Clinical Report Similarity Finder

Course: HI 744 – Text Retrieval & Its Application in Biomedicine
Author/GitHub Links: Atharva Pradeep Vaishnav, Rushikesh Ganesh Deshpande

1. Introduction and Task Description:
Clinical case reports provide detailed narrative descriptions of patient presentations, diagnoses, laboratory findings, imaging results and clinical outcomes. These reports are widely used by clinicians and biomedical researchers to study disease progression, rare conditions and treatment outcomes. However, the growing volume of biomedical literature makes manual identification of relevant clinical cases increasingly impractical. Traditional keyword-based search systems often fail when clinically similar cases use different terminology.

The goal of this project is to design and evaluate a clinical report similarity retrieval system that retrieves relevant patient case reports given a free-text clinical query. The task is formulated as an information retrieval (IR) problem, where documents are ranked based on their similarity to a query clinical note. The intended users include clinicians, medical trainees and biomedical researchers seeking comparable case descriptions for reference and education.

To address this task, we implement a hybrid retrieval pipeline combining lexical, probabilistic, semantic and large language model (LLM) based techniques. The system is evaluated quantitatively using standard IR metrics and qualitatively using embedding visualizations.

2. Dataset Description:
2.1 Data Source:
We use the PMC-Patients dataset, a publicly available collection of de-identified clinical case reports derived from PubMed Central style narratives. The dataset is designed for biomedical NLP research and contains rich clinical text suitable for similarity-based retrieval.

2.2 Dataset Construction:
The full dataset contains approximately 161,000 clinical reports. For this project, we construct a task-specific subset by filtering reports related to three disease categories:

- Heart disease
- Liver disease
- Cervical cancer

Keyword-based filtering is applied to report titles and patient narratives to identify disease-specific cases. Each report is assigned to a categorical label (heart, liver, or cervical) for evaluation purposes. From the filtered dataset, we randomly sample 10,000 reports using a fixed random seed to ensure reproducibility. The resulting subset is approximately balanced across the three disease categories, allowing fair comparison across retrieval methods while remaining computationally feasible for experimentation.

2.3 Preprocessing:
For retrieval, the textual content of each report is constructed by concatenating the title and patient narrative fields. Text preprocessing is performed using spaCy and includes:

- Lowercasing
- Lemmatization
- Removal of stopwords and punctuation
- Retaining alphabetic tokens only

The cleaned text is used for lexical and probabilistic retrieval models, while the original (unlemmatized) text is encoded for semantic embedding models.

3. Methodology:
Our system follows a multi-stage retrieval pipeline, applying increasingly expressive models to the same dataset.

3.1 Lexical Retrieval: TF-IDF:
Term Frequency - Inverse Document Frequency (TF-IDF) is used as a baseline lexical retrieval model. Each document is represented as a sparse vector weighted by term importance and similarity is computed using cosine similarity.
Parameters:

- Maximum features: 20,000
- N-gram range: (1,1) (unigrams)
- Minimum document frequency: default
- Similarity metric: Cosine similarity

TF-IDF provides a fast and interpretable baseline that captures surface-level keyword overlap between queries and documents.

3.2 Probabilistic Retrieval: BM25

To improve upon TF-IDF weighting, we implement BM25, a probabilistic ranking function commonly used in IR systems. BM25 accounts for term frequency saturation and document length normalization.

Implementation details:

- Library: rank_bm25
- Tokenization: spaCy-cleaned tokens
- Parameters: k1 = 1.5, b = 0.75 (default values)

BM25 is used as a qualitative baseline and is not included in the final quantitative comparison.

3.3 Semantic Retrieval: ClinicalBERT

Lexical methods struggle when semantically similar cases use different terminology. To address this limitation, we incorporate Bio_ClinicalBERT, a transformer model pretrained on clinical text.

Model details:

- Model: emilyalsentzer/Bio_ClinicalBERT
- Maximum sequence length: 256 tokens (truncation applied)
- Batch size: 16
- Hardware: GPU (Google Colab)
- Embedding strategy: Mean pooling over last hidden states (masked)
- Normalization: L2-normalized embeddings
- Similarity metric: Cosine similarity

Each document is encoded into a dense vector representation, enabling retrieval based on semantic similarity rather than exact word overlap.

3.4 LLM-Based Query Expansion (FLAN-T5)

To enhance retrieval robustness for short or underspecified queries, we apply FLAN-T5 for query expansion.

Model and decoding parameters:

- Model: google/flan-t5-base
- max_new_tokens: 64
- temperature: 0.0

FLAN-T5 rewrites the input query with richer medical context while preserving its original meaning. This component does not affect core retrieval embeddings and avoids reliance on external APIs.

