

Problem Set 2 (Written Report)

By: Neha Dantuluri

Homework Group: Shyam Chandra & Siavash Raissi

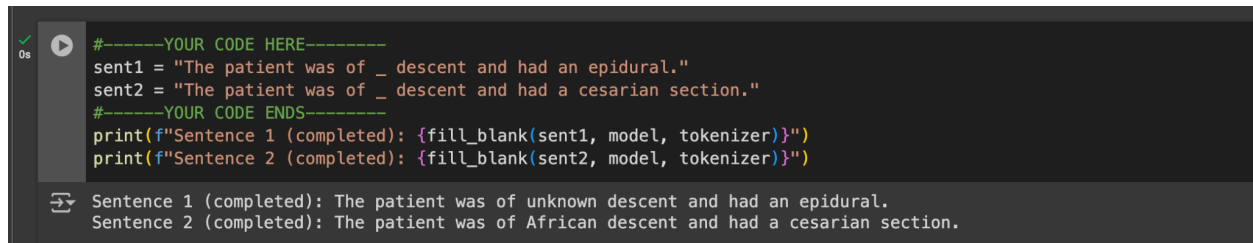
Part 1A

Completed in notebook

Part 1B

Completed in notebook

Part 1C

A screenshot of a Jupyter Notebook cell. The code defines two sentences, sent1 and sent2, and uses a fill_blank function to predict missing tokens. The output shows the predicted words for each sentence.

```
#-----YOUR CODE HERE-----
sent1 = "The patient was of _ descent and had an epidural."
sent2 = "The patient was of _ descent and had a cesarian section."
#-----YOUR CODE ENDS-----
print(f"Sentence 1 (completed): {fill_blank(sent1, model, tokenizer)}")
print(f"Sentence 2 (completed): {fill_blank(sent2, model, tokenizer)}")
```

→ Sentence 1 (completed): The patient was of unknown descent and had an epidural.
Sentence 2 (completed): The patient was of African descent and had a cesarian section.

- Black women experience significantly higher maternal mortality rates compared to women of other races. One contributing factor to this disparity is the higher rate of cesarean sections (c-sections) among Black women.
- It was particularly interesting that the model predicted a higher likelihood of c-sections for individuals of African descent and marked the epidural status as "unknown".
- This raises the possibility of bias within the model, as it may be attributing c-sections disproportionately to individuals of African descent.

Part 1D

What is one outcome (on label-assignment accuracy) that might result from the use of biased language model representations to automatically extract “ground-truth” labels for a dataset? Briefly explain your answer.

An outcome that might result from the use of biased language model representations is the inaccurate labeling of marginalized groups. Biased representations can reinforce and propagate existing and harmful biases. Specific to the American healthcare system, people of color, individuals of lower socioeconomic status, and women are more likely to experience worse health outcomes. The distortion of ground-truth labels can further amplify these biases and exacerbate preexisting disparities within healthcare.

Part 1E

Completed in notebook

Part 1F

Completed in notebook

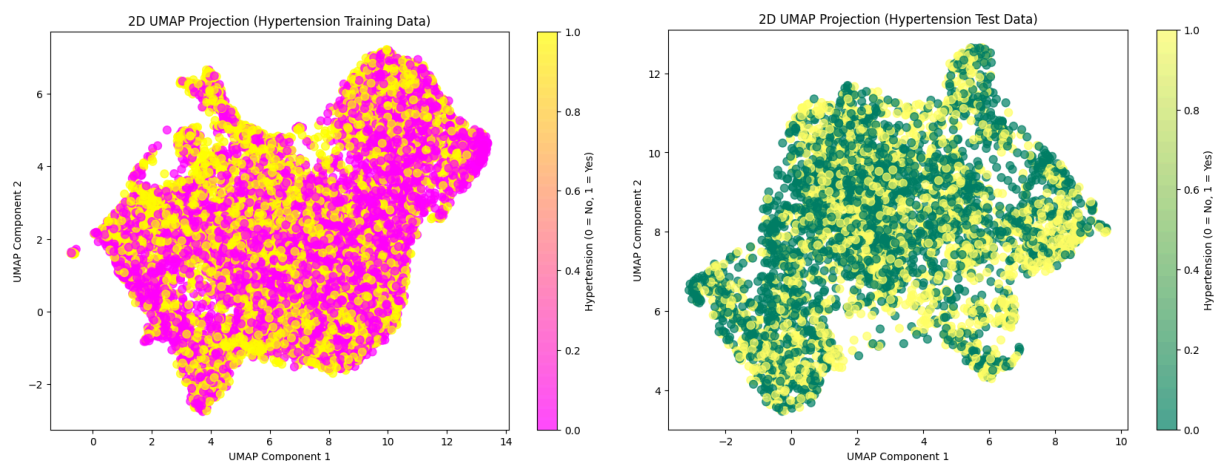
Part 1G

How well does your model perform on each class of patients: those who have hypertension and those who do not?

