Encoding symptoms from WebMD text using an LLM

By: Ritika and Kaitlin

Initial Processing

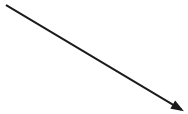
Text from WebMD

ADVERTISEMENT

You might also feel weak, dizzy, or like you're going to pass out, and you could start sweating a lot. Sometimes, you'll also have mild pain in your jaw, neck, back, or arms. Plus, you may have trouble breathing.

What is a silent heart attack?

As the name suggests, a silent heart attack is one that happens without any obvious signs usually related to heart attacks, such as dizziness, a faster or irregular heartbeat (palpitations), trouble breathing, and anxiety. It’s hard for you to know if you’re having a silent heart attack because it happens without warning.

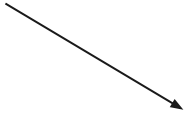


Processed text

What Are the Symptoms of Acne?

The symptoms of acne are:

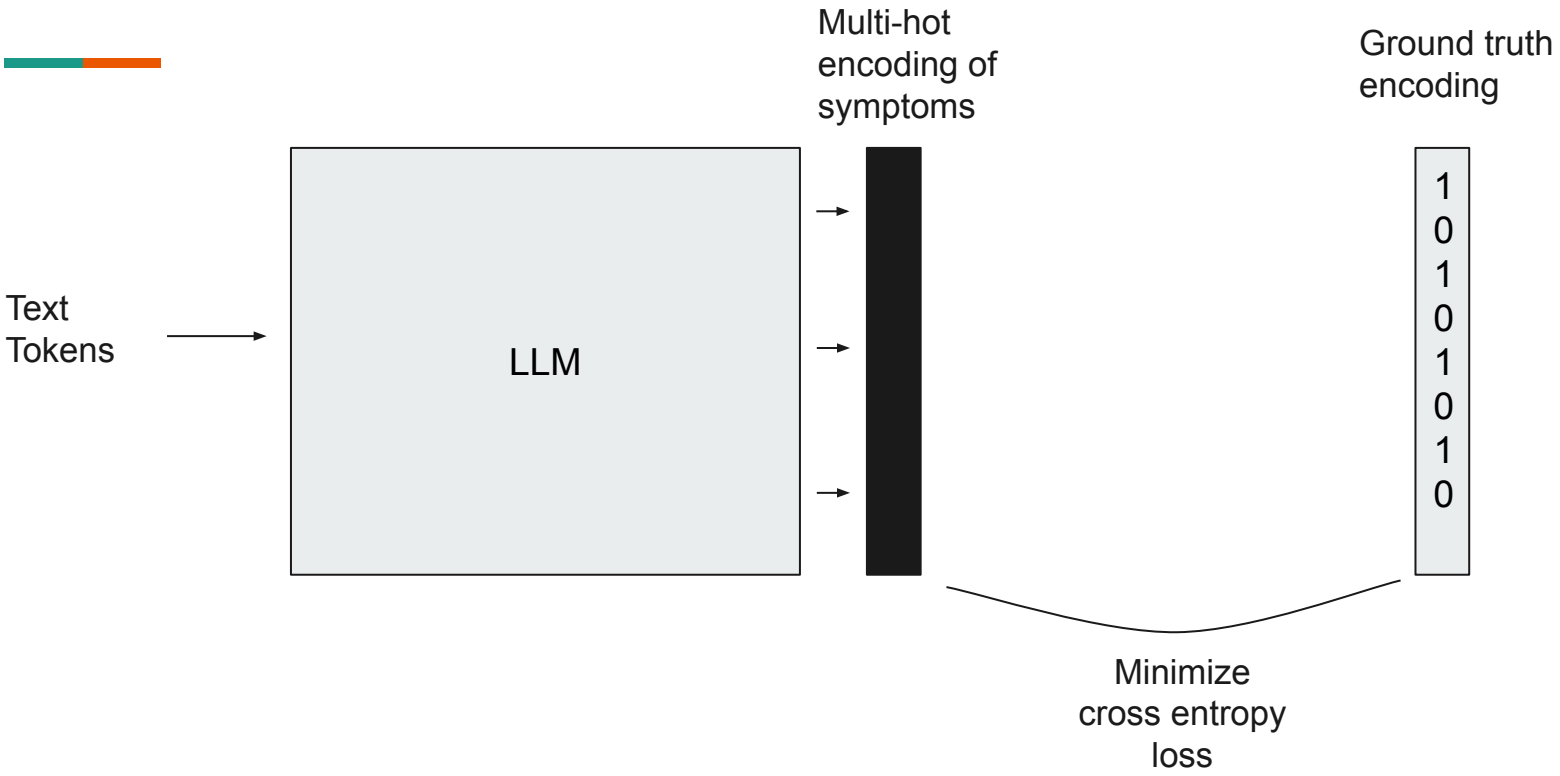
Persistent, recurrent red spots or swelling on the skin, generally known as pimples; the swelling may b
pus. They typically appear on the face, chest, shoulders and/or neck, or upper portion of the back.
Dark spots with open pores at the center (blackheads)
Tiny white bumps under the skin that have no obvious opening (whiteheads)
Red swellings or lumps (known as papules) that are visibly filled with pus
Nodules or lumps under the skin that are inflamed, fluid-filled, and often tender; these nodules may be
across.