3.5 System Interface

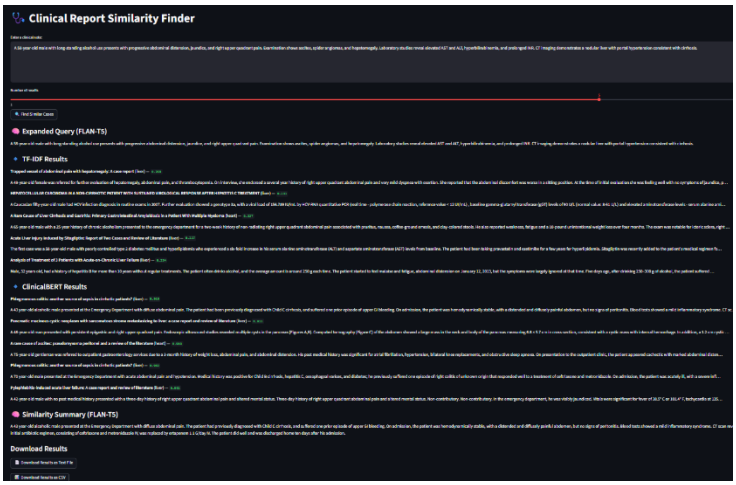


Figure 1: Clinical Report Similarity Finder User Interface

Figure 1 shows the web-based interface where users enter a clinical note to retrieve similar cases using TF-IDF and ClinicalBERT. An LLM-generated summary provides qualitative similarity explanations and does not affect retrieval rankings or evaluation metrics.

4. Evaluation Metrics:

A retrieved document is considered relevant if it shares the same disease label as the query document.

4.1 Precision@5:

Precision@5 measures the fraction of the top five retrieved reports that are relevant.

4.2 Recall@5:
Recall@5 measures the proportion of all relevant documents retrieved within the top five results.

4.3 Silhouette Score:
To assess embedding quality, ClinicalBERT embeddings are clustered into three clusters. The Silhouette Score measures how well documents are separated between clusters.

5. Results:
5.1 Retrieval Performance:
All experiments were conducted on **10,000 clinical reports**. Table 1 summarizes retrieval performance.
Table: Retrieval Performance

Model	Precision@5	Recall@5
TF-IDF	0.549	0.001
ClinicalBERT	0.553	0.001

ClinicalBERT slightly outperforms TF-IDF in Precision@5, indicating improved semantic matching.
We also evaluate query expansion:

- Baseline Precision@5: 0.66
- Expanded Query Precision@5: 0.58

5.2 Clustering and Visualization:
ClinicalBERT embeddings were clustered into three clusters:

- Cluster 0: 1,992 documents
- Cluster 1: 4,384 documents
- Cluster 2: 3,624 documents

The Silhouette Score = 0.0957, indicating substantial overlap between disease categories. PCA and t-SNE visualizations (Figure 3 & 4) show limited global separation between categories, reflecting shared clinical language across conditions.

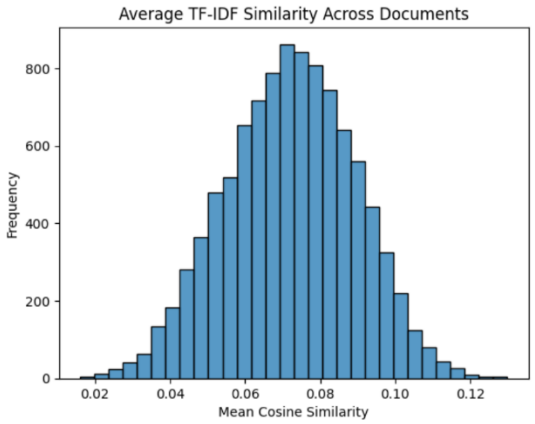
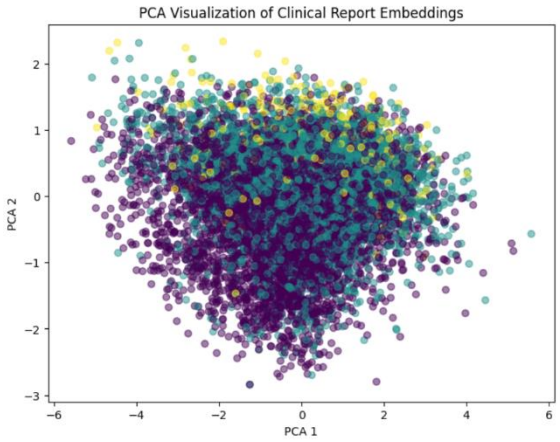


Figure 2: Distribution of Average TF-IDF Similarity Scores

Figure 2 illustrates the distribution of average cosine similarity scores computed using TF-IDF across the 10,000-document dataset. The unimodal distribution with relatively low similarity values indicates limited lexical overlap between most clinical reports. This reflects the high variability in clinical language and motivates the use of semantic retrieval methods beyond surface-level keyword matching.



