation metrics

```
Macro-F1 score for training and validation dataset, Epoch 1 = 0.785439
Macro-F1 score for training and validation dataset, Epoch 2 = 0.781905
Macro-F1 score for training and validation dataset, Epoch 3 = 0.781404
```

3 Evaluation metrics on test dataset

```
Macro-F1 score on test dataset = 0.6782473424941665
```

```
Macro-F1 score on test dataset = 0.8052854649713289
```

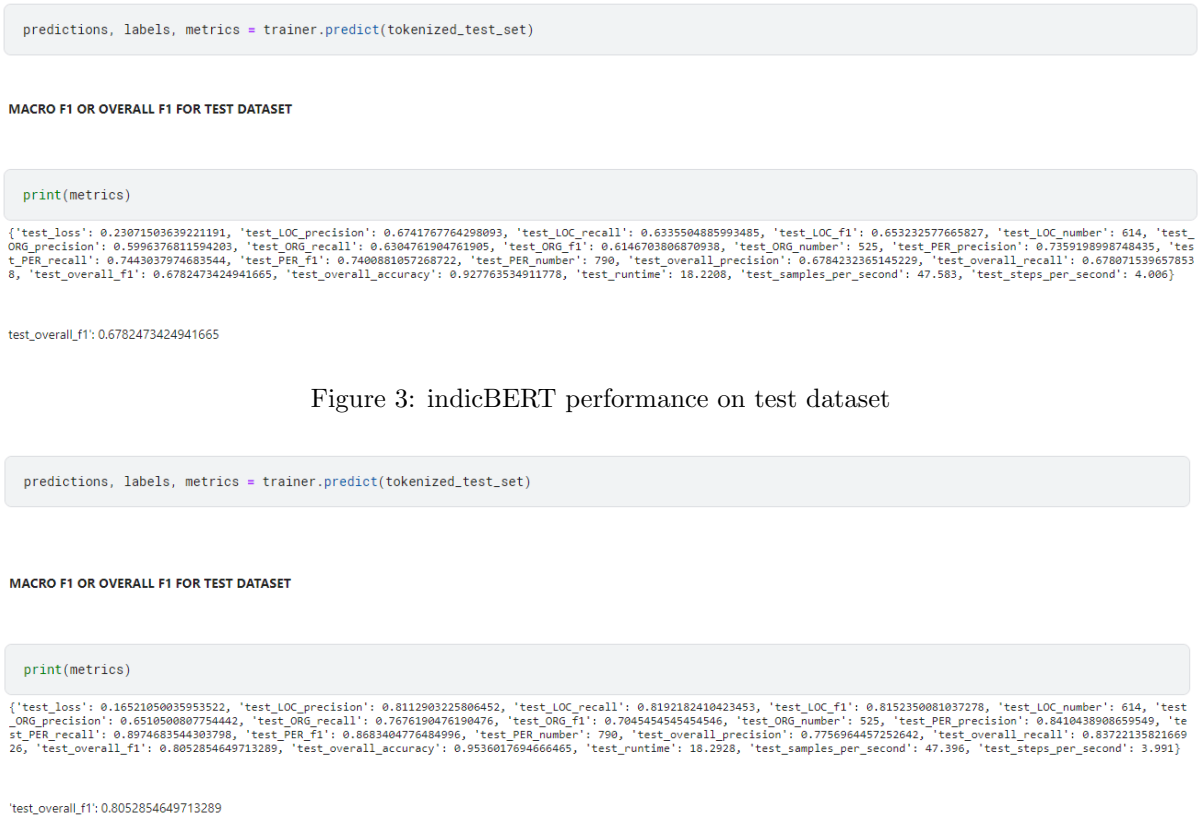

Figure 3: indicBERT performance on test dataset

Figure 4: indicNER performance on test dataset

4 Comparison of both models

Training and Validation Dataset

- IndicNER consistently outperforms IndicBERT across all epochs. IndicNER achieves higher macro-F1 scores, indicating better overall performance in recognizing named entities during training and validation.
- IndicBERT shows improvement from epoch to epoch, but its performance remains lower compared to IndicNER.

Test Dataset

- IndicNER also outperforms IndicBERT on the test dataset, with a higher macro-F1 score. This suggests that IndicNER's superior performance observed during training and validation extends to unseen data, indicating its effectiveness and robustness in named entity recognition tasks.

In summary, based on the provided macro-F1 scores, IndicNER demonstrates superior performance over IndicBERT for named entity recognition tasks on Hindi language.

