Homework 2 Report

Author: Yanran Li

Collaborators: Kexuan Liang, Yichen Wang

Part 1.

(c).

```
Sentence 1 (completed): The renal dysfunction score changes over time among the
black African population.
Sentence 2 (completed): The cardiovascular dysfunction score changes over time
among the elderly women population.
```

(d).
One outcome that might result from the use of biased language model representations to automatically extract "ground-truth" labels for a dataset is that it might biased the readers' thought about the settings. Specifically, in a clinical setting, doctors may give wrong judgement after seeing the results from the biased language model. Another example rests that the severity of heart failure may be overestimated for black patient, which will lead unreasonable predicted prognosis and adjustments from doctors' treatments.

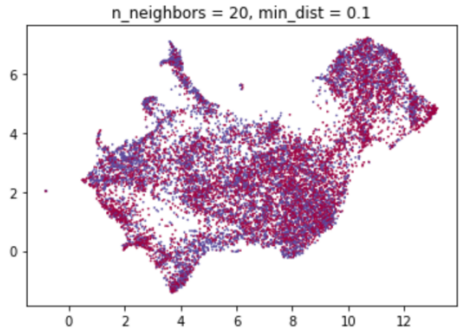
(g).
Metrics for both training and test sets are:
(On the table below, for Precision, Recall and F1-score, the first column represents the patients do not have hypertension and the second column represents the patients who have hypertension.)

index	Accuracy	Precision	Recall	F1-score
Train	0.7068003557855583	[0.75715833 0.65171013]	[0.70398619 0.71040975]	[0.72960477 0.67979513]
Test	0.6524375454765947	[0.7087156 0.5892949]	[0.65941101 0.64325843]	[0.68317488 0.61509535]

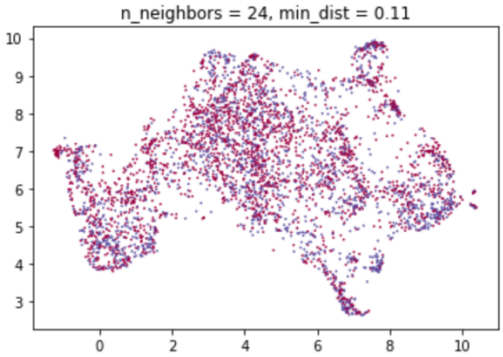
Our logistic regression models on both the training and test sets are not very satisfactory, since the accuracy score is only 0.707 in the training set and 0.652 in the test set.

Our logistic regression model perform better in patients do not have hypertension than those who do not have. Although the Recall scores for both patients who do not have hypertension than those who do not have are similar in train and test sets, the Precision scores and F1 scores are much higher for no hypertension ones than hypertension ones.

(h).
For training set, after tuning the different values of n_neighbors, and min_dist, the best resulting plot from my tuning is:



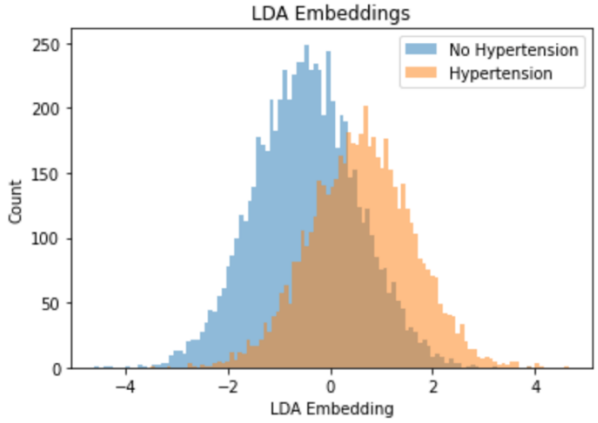
For test set, after tuning the different values of n_neighbors, and min_dist, the best resulting plot from my tuning is:



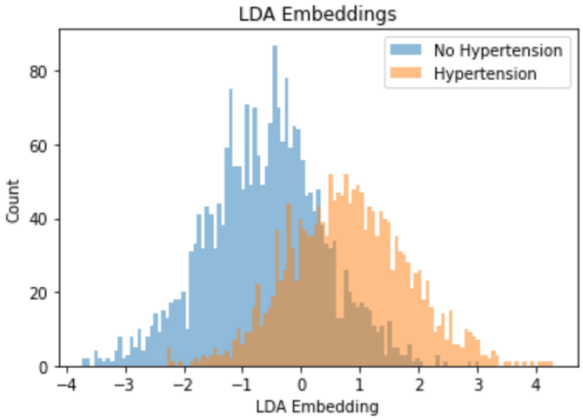
(i).
From the UMAP plots, the classes don't look linearly separable. Different color points seem to mix with each other.

From (g), our logistic regression models have relatively low accuracy score (0.707 in the training sets and 0.652 in the test sets) and can not perform well in predicting hypertension. Thus, the plots can support the logistic regression model's hypertension prediction results.

(j).
Train:



Test:



(k) From the histogram plots, the classes don't look linearly separable. Discrepancies between 2 groups in both train and test sets are relatively inapparent, and there are considerable overlapping embeddings between the two groups.

The plots can support the logistic regression model's hypertension prediction results since the logistic regression models do not have satisfactory accuracy scores (0.707 in the training sets and 0.652 in the test sets) in predicting hypertension..

(l) ClinicalBERT embeddings doesn't seem like a good choice for automatic label extraction for hypertension. From our above discussions, plots that projected high-dimensional data down to low-dimensional data showed the hypertension classes are not linearly separable, which may lead to low accuracy in prediction since little useful information is available for distinguishing the hypertension statuses. In addition, the performances of our logistic models in (g) are not satisfied for label extraction. Besides, the potential societal biases may educed by ClinicalBERT embeddings are also a shortcoming that we should consider when we are seeking a good choice for automatic label extraction for hypertension.