Problem Statement

Semantic* grouping of clinical studies for retrieval & strategic insights

Abstract

Overview

Problem Overview

- **The task involved developing an NLP-based semantic search system for clinical trials, aimed at retrieving the most relevant trials based on user queries. It required processing textual data like study titles, outcomes, and eligibility criteria to compute semantic similarity and provide explainable results.**

Solution Overview

1) Generate separate semantic embeddings for each clinical trial text field (e.g., study titles, outcomes, eligibility criteria) using the BioMedBERT model, leveraging its domain-specific capabilities for biomedical text.

2)Use a mean pooling strategy to aggregate token-level embeddings into sentence-level representations.

3)Compute cosine similarity to rank the top 10 relevant trials.

4)Use SHAP and LIME for global and localized explainability of similarity scores.

5)Validate and visualize results to ensure relevance and improve performance.

Approach

Embedding Generation:

- Generate separate embeddings for each clinical trial text field using **BioMedBERT**, chosen for its specialized understanding of biomedical text.
- Apply a **mean pooling strategy** to convert token-level embeddings into sentence-level representations.

Similarity Computation:

- Compute cosine similarity between query embeddings and clinical trial field embeddings.
- Aggregate similarity scores from different fields to create a comprehensive ranking of trials.

Explainability Analysis:

- Apply **LIME** for localized explanations to interpret the relevance of individual queries and highlight contributing text segments.

Validation and Evaluation:

- Perform **Top-K Accuracy** and **Mean Reciprocal Rank (MRR)** evaluations using query titles from the original dataset.
- Validate performance using a random sample of dataset entries to ensure robustness and generalizability.

Results and Limitations

Performance Metrics:

- **Top-10 Accuracy**: Achieved a score of **99.6%** on a subset of 100 random entries, indicating the model's ability to consistently retrieve the correct clinical trial within the top 10 ranked results.
- **Mean Reciprocal Rank (MRR)**: Achieved an MRR of **0.962**, showcasing the high relevance of top-ranked trials and the model's ability to prioritize correct trials effectively.

Limitations of Mean Pooling Strategy

Reason for Selection: Mean pooling was chosen as the embedding aggregation strategy due to its simplicity and effectiveness in capturing the overall semantic meaning of token-level embeddings, particularly for domain-specific models like BioMedBERT.

Advantages: It ensures a balanced representation by averaging the embeddings, making it robust to variations in text length and noise in individual token embeddings.

Limitations: Mean pooling may lose finer-grained contextual information compared to more advanced techniques like attention-based pooling.

Introduction and Model Choice

Introduction

Background

Clinical trials are pivotal for evaluating the safety and efficacy of medical interventions, generating vast datasets that include study titles, outcomes, and criteria. Efficiently retrieving relevant trials from these datasets is crucial for researchers and healthcare professionals. However, traditional keyword-based search systems struggle to capture the semantic relationships in the data, making trial identification time-consuming and error-prone.

Problem Statement

The given clinical trial dataset includes study titles, primary and secondary outcome measures, and detailed criteria for participant selection. The challenge lies in building an efficient search system that matches queries, such as trial titles or criteria, to relevant entries. Additionally, it is vital to provide interpretable results to ensure trust and transparency. This study focuses on leveraging a sentence-transformer-based similarity search system and integrating explainability techniques like SHAP to enhance result quality and interpretability.

Model Selection

BioMedBERT:
BioMedBERT is a domain-specific model pretrained on biomedical data, designed to understand clinical trial text more effectively than general-purpose NLP models. It generates embeddings that accurately represent clinical trial data, enabling meaningful comparisons with user queries.

Mean Pooling Strategy:
Mean pooling aggregates token embeddings by averaging them to create fixed-size representations of variable-length text. This method is computationally efficient but may overlook important semantic nuances by treating all tokens equally.

Cosine Similarity:
Cosine similarity measures the semantic closeness between user queries and clinical trial text, making it an efficient and simple metric for retrieving relevant trials from large datasets.

LIME (Explainability):
LIME offers query-specific insights into the text's influence on similarity scores, ensuring transparency and building trust in the retrieval process.

Methodology

End to End ML Pipeline: Technical Flow

Data Collection

- **Source**: Clinical trial data is extracted from `usecase_1.csv`, including Study Title, Primary and Secondary Outcome Measures. Eligibility criteria are merged using `eligibilities.txt` based on `NCT_ID`.
- **Content**: The dataset comprises metadata fields like Study Title, Outcome Measures, and Eligibility Criteria.