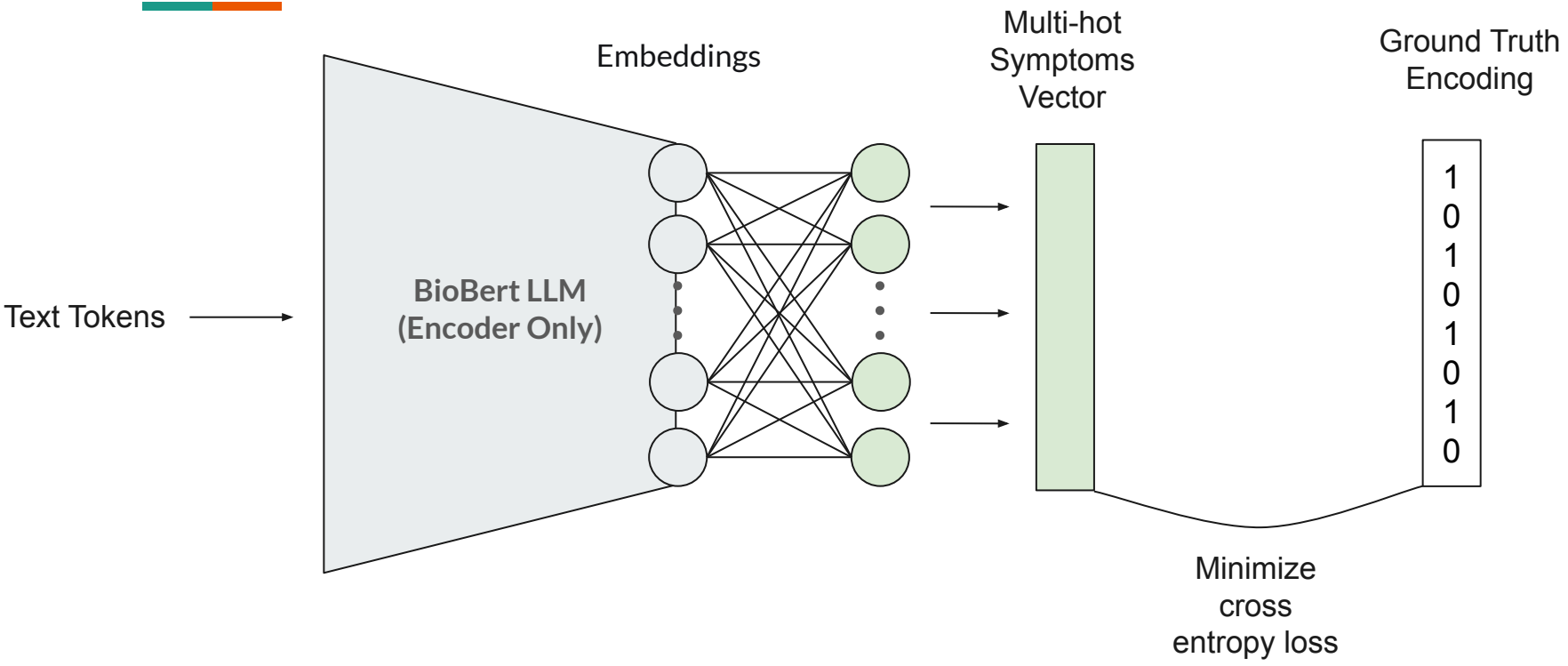
Tokenized

```
Disease 3: allergies
Text: What Are Allergies?
Allergies are when your immune system overreacts to something called an allergen. An allergen is a foreig
Label: [0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
Symptoms present:
 - continuous_sneezing
 - shivering
 - chills
 - watering_from_eyes
```

Model


Multi-hot encoding of symptoms
Ground truth encoding
Text Tokens
LLM
1
0
1
0
1
0
1
0
Minimize cross entropy loss

Model


Text Tokens
BioBert LLM
(Encoder Only)
Embeddings
Multi-hot
Symptoms
Vector
Ground Truth
Encoding
1
0
1
0
1
0
1
0
Minimize
cross
entropy loss

