Appendix 1: Hyperparameter Tuning: Fresh Training of BERT-based LLM with Synthetic Data

We also experimented with several variations of the LLM training hyperparameters during the fresh training of BERT with synthetic data. The tuning experiments were conducted with train, validation and test split of 80%, 10%, 10% and test data evaluation on 100 test points was kept fixed.

Exp ID	Learning Rate	Batch Size	Epochs	Eval Accuracy	Eval Precision	Eval Recall	Eval F1
1	2e-5	8	5	0.9991	0.9897	0.9960	0.9928
2	2e-5	16	5	0.9991	0.9892	0.9964	0.9928
3	**2e-5**	**32**	**5**	**0.9992**	**0.9892**	**0.9967**	**0.9929**
3	**2e-5**	**32**	**5**	**0.9992**	**0.9892**	**0.9967**	**0.9929**
4	2e-5	32	7	0.9992	0.9892	0.9967	0.9929
5	2e-5	32	10	0.9992	0.9899	0.9960	0.9929
3	**2e-5**	**32**	**5**	**0.9992**	**0.9892**	**0.9967**	**0.9929**
6	1e-5	32	5	0.9991	0.9895	0.9962	0.9928
7	4e-5	32	5	0.9991	0.9892	0.9964	0.9928
8	2e-5 linear	32	5	0.9991	0.9890	0.9967	0.9928
9	2e-5 cosine	32	5	0.9991	0.9890	0.9967	0.9928

Observations and Analysis - There was very small difference in the F1 score in all the experiments. Batch size 32 achieved the best F1 scores (0.9929) outperforming batch sizes of 8 and 16. Training for 5, 7 and 10 epochs resulted in the same F1 score of 0.9929, selecting 5 Epochs as it is faster. Learning rates of 1e-5 and 4e-5, along with dynamic learning rates both linear and cosine had the same F1 score of 0.9928. The fixed learning rate of 2e-5 was marginally better with an F1 score of 0.9929.

Best Performing Configuration - Learning Rate: 2e-5 (fixed learning rate) Batch Size: 32 Epochs: 5

Appendix 2: Aligning labels with Tokens

Challenges with KindLab Implementation on HuggingFace - The Huggingface models available as part of KindLab did not have a Tokenizer file along with it. Hence, the tokenizer of the base classifier, namely, Bert, Roberta and Distilbert were used.

```
tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")
```

One challenge with using the tokenizer of the base classifier relates to the alignment of labels. The list of label information was available in the Config file.

```
label_list = ["O", "AGE", "CONTACT", "DATE", "ID", "LOCATION", "NAME", "PROFESSION"]
```

We need to tokenize text and align character-based entity labels with the corresponding tokens produced by BERT tokenizer. For token-level classification tasks like Named Entity Recognition (NER), each token needs a corresponding label. Initially, every token is labeled as "O" (which means "Outside" in BIO tagging—no entity).

```
labels_aligned = ["O"] * len(tokenized_inputs["input_ids"])
```

And to align Entities to Tokens:

- Loop through each labeled entity (start, end, label).
- For each token (looping with enumerate(offset_mapping)):
    - It checks **if the token overlaps** with the entity span (character ranges):

  if offset_end > start and offset_start < end:

        - If yes, it **updates the token's label** from "O" to the actual entity label (e.g., "NAME").
        - Spaces in the label (if any) are replaced with underscores.

Returns a dictionary with:

- input_ids: the token IDs for the text.
- attention_mask: the mask showing which tokens are padding vs. real text.
- labels: a list of label IDs aligned with the tokens (using a label2id dictionary to map label names to numeric IDs, defaulting to 0 if not found).