Embedding Generation

- **Model**: Use the `preMedBERT` model, fine-tuned for clinical trial and biomedical text, to generate embeddings for each metadata field and the user query.
- **Pooling Strategy**: Apply a mean pooling strategy over the token embeddings produced by `preMedBERT` to create fixed-length vector representations.
- **Output**: Separate embeddings are generated for Study Title, Primary Outcome, Secondary Outcome, and Criteria, as well as for the user query.

Similarity Computation

- **Metric**: Compute cosine similarity between the query embedding (based on Study Title) and embeddings of each metadata field.
- **Weighted Scoring**: Assign weights to metadata fields based on their relevance to the query (Study Title: 0.7, Criteria: 0.3, Outcome Measures: 0.2 and 0.1).
- **Ranking**: Aggregate similarity scores to rank clinical trials, retrieving the top 10 most relevant results.

Explainability Integration

- **LIME**: Highlight the influence of specific features on the relevance of individual trials for a given query, offering local interpretability.

Model Evaluation Metrics

  - **Mean Reciprocal Rank (MRR)**: Assess the system's ability to rank relevant trials high in the search results.
  - **Top-K Accuracy**: Measure the proportion of queries where a relevant trial appears in the top K results.

Output

- **Ranked Trials**: Present the top 10 ranked trials, their similarity scores, and SHAP/LIME-based explanations for enhanced transparency.
- **User Interaction**: Provide visualizations and interpretability features to help users understand the ranking process.

Most similar trials:

	NCT Number	Study Title
50534	NCT04176536	A Study in People With Normal Kidney Function ...
0	NCT03302091	A Study in People With Normal Kidney Function ...
3933	NCT05718648	A Study to Test How BI 1015550 is Taken up in ...
67930	NCT05613036	A Study in People With Advanced Cancer to Test...
14514	NCT05718843	A Study to Test How Iclepertin is Taken up in ...
101096	NCT06352411	A Study to Test How BI 456906 is Taken up in t...
7825	NCT05863130	A Study in Healthy Men to Test How BI 764198 i...
103813	NCT05515328	A Study in Healthy Men to Test How BI 685509 i...
86287	NCT05421338	A Study in Healthy Men to Test How BI 456906 i...
32584	NCT05833035	A Study in Healthy Men to Test How BI 1291583 ...

Sample Similarity Retrieval Output for NCT03302091

Results and visualization

Model Outcomes (Evaluation Metric Scores)

Top-K Accuracy (99.6%):

- Top-K accuracy measures how often the correct answer appears within the top K (10 here) predictions.
- This indicates that for **99.6% of the queries**, the correct entry (the "gold standard") is present in the top 10 results returned by the model.

Mean Reciprocal Rank (MRR): 0.9626

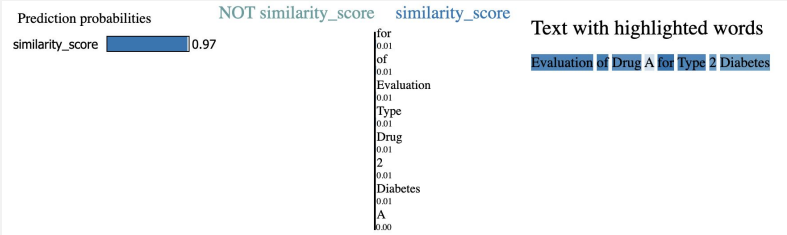
- The MRR reflects how high the correct result ranks in the predictions. A value of **0.96** means that, on average, the correct result is **very close to the top** (i.e., close to the 1st position in most cases).
- For reference:
  - An MRR of **1.0** means the correct result is always ranked 1st.
  - An MRR of **0.96** indicates that the system almost always ranks the correct result within the top 1 or 2 positions.

```
Top-10 Accuracy (Random Sample): 0.996
Mean Reciprocal Rank (MRR, Random Sample): 0.9626000344325942
```

Note: Here 100 random trials from the dataset were selected for evaluation

Explainability

LIME Explanations for the query text: “Evaluation of Drug A for Type 2 Diabetes”:



- The similarity score is **predominantly driven by the overall context of the query** rather than specific individual words.
- Every word in the query contributes equally and minimally to the prediction. This might suggest that:
  - The model relies on **global semantic similarity** between embeddings rather than specific token-level information.
  - The embedding model or similarity metric might smooth out word-level importance, leading to this uniform contribution.