Chosen optimizer, loss function, pretrained LLM model

Optimizer: AdamW

- handles weight decay regularization separately from the optimization steps, improving generalization and training stability in deep learning models

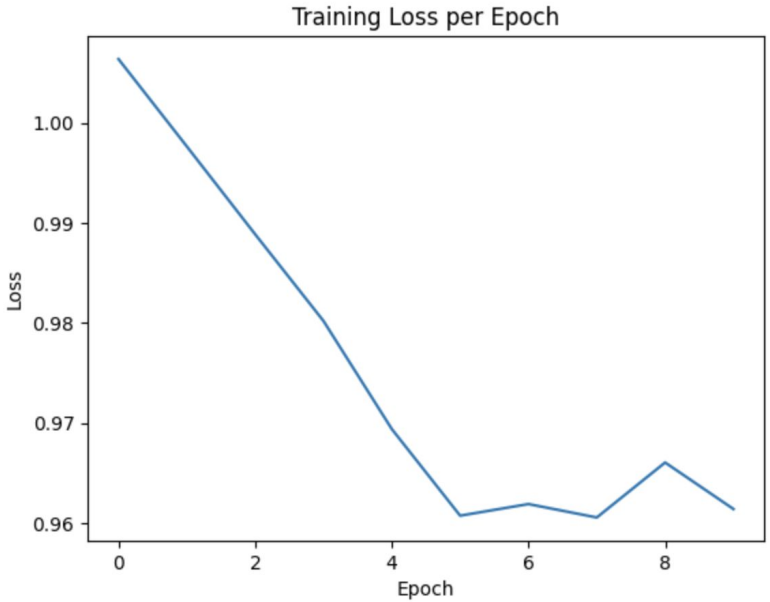
Loss function: BCEWithLogitLoss

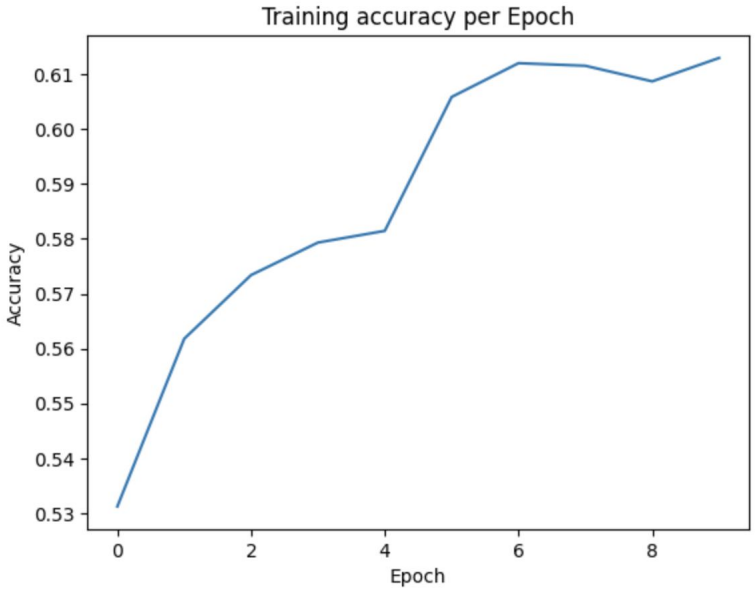
- combines a sigmoid layer with the binary cross-entropy loss in one single class, making it numerically more stable