5 Evaluation metrics of chatgpt on 25 sentences

```
Class: O
Precision: 0.9336823734729494
Recall: 0.9870848708487084
F1-score: 0.9596412556053812

Class: B-PER
Precision: 0.7222222222222222
Recall: 0.7222222222222222
```

F1-score: 0.7222222222222222

Class: I-PER
Precision: 0.5882352941176471
Recall: 0.625
F1-score: 0.6060606060606061

Class: B-LOC
Precision: 0.8571428571428571
Recall: 0.8571428571428571
F1-score: 0.8571428571428571

Class: I-LOC
Precision: 0
Recall: 0
F1-score: 0

Class: B-ORG
Precision: 1.0
Recall: 0.8888888888888888
F1-score: 0.9411764705882353

Class: I-ORG
Precision: 0.8571428571428571
Recall: 1.0
F1-score: 0.923076923076923

Class: B-MISC
Precision: 0.25
Recall: 0.045454545454545456
F1-score: 0.07692307692307693

Class: I-MISC
Precision: 0.5
Recall: 0.15
F1-score: 0.23076923076923075

Macro F1-score: 0.5907791824876147
Accuracy: 0.9088098918083463

6 Evaluation metrics of indicBERT on 25 sentences

Metrics for O:
Precision: 0.9026
Recall: 1.0000
F1 Score: 0.9488

Metrics for B-PER:
Precision: 0.5882
Recall: 0.6250
F1 Score: 0.6061

Metrics for I-PER:
Precision: 0.9231
Recall: 0.7500
F1 Score: 0.8276

Metrics for B-LOC:

```
Precision: 0.9091
Recall: 0.7692
F1 Score: 0.8333

Metrics for I-LOC:
Precision: 0.0000
Recall: 0.0000
F1 Score: 0.0000

Metrics for B-ORG:
Precision: 0.7143
Recall: 0.5556
F1 Score: 0.6250

Metrics for I-ORG:
Precision: 0.5000
Recall: 0.3333
F1 Score: 0.4000

Metrics for B-MISC:
Precision: 0.0000
Recall: 0.0000
F1 Score: 0.0000

Metrics for I-MISC:
Precision: 0.0000
Recall: 0.0000
F1 Score: 0.0000

Accuracy: 0.8893
Macro F1 Score: 0.4712
```

7 Evaluation metrics of indicNER on 25 sentences

```
Metrics for O:
Precision: 0.9241
Recall: 0.9878
F1 Score: 0.9549

Metrics for B-PER:
Precision: 0.7000
Recall: 0.7778
F1 Score: 0.7368

Metrics for I-PER:
Precision: 0.7000
Recall: 1.0000
F1 Score: 0.8235

Metrics for B-LOC:
Precision: 0.7333
Recall: 0.6111
F1 Score: 0.6667

Metrics for I-LOC:
Precision: 0.0000
Recall: 0.0000
F1 Score: 0.0000
```

```
Metrics for B-ORG:
Precision: 0.5455
Recall: 0.6000
F1 Score: 0.5714

Metrics for I-ORG:
Precision: 1.0000
Recall: 0.5000
F1 Score: 0.6667

Metrics for B-MISC:
Precision: 0.0000
Recall: 0.0000
F1 Score: 0.0000

Metrics for I-MISC:
Precision: 0.0000
Recall: 0.0000
F1 Score: 0.0000

Accuracy: 0.8978
Macro F1 Score: 0.4911
```

8 Comparison of both models with chatgpt

In comparing the performance of ChatGPT, IndicBERT, and IndicNER for named entity recognition tasks, several key observations emerge.

- CHATGPT: This model demonstrates strong precision and recall for the "O" class, indicating its ability to accurately identify non-named entities. However, its performance varies across named entity classes, with lower precision, recall, and F1-scores observed for most classes compared to IndicBERT and IndicNER. Notably, ChatGPT provides metrics for miscellaneous entities (MISC), which are not classified by IndicBERT and IndicNER. Overall, ChatGPT achieves the highest macro F1-score of 0.5908, showcasing its effectiveness in named entity recognition tasks.

- indicBERT: While IndicBERT also exhibits high precision and recall for the "O" class, its performance for named entity classes varies. Despite achieving moderate precision and recall for some classes, such as "I-PER", "B-LOC", and "B-ORG", IndicBERT falls short in achieving comparable performance to ChatGPT, particularly in classifying miscellaneous entities (MISC). This limitation contributes to its lower overall macro F1-score of 0.4712.

