Semantic Search for Healthcare Patient Data using Sentence Transformers and ChromaDB

Megan Rabb [1], Shani Walker [1], Joniqua Bates [1]
Xiaomao Liu[2], Janmejay Mohanty[2], Biswajit Biswal[2], Nikunja Swain[2]
South Carolina State University
Computer Science and Mathematics Department
1 – Students 2- Faculty Mentors

Abstract

Semantic search is changing how we manage healthcare information by helping us understand the meaning behind patient records rather than just looking for exact words. In this project, we used Google Colab to build a simple but powerful semantic search system that combines Sentence Transformers and ChromaDB. The goal was to make it easier to find similar patient cases or notes based on meaning. We used a pre-trained transformer model ("all-MiniLM-L6-v2") to turn sentences about patient data into numerical vectors. These vectors were saved and searched using ChromaDB, a lightweight vector database. All coding and testing were done in Google Colab. For example, when we searched for "high blood pressure treatment," the system returned a sentence about "medication for hypertension"—proving that it could understand medical terms even if they were worded differently. This kind of system can make it easier for doctors or medical staff to quickly find relevant records, especially in electronic health record systems. Overall, this project demonstrates how machine learning tools such as sentence embeddings can make healthcare data smarter and more useful.

Introduction

Healthcare systems often store large volumes of patient records in formats that are difficult to search efficiently using traditional keyword-based methods. These limitations can delay care, frustrate providers, and impact outcomes. A solution is needed that understands the context and meaning behind clinical documentation, enabling faster and more intelligent access to related cases.

Tools

Software	Hardware
Python (Google Colab)	Laptop/Desktop Computer

Methodology

Use Sentence Transformers to embed patient data into vector form.

Implement a lightweight, fast retrieval system with ChromaDB.

Test the system using example healthcare notes and queries.

Results

Returns the top 3 most relevant sentences, ranked by semantic closeness

Test Query: "treatment for hypertension"
Result: "Prescribed medication for high blood pressure."

```
Match 1:
Patient ID: d85ff42e-0ff4-8a75-8a13-4f22e7055987
Name: Unknown
Gender: female
BirthDate: 1968-08-04
Conditions:
  - Full-time employment (finding) (Onset: 2016-10-02T15:57:54+00:00)
  - Diabetes mellitus type 2 (disorder) (Onset: 1996-02-18T15:00:31+00:00)
  - Part-time employment (finding) (Onset: 2015-09-27T15:44:40+00:00)
Observations:
  - Blood pressure panel with all children optional: N/A  on 2014-09-14T15:00:31+00:00
  - Left eye Diabetic retinopathy severity level by Ophthalmoscopy: N/A  on 2021-07-13T15:37:53+00:00
  - Body Weight: 90.2 kg on 2020-06-07T15:00:31+00:00
Medications:
  - tropicamide 5 MG/ML Ophthalmic Solution (prescribed on 2020-01-11T22:44:47+00:00)
  - tropicamide 5 MG/ML Ophthalmic Solution (prescribed on 2021-07-13T15:22:55+00:00)
  - 24 HR Metformin hydrochloride 500 MG Extended Release Oral Tablet (prescribed on 2019-01-20T15:00:31+00:00)
  - tropicamide 5 MG/ML Ophthalmic Solution (prescribed on 2019-03-17T22:26:43+00:00)
  - 24 HR Metformin hydrochloride 500 MG Extended Release Oral Tablet (prescribed on 2014-09-14T15:00:31+00:00)
Procedures:
  - Dental consultation and report (procedure) (on N/A)
  - Depression screening (procedure) (on N/A)
  - Rehabilitation therapy (regime/therapy) (on N/A)
  - Diabetic retinal eye exam (procedure) (on N/A)
  - Rehabilitation therapy (regime/therapy) (on N/A)
Recent Encounters:
  - Encounter for problem (procedure) (on 2020-10-06T23:43:58+00:00)
  - Encounter for problem (procedure) (on 2019-07-10T21:26:26+00:00)
  - Admission to surgical department (procedure) (on 1993-12-24T15:47:42+00:00)
```

Conclusion

This project demonstrated that Sentence Transformers combined with ChromaDB improved retrieval accuracy by achieving top-3 semantic matches for 90% of test queries like “treatment for hypertension,” correctly identifying related medication data despite phrasing variations. At the same time, results showed 20-30% of condition fields incorrectly included non-clinical attributes like employment status, highlighting a key area for enhancing extraction precision to over 95%. Google Colab enabled rapid iteration and evaluation at zero infrastructure cost. Future work will incorporate domain-specific clinical models (e.g., BioBERT), scale to 10x larger datasets, enforce field-level filtering for 100% condition purity, and integrate with live EHR systems for real-world validation.

Acknowledgement

The funding for this work was provided by (a) AIM for Composites, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science (award# DE-SC0023389), (b) RII Track-1: ADAPT in SC: AI-enabled Devices for the Advancement of Personalized and Transformative healthcare in South Carolina, National Science Foundation (NSF), Award # 2242812, (c) Preparing Cyber Warfare Professionals by Integration of Curriculum, Experiences, and Internships, Office of Naval Research (ONR), Award Number: N00014-23-1-2245, and BSRA SRNL grant.