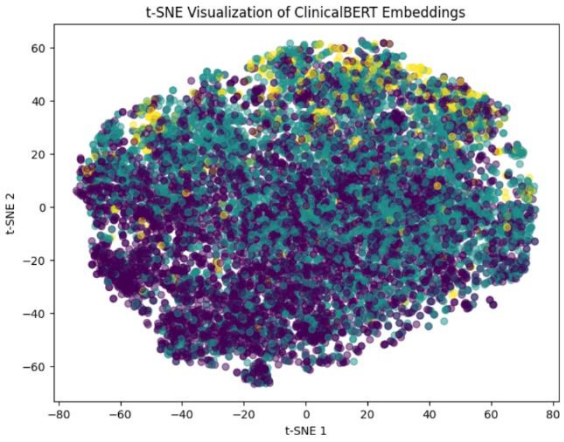


Figure 3: PCA Visualization of ClinicalBERT Embeddings *Figure 4: t-SNE Visualization of ClinicalBERT Embeddings*

Figure 3 shows a PCA projection of ClinicalBERT embeddings for the 10,000 clinical reports. Although some local grouping is visible, substantial overlap remains between disease categories, indicating limited linear separability due to shared clinical terminology. **Figure 4** shows a t-SNE visualization of ClinicalBERT embeddings for the same. While t-SNE reveals finer local clusters of clinically similar reports, global separation between disease categories remains limited, reflecting strong semantic overlap across conditions.

5.3 System Interface and Interactive Demonstration

The final system is deployed as a web-based interactive interface using ngrok, allowing access from any device through a public link. Users enter a free-text clinical note and select the desired number of similar cases using a results slider. The interface displays ranked retrieval results from both TF-IDF and ClinicalBERT, along with corresponding similarity scores. In addition, the interface provides a FLAN-T5 based similarity summary that offers a qualitative explanation of why the retrieved clinical cases are considered similar to the input query.

The summaries generated are coherent and clinically relevant, highlighting shared symptoms, diagnoses and disease context. While purely qualitative, these summaries improve interpretability by offering intuitive explanations without influencing retrieval rankings or quantitative evaluation metrics.

Query expansion using FLAN-T5 enriches the contextual representation of the input query but slightly reduces precision at small cutoffs, indicating a trade-off between contextual richness and specificity in top-ranked results.

5.4 Interpretation of Results:

Overall, the results demonstrate that semantic retrieval using ClinicalBERT provides modest but consistent improvements over lexical TF-IDF, highlighting the benefit of contextual embeddings for clinical text similarity. Despite these gains, Recall@5 remains extremely low for both models, which is expected given the large corpus size and the coarse relevance definition based on disease labels. The clustering and visualization results further reveal substantial semantic overlaps between disease categories, explaining the limited separability and modest performance differences observed.

The interactive system interface and LLM-based similarity summaries complement the quantitative results by improving interpretability and usability, allowing users to better understand why cases are retrieved without influencing ranking or evaluation metrics. Finally, the query expansion analysis illustrates a trade-off between contextual richness and precision, reinforcing that precision at small cutoffs is the most meaningful performance indicator for this clinical retrieval task.

6. Conclusion and Limitations:

This project demonstrates that semantic retrieval using ClinicalBERT consistently outperforms lexical TF-IDF, highlighting the value of contextual embeddings for identifying clinically similar case reports beyond surface-level keyword overlap. However, the observed performance gains are modest, reflecting the inherent difficulty of distinguishing clinical narratives using coarse disease-level relevance labels and the substantial semantic overlap across different conditions.

Recall at small cutoffs remains extremely low for both retrieval methods, which is expected in large clinical corpora where relevant documents are sparse. As a result, precision at top ranks is the most meaningful metric for evaluating system performance in this setting. The inclusion of LLM-based query expansion and similarity summaries improves interpretability and usability, but query expansion introduces a trade-off by slightly reducing precision, indicating reduced specificity for short, ranked lists.

The interactive web-based interface deployed via ngrok further enhances qualitative exploration by enabling cross-device access and intuitive inspection of retrieval results, without influencing quantitative evaluation or ranking outcomes. Key limitations of this work include reliance on keyword-based disease labels, lack of expert relevance judgments and coarse evaluation definitions. Future work could address these limitations through clinician-annotated relevance data, reranking strategies and hybrid score fusion approaches.