Retrieving Semantically Similar Clinical Trials

Team : Malaai

Pushpender Singh

Priyadarshi Anand

Ujjawal Saini

Rahul Yadav

Tushar Jindal

Problem Understanding

Objective

To efficiently retrieve semantically similar clinical trials using advanced natural language processing (NLP) and scalable search techniques.

Challenges

- **Data Heterogeneity:** Clinical trials contain diverse information across multiple fields (e.g.,study title,).
- **Semantic Understanding:** Need to capture meaning beyond keyword matching.
- **Scalability:** Efficient retrieval from a large dataset of trials.

Methodology

Data Preprocessing

- Source: Extracted data from eligibilities.txt to create a new column named "criteria."
- Dataset: Combined "criteria" with other trial fields (e.g., title, conditions, outcomes) to create a unified dataset for analysis.

Model Selection

Transformer Model: sentence-transformers/all-mpnet-base-v2

- Pre-trained for semantic similarity tasks.
- Encodes textual data into high-dimensional dense embeddings.

Retrieval Mechanism

FAISS (Facebook AI Similarity Search):

- Optimized for scalable and fast nearest-neighbor searches.
- Supports high-dimensional vector indexing and retrieval.

Approach

1

Preprocessing

Unified fields into a single textual representation for each trial. Normalized and cleaned text to improve model performance.

2

Embedding Generation

Used sentence-transformers/all-mpnet-base-v2 to encode text into semantic embeddings. Captured the meaning of each trial in a dense numerical vector.

3

FAISS Indexing

Indexed embeddings using FAISS for cosine similarity-based retrieval. Normalized embeddings to ensure consistent similarity calculations.

4

Query Handling

Input query text is encoded using the same transformer model. Normalized query embedding is searched in the FAISS index to retrieve the top k similar trials.

Results

- **Similarity Score:** Achieved a similarity score between 0.8 and 0.9, indicating strong semantic relevance.
- **Performance:** Achieved rapid retrieval times, even for large datasets.

Conclusion

1 Key Takeaways

Combined state-of-the-art NLP with scalable search for efficient trial retrieval. Achieved meaningful semantic understanding of diverse clinical trial data. Demonstrated the effectiveness of FAISS for handling large datasets.

2 Future Work

Integrate user feedback to refine query relevance. Expand preprocessing to include additional metadata fields. Explore real-time updates to the FAISS index for dynamic datasets.