Pretrained LLM: BioBert

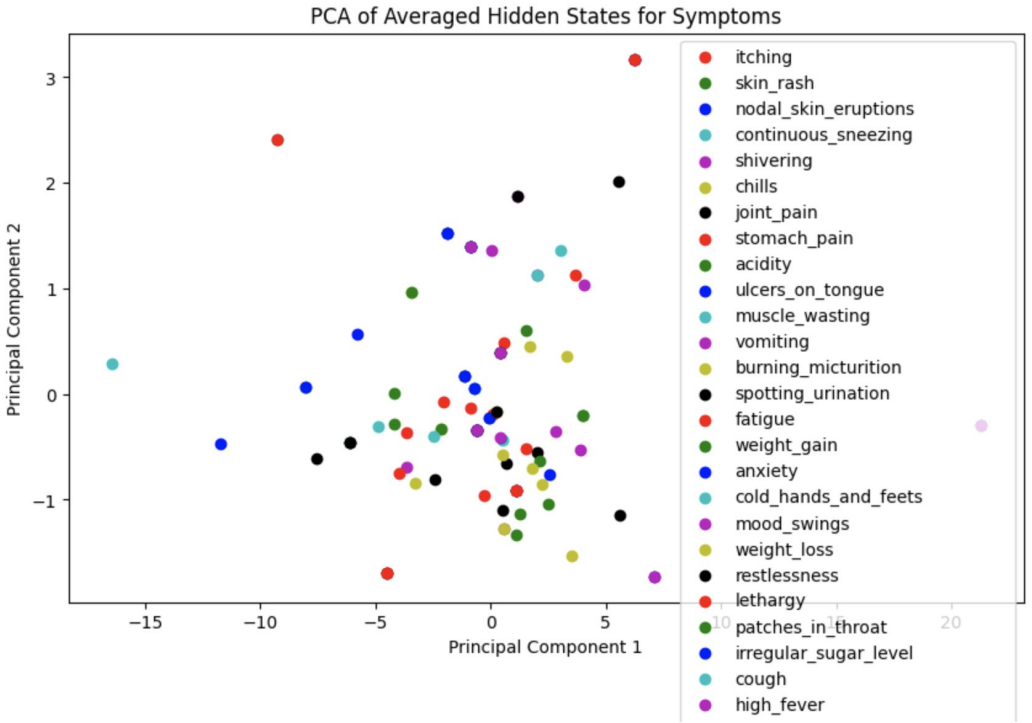
- language representation model pre-trained on large-scale biomedical corpora
- designed for biomedical knowledge for tasks like disease prediction, drug discovery, and medical question answering

Accuracy of symptom prediction (train)





PCA


PCA of Averaged Hidden States for Symptoms
Principal Component 2
Principal Component 1
3
2
1
0
−1
−15
−10
−5
0
5
10
15
20
itching
skin_rash
nodal_skin_eruptions
continuous_sneezing
shivering
chills
joint_pain
stomach_pain
acidity
ulcers_on_tongue
muscle_wasting
vomiting
burning_micturition
spotting_urination
fatigue
weight_gain
anxiety
cold_hands_and_feets
mood_swings
weight_loss
restlessness
lethargy
patches_in_throat
irregular_sugar_level
cough
high_fever

Testing Results

```
predicted disease symptoms:  ['skin_rash', 'continuous_sneezing', 'shivering', 'chills', 'joint_pain', 'stomach_pain', 'acidity', 'vomiting',
Ground truth symptoms:  ['skin_rash', 'pus_filled_pimples', 'blackheads', 'scurring']
disese:  acne
Number of correct symptoms (that the disease has):  2
Number of incorrect symptoms (that the disease has):  2
Model got the following symptoms correct:  ['skin_rash', 'scurring']
Model missed the following symptoms:  ['pus_filled_pimples', 'blackheads']
```

```
Accuracy: 0.0, Precision: 0.04597701149425287, Recall: 0.48, F1-Score: 0.08391608391608392
```

Next Steps:

- more diversity in the training data
- adjust more hyperparameters like the learning rate, #of epochs, class weighting, dropout rate

Potential problem: the WebMD information may not closely align with our ‘ground truth’ symptoms

Comparison of Predicted Symptoms

Method	Disease (train set)	Symptoms	Similarity
Kaggle (actual ground truth)	heart_attack	Vomiting, breathlessness, chest_pain, sweating	0.071
Kaggle (“ground truth”)	heart_attack	'joint_pain', 'vomiting', 'yellowish_skin', 'dark_urine', 'nausea', 'loss_of_appetite', 'abdominal_pain', 'diarrhoea', 'mild_fever', 'yellowing_of_eyes', 'muscle_pain'	0.071
BioBert LLM	heart_attack	'joint_pain', 'vomiting', 'fatigue', 'high_fever', 'sweating', 'dehydration', 'yellowish_skin', 'dark_urine', 'nausea', 'loss_of_appetite', 'abdominal_pain', 'yellowing_of_eyes', 'malaise', 'congestion', 'swollen_legs', 'excessive_hunger', 'irritability', 'history_of_alcohol_consumption'	Actual Ground Truth=0.1 “Ground Truth” = 0.38
GPT4	heart_attack	"Chest_pain", "neck_pain", # as an approximation for jaw pain, "indigestion","sweating","vomiting", "dizziness","Fatigue","breathlessness","Fast_heart_rate", "anxiety"	Actual Ground Truth = 0.4 “Ground Truth” = 0.05 BioBertLLM = 0.12

Principles related

1. Distributed Computation
    a. Computation is split up across different nodes

2. Prediction
    a. Predict health outcome based on text

3. Attention
    a. LLM

Future work

1. Debug Dataset problems
2. Reverse process:
    a. Input symptoms and try to create a text file with the descriptions