- indicNER: Similar to ChatGPT and IndicBERT, IndicNER demonstrates high precision and recall for the "O" class. Its performance for named entity classes exhibits variability, with certain classes, such as "I-PER" and "I-ORG", achieving relatively higher precision and recall compared to others. However, IndicNER also does not classify miscellaneous entities (MISC), contributing to its lower overall macro F1-score of 0.4911 compared to ChatGPT.

In summary, both IndicBERT and IndicNER show promising performance in certain aspects of named entity recognition tasks, such as precision and recall for specific classes. However, their inability to classify miscellaneous entities (MISC) limits their overall effectiveness compared to ChatGPT, which provides metrics for all named entity classes

9 Hyperparameters Tuned

- **Batch Size**
  The batch size, an essential parameter in training neural networks, dictates the number of samples processed in each training iteration. Balances computational efficiency and model generalization.

  i)Batch size:8
  When I tried to train my NER model on batch size 8 and 2 epochs, more than half of my output space on Kaggle got full on 0.2/2 epochs only, so I had to stop the training.

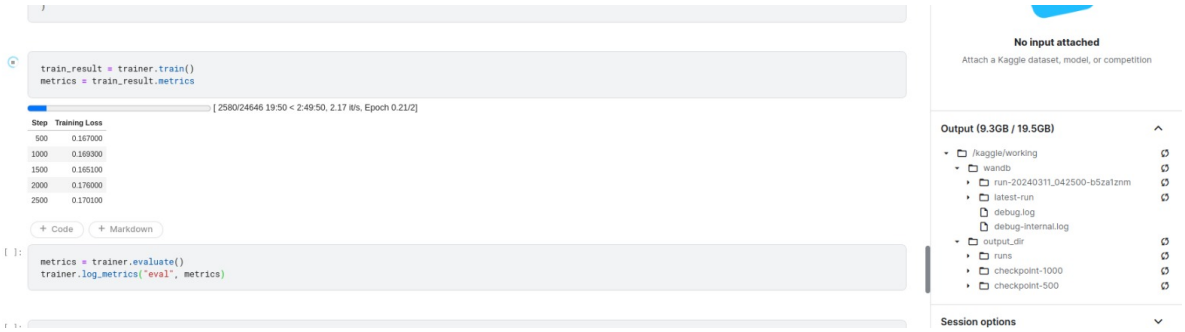

Figure 5: training on batch size=8 and 2 epochs

I got accuracy 0.88 on indicBERT on batch size 8.

ii)Batch size:12
I was able to train my model successfully and got accuracy of 0.915135 on indicBERT and 0.947059 on indicNER on batch size 12.

iii)Batch size:16
On batch size 16, accuracy of models decreased, maybe because of overfitting. Hence I decided to not increase batch size further.

CHOSEN OPTIMAL VALUE OF BATCH SIZE = 12

- **Learning Rate**
  The learning rate is one of the most critical hyperparameters in training machine learning models, including neural networks. Its significance lies in its crucial role in determining the size of the steps taken during optimization, influencing the trajectory of the training process and ultimately impacting the performance and convergence of the model.

  i) Learning Rate:0.00002
  On using learning rate = 2e-5, I got accuracy of 0.915135 on indicBERT and 0.947059 on indicNER on batch size 12.

  ii) Learning Rate:0.00003
  On using learning rate = 3e-5, accuracy of both models decreased.

CHOSEN OPTIMAL VALUE OF LEARNING RATE = 0.00002

- **Number of epochs**
  The number of epochs determines how many times a model iterates through the training dataset, impacting its convergence and performance. More epochs can improve model accuracy but may lead to overfitting if not balanced. However, too few epochs may result in underfitting, compromising model effectiveness. Therefore, selecting an optimal number of epochs is crucial for achieving

a balance between model performance and computational efficiency.

i) Number of epochs:3
On number of epochs=3, batch size=12 and learning rate=0.00002, I got accuracy of 0.915135 on indicBERT and 0.947059 on indicNER.

ii) Number of epochs:2
On number of epochs=3, batch size=12 and learning rate=0.00002, I got accuracy of 0.886371 on indicBERT and 0.902431 on indicNER.

