Determining multi-hot encoding of symptoms from WebMD text.

Abstract:

This study presents a novel approach to determining medical symptoms from WebMD text using a large language model (LLM), specifically BioBert. We aim to address the challenge of effectively processing natural language text to extract relevant medical symptoms, a task critical in healthcare and medical research. The study begins with the preprocessing of WebMD text; this involves the extraction and tokenization of relevant symptom data. We then utilize a BioBert model, a variant of BERT specialized in biomedical text, that is fine-tuned using the AdamW optimizer and BCEWithLogitLoss for enhanced precision. A logistic regression model aids in predicting symptoms from the encoded text.

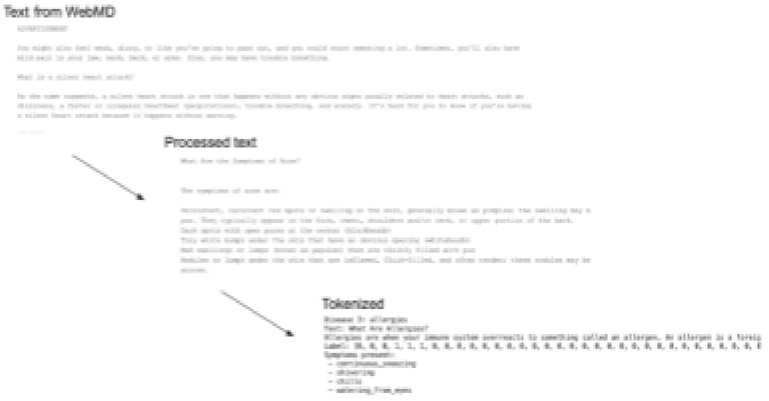
Our primary contribution lies in the application of LLMs to medical text, demonstrating the potential of machine learning to understand and process healthcare-related information. The approach shows promise in disease prediction and symptom analysis, opening new pathways in automated medical diagnostics and patient care. Challenges such as overfitting were addressed by diversifying training data and adjusting hyperparameters, ensuring better model performance. Future work includes further debugging of dataset issues and exploring reverse processing for symptom-based text generation.

Additional Details:

The next section outlines the stages of our model's creation. It starts with the transformation of WebMD medical text, namely, scraped web pages containing information about particular diseases, into data that our model can understand. This is done by breaking down the text into smaller pieces, a process known as tokenization. The figures will show each step in this process, demonstrating how our model learns to identify and categorize medical symptoms from the text data**.**

The initial step consists of text preprocessing. During this, raw medical text from WebMD is converted into a processed, tokenized format. This illustrates the transformation of a symptom description into a structured, tokenized list, which is the preprocessing necessary for the LLM to interpret and analyze the data **(Fig. 1)**.

FIGURE 1



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 flutter 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.

[illegible]
```

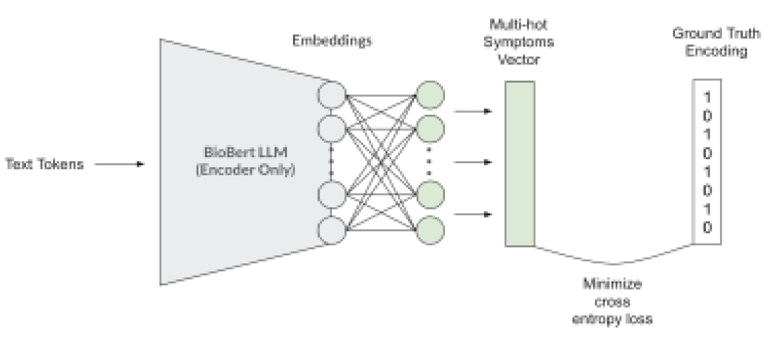
Processed text

```
What Are the Symptoms of Acne?


The symptoms of acne are:

[illegible]
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
severe.
```

Tokenized

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

FIGURE 2


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

Next, the tokenized text is fed into the BioBert LLM, which is pretrained on medical text. This LLM outputs a matrix of logits that is then removed, and the model produces a new output layer of a multi-hot encoding of symptoms **(Fig. 2)**. We then use cross entropy loss to minimize the error between the ground truth multi-hot symptom encoding and the encoding produced by the LLM. Specifically, we fine-tuned the LLM with the AdamW optimizer and used BCEWithLogitLoss to measure loss. After fine-tuning, the model returns a multi-hot encoding of symptoms from each text input.

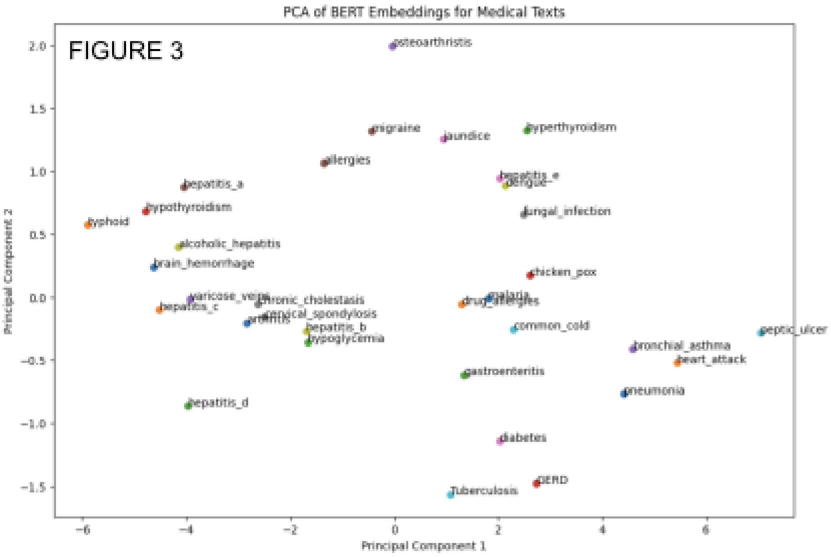


FIGURE 3

Fig. 3 shows a 2D scatter plot resulting from a principal component analysis (PCA) on the hidden states of the LLM after processing the symptom text. The vast distribution of data points indicates the variation in the model's internal representations of the diseases. There are no clear clusters in this plot, indicating that our model doesn't typically see strong correlations between diseases. However, there are instances where certain diseases cluster close together, such as Hepatitis E and Dengue. Although their symptom set isn't identical, they are both viral infections that share multiple symptoms. That said, not all closely positioned points have a similar symptom set, indicating that our model might not have the most accurate representation of all diseases.

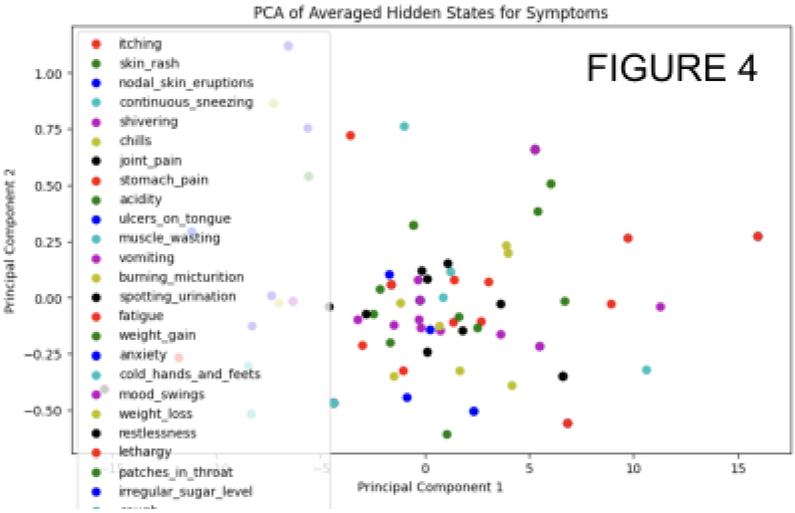


FIGURE 4

Fig. 4 builds upon the previous figure by averaging the hidden states for each symptom and plotting them in the same principal component space. Each dot, colored uniquely per symptom, represents the aggregated representation of a symptom across multiple instances. This visualization is particularly insightful for the project as it demonstrates the model's ability to consistently encode the same symptoms in a similar manner, a fundamental requirement for accurate symptom recognition in medical diagnosis. In addition, certain points, such as 'painful walking,' and 'movement stiffness,' exist very close to each other, showing that our model considers these symptoms to have similar representations. This implies that our model has some capability to understand textual descriptions and abstract them into a high-dimensional space where similar symptoms are near each other, facilitating differentiation and classification tasks.

Overall, although the model is able to exhibit some understanding of disease and symptom relationships and correctly extract certain disease symptoms from text, the model is unable to consistently and precisely predict the multi-hot symptom set for an unseen disease, as indicated by its performance on the test set: Accuracy: 0.0, Precision: 0.03, Recall: 0.08, F1-Score: 0.04. The project presents an exploration into medical diagnostics, where LLMs interpret and encode symptoms from text with some accuracy. The consistent clustering of symptom representations via PCA represents the model's understanding of medical data. Future steps will focus on refining dataset integrity and reverse engineering the encoding process for generating text-based narratives from multi-hot encoding of symptoms. This will hopefully further bridge the gap between artificial intelligence and practical healthcare applications.