Appendix 3: Google Colab Notebooks and Saved Models

Dataset

The synthetic dataset consists of 2 JSON files - annotations.json and discharge_summaries.json
The discharge summaries file consists of a document_id and the text of the medical record.
The annotations file contains the annotation_id, start_index, stop_index, entity, entity_type, and the document_id.
<u>Dataset Statistics:</u>

Metric	Value
Number of Documents	1000
Number of Annotations	9000
Average Annotations per Document	9.0
Average Document Length (words)	205.9

Entity Type	Count
NAME	2000
AGE	1000
DATE	3000
IDNUM	1000
LOCATION	1000
PHONE	1000

"text": "Name: Ashley Wolfe   Unit No: 1110277\nAdmission Date: 23/08/2023   Discharge Date: 25/08/2023\nDate of Birth: 13/07/2009   Age: 57   Sex: F\nService: Paediatrics\nAttending: Paula Sutton\n\nChief Complaint: Patient presented with complaints relevant to paediatrics evaluation.\n\nHistory of Present Illness:\nAshley Wolfe, a 57-year-old logistics and distribution manager from Robertburgh, KS, was admitted with a several-day history of symptoms requiring paediatrics evaluation.\nSymptoms were progressive and included complex features related to the underlying condition. On evaluation, the patient reported symptom details including variations, associated features, and previous management attempts.\n\nPhysical Examination and Diagnostic Findings:\nThorough physical evaluation was conducted, revealing relevant signs supportive of the working diagnosis.\nLaboratory studies, imaging (e.g., CT/MRI/Ultrasound/X-ray), and specialty consultations were pursued to refine diagnosis.\n\nDiagnosis:\nBased on history and investigations, the final diagnosis was determined as a condition commonly managed in paediatrics with consideration of differential diagnoses.\n\nTreatment and Hospital Course:\nThe patient underwent appropriate pharmacologic and/or procedural interventions. Response to treatment was closely monitored.\nPain control, supportive therapy, and targeted interventions were used. The patient remained hemodynamically stable throughout.\n\nDischarge Plan and Follow-Up:\nPatient was discharged in stable condition. Advised to continue medications and follow up at the outpatient paediatrics clinic.\nFollow-up appointment scheduled at: Robertburgh, KS clinic. Contact: 053-606-5681x5409."

{   "annotation_id": "T1",   "start": 6,   "stop": 18,   "entity": "Ashley Wolfe",   "entity_type": "NAME",   "document_id": "ex_ds_1" },
{   "annotation_id": "T2",   "start": 130,   "stop": 132,   "entity": "57",   "entity_type": "AGE",   "document_id": "ex_ds_1" }

<u>Fine Tuning Pre-trained Models and Fresh training:</u>

We explored two approaches:

a. Fine tune the three pre-trained models available along with original paper, with the synthetic data and measure the performance.
b. Train the three base models BERT, RoBERTa, DistilBERT with the synthetic data and measure the performance.

Each of the 3 model files contain both fine tuning of the pretrained model and fresh training with the synthetic data.

1. Bert_KindLab_Predict_FinetuneLLM.ipynb
2. Roberta_KindLab_Predict_FinetuneLLM.ipynb
3. Distilbert_KindLab_Predict_FinetuneLLM.ipynb

2 additional LLM models were trained as an extension of this project. Since there were no pretrained models, only fresh training with the synthetic data was performed.

4. Electra_LLM.ipynb
5. Deberta_LLM.ipynb

Hyperparameter Tuning for BERT

We experimented with several variations of the LLM training hyperparameters during the finetuning of the pre-trained model KindLab/bert-deid. In the hyperparameter tuning, we explored the impact of varying learning rates, batch sizes, and training epochs on evaluation metrics such as precision, recall, F1-score, and accuracy.

1. Hyperparameter_Fresh_BertLLM.ipynb- This file contains the hyperparameter tuning done for the fresh training with synthetic data set.
2. Hyperparameter_KindLabFinetune_Bert.ipynb - This file contains the hyperparameter tuning done for the fine tuning of the pre-trained models with synthetic data set.
3. Bert_CPU_Training_Fresh.ipynb- This file contains the fine tuning of the BERT model with CPU as the runtime environment.

Saved Models

The fresh training with the synthetic data set are saved in the 5 folders below

bert_base_cased_finetuned

deberta_base_cased_finetuned

distilbert_base_finetuned

Electra_base_cased_finetuned

Roberta_base_finetuned

The models obtained after fine-tuning the pretrained models with the synthetic dataset are saved in the 3 folders below

kindlab_bert_deid_finetuned

kindlab_distilbert_deid_finetuned

kindlab_roberta_deid_finetuned