CHOSEN OPTIMAL VALUE OF BATCH SIZE = 12

10 Output of both models

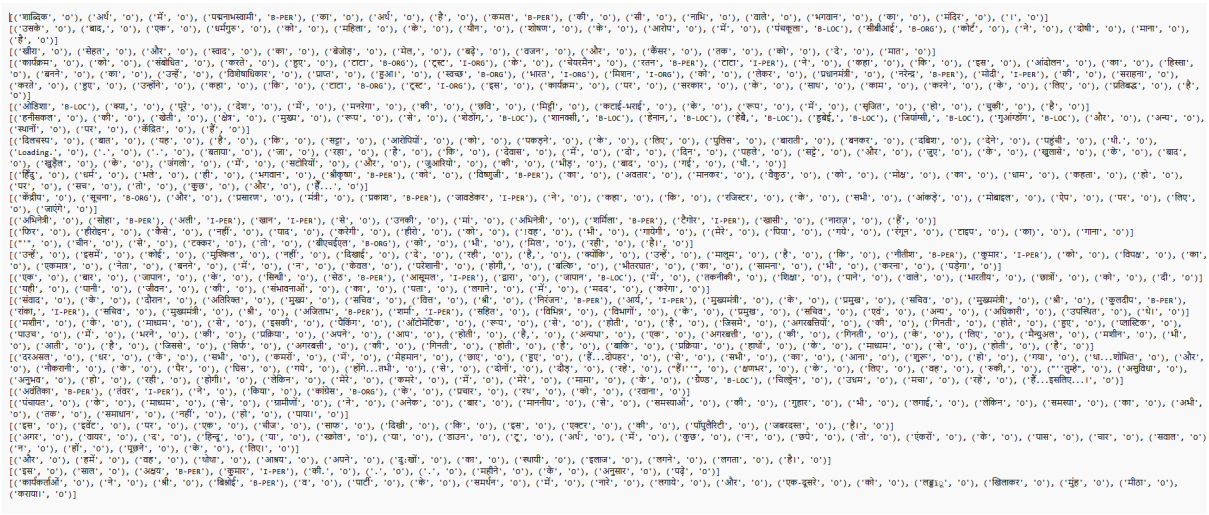

Figure 6: indicBERT

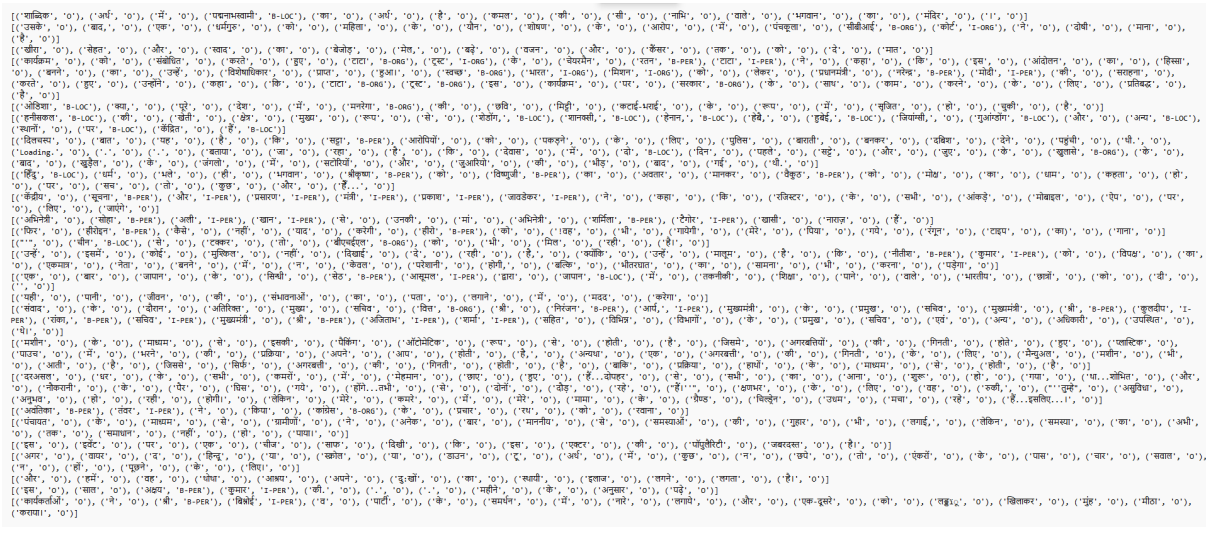

Figure 7: indicNER

11 Conclusion

- **Model Performance Comparison:**
  IndicBERT and IndicNER showed varying performance on the named entity recognition task, with IndicNER exhibiting higher accuracy overall.

- **ChatGPT's Role:**
  ChatGPT performed better than NER models because it also considered classes B-MISC and I-MISC, unlike NER models. If chatGPT also wouldn't have considered MISC classes, then specialized NER models would have given better metrics.

- **Hyperparameter Significance:**
  The significance of hyperparameters such as learning rate and number of epochs was evident in their impact on training dynamics, convergence, and model performance.