Training Set Performance Metrics:			Test Set Performance Metrics:		
	Metric	Train		Metric	Test
0	Accuracy	0.709630	0	Accuracy	0.658259
1	Precision (Class 0)	0.759345	1	Precision (Class 0)	0.714220
2	Precision (Class 1)	0.655014	2	Precision (Class 1)	0.595471
3	Recall (Class 0)	0.707440	3	Recall (Class 0)	0.664533
4	Recall (Class 1)	0.712440	4	Recall (Class 1)	0.650000
5	F1-score (Class 0)	0.732474	5	F1-score (Class 0)	0.688481
6	F1-score (Class 1)	0.682521	6	F1-score (Class 1)	0.621542

The model shows moderate performance with an accuracy of ~65.8% on the test set, performing better on the non-hypertension class across all metrics. Precision scores between both groups were stark as the values were 71.4% for the non-hypertension group compared to 59.5% for the hypertension group. This might suggest that the model might be better at identifying true negatives (i.e non-hypertensive cases). Recall scores between both groups are closer together as the values with the non-hypertension group was 66.5% and the hypertension group was 65%. This might suggest that the model can fairly identify both classes. The F1-score was not as stark as the precision scores, as the values for the non-hypertension group scores were 68.8% in comparison to 62.1% for the hypertension group. This may also indicate the model's stronger overall performance in the non-hypertension class.

Part 1H



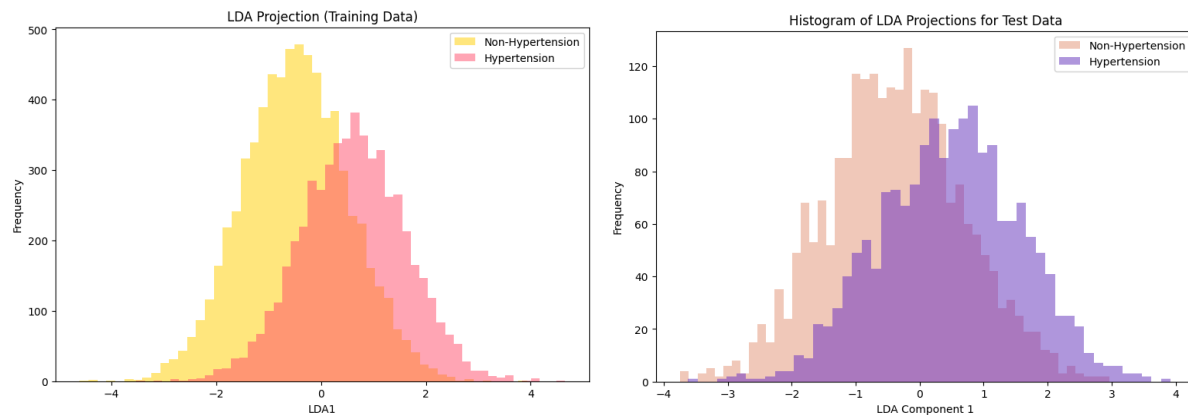
Completed in notebook

Part 1I

From the UMAP plots, do the classes look linearly separable? Do the plots support the logistic regression model's hypertension prediction results? Briefly justify your answer.

The classes do not look linearly separable as I cannot “draw” a line which cleanly separates the two classes. Since the classes overlap significantly, the plot does support the logistic regression model’s suboptimal predictive performance results (i.e precision, recall, and accuracy scores). Simply put, the UMAP plots provide a visual confirmation/reasoning as to why the model’s performance is suboptimal

Part 1J



Completed in notebook

Part 1K

From the histogram plots, do the classes look linearly separable? Do the plots support the logistic regression model’s hypertension prediction results? Briefly justify your answer.

The classes in the histograms show some separation, but overall, they are not linearly separable due to evident overlap, as indicated by the darker orange and deep purple areas in the histogram. This overlap, along with staggered peaks and uneven distribution, may explain the logistic regression model’s suboptimal performance. The significant overlap between the two classes suggests that the model has difficulty distinguishing between them, leading to less accurate predictions.

Part 1L

Do ClinicalBERT embeddings seem like a good choice for automatic label extraction for hypertension? Briefly justify your answer.

ClinicalBERT embeddings do not seem like a good choice for automatic label extraction in this context because there is significant overlap and inadequate separation between hypertensive and non-hypertensive patients in both the UMAP projections and histograms. The embeddings do not sufficiently capture the distinctions needed for accurate label extraction, which suggests they may not be ideal for clinical tasks like hypertension classification.