CMPE 259

Smart Symptom and Triage Assistant
Final Report

Vaibhavi Savani - 017456972

Date: 12/08/2024

Problem Definition

In today's healthcare landscape, individuals often face challenges in identifying the causes of their symptoms and accessing reliable, evidence-based information. The lack of accessible tools for symptom analysis can lead to increased anxiety, misdiagnosis, or delayed consultation.

The Smart Symptom and Triage Assistant seeks to empower users by providing a platform for symptom analysis. By leveraging a comprehensive dataset of symptoms and diseases and advanced natural language processing, the system aims to:

- Assist users in understanding potential causes of their symptoms.
- Provide personalized insights based on medical knowledge.
- Deliver evidence-based information about symptoms and conditions.

This assistant aspires to help individuals better understand their health and make informed decisions.

Dataset

The dataset was curated through web scraping, involving:

- **Step 1: Gathering Diseases** from the NHS website: https://www.nhs.uk/conditions/
- **Step 2: Collecting Symptoms** from the Mayo Clinic website: https://www.mayoclinic.org/search

Dataset Summary

- **Number of Diseases**: 788
- **Number of Symptoms (Words)**: 46,312
- **Data Composition**: Disease names, associated symptoms, and metadata.

Methodologies

1. **MiniLM-Based Approach**
    - **Embedding and Retrieval**: Uses sentence-transformers/paraphrase-MiniLM-L6-v2 for semantic embeddings of symptoms and diseases. These embeddings are stored in FAISS for efficient retrieval.
    - **Strengths**: Efficient and scalable for large datasets.
    - **Limitations**: Lacks precision for nuanced or rare medical cases.
2. **BioBERT-Based Approach**
    - **Domain-Specific Embedding**: Tailored for biomedical text, offering improved retrieval accuracy for medical queries.
    - **Strengths**: Superior performance for medical applications.
    - **Limitations**: High computational demands hinder real-time performance.
3. **GPT-Neo-Based Approach**
    - **Language Model Capabilities**: Fine-tuned on question-answer pairs for symptom-disease prediction.
    - **Challenges**: Reliance on fine-tuning and high-quality datasets for effective predictions.

4. **LLaMA-Based Approach**
    - **Model Overview**: LLaMA-2 (meta-llama/Llama-2-7b-chat-hf) excels in handling complex prompts and long input sequences.
    - **Strengths**: High accuracy and robust contextual understanding.
    - **Challenges**: Computationally intensive.

Optimization Techniques

- **Prompting**: Dynamic prompts incorporate retrieved context and explicitly guide models to generate disease names or fallback responses like "I don't know the answer."
- **Hyperparameter Tuning**: Parameters like top_k, temperature, and token length were optimized through grid search, improving accuracy to over 78%.
- **Retrieval-Augmented Generation (RAG)**: Dynamically retrieves context using FAISS, enhancing inputs for the models.

Evaluation

To evaluate the models' performance, a dataset of questions was generated using GPT-Neo based on disease-symptom pairs. The prompt used for generating the questions aimed to mimic realistic inquiries a patient might ask a doctor or caregiver to better understand their condition, treatment options, and management strategies.

The process involved:

- Randomly selecting symptoms associated with diseases from the dataset.
- A total of 46 questions were generated and used as the evaluation set across all models.

Model Evaluation Results

The models were tested on their ability to predict diseases based on the evaluation questions. The accuracy results are as follows:

- **BioBERT**: Accuracy: **65.22%**
  BioBERT's domain-specific embeddings provided decent accuracy but struggled with generalization for nuanced cases or symptoms with overlapping diseases.
- **MiniLM**: Accuracy: **78.26%**
  MiniLM's efficiency and strong general-purpose embeddings enabled it to outperform BioBERT, though it lacked domain-specific insights for complex cases.
- **GPT-Neo**: Accuracy: **45.65%**
  GPT-Neo struggled to provide consistent and accurate predictions without robust context retrieval, highlighting its reliance on fine-tuning and prompt engineering.
- **LLaMA-2**: Accuracy: **89.56%**
  LLaMA-2 achieved the highest accuracy, showcasing its ability to handle large input sequences and complex symptom-disease relationships effectively. However, its performance is also tied to higher computational demands.

Conclusion

The evaluation results indicate that the **LLaMA-2** model outperformed all other approaches, achieving an accuracy of **89.56%**, making it the most reliable model for predicting diseases based on symptoms and context. **MiniLM** followed with a respectable **78.26%**, demonstrating its scalability and efficiency despite being a general-purpose model.

While **BioBERT** provided decent performance (**65.22%**), its domain specialization did not fully compensate for the need for robust context retrieval. **GPT-Neo**, with the lowest accuracy (**45.65%**), struggled to generalize effectively, emphasizing the importance of enriched datasets and enhanced prompts for its application.

The findings highlight that:

- **LLaMA-2** is best suited for tasks requiring high accuracy and detailed contextual understanding, albeit at a higher computational cost.
- **MiniLM** offers a good balance between performance and efficiency, making it suitable for real-time applications.
- Further optimization, such as dataset enrichment and improved RAG integration, could enhance the performance of models like BioBERT and GPT-Neo, especially in handling edge cases and ambiguous symptom descriptions.

These results provide valuable insights for deploying the **Smart Symptom and Triage Assistant** in practical healthcare scenarios.

Future work

1. **Dataset Enrichment**:
   Expand the dataset to include a wider range of diseases, symptoms, and real-world patient questions. Incorporate rare diseases and less common symptoms to improve the system's generalization capabilities and handle edge cases more effectively.

2. **Model Specialization**:
   Experiment with domain-specific models like BioBERT or fine-tuned LLaMA versions trained exclusively on medical datasets. This could enhance the accuracy and reliability of predictions for nuanced or ambiguous medical scenarios.

3. **Real-Time Deployment**:
   Develop a scalable and responsive deployment pipeline. Integrate the system into real-time applications like chatbots or virtual assistants, ensuring fast inference and seamless interaction with users in various healthcare settings.

References

1. NHS. (n.d.). *Health conditions*. Retrieved December 8, 2024, from https://www.nhs.uk/conditions/
2. Mayo Clinic. (n.d.). *Diseases and conditions*. Retrieved December 8, 2024, from https://www.mayoclinic.org/diseases-conditions
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. In *NeurIPS 2020*. arXiv. https://arxiv.org/abs/2005.11401
4. Meta AI. (n.d.). *LLaMA 2: Open foundation and fine-tuned chat models*. Retrieved December 8, 2024, from https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/
5. Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (n.d.). *GPT-Neo: Large scale autoregressive language modeling with Mesh-TensorFlow*. Zenodo. https://doi.org/10.5281/zenodo.5297715
6. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained Transformers*. arXiv. https://doi.org/10.48550/arXiv.2002.10957
7. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). *BioBERT: A pre-trained biomedical language representation model for biomedical text mining*. arXiv. https://doi.org/10.48550/arXiv.1901.08746