Modeling

The following is the automated pipeline of the models from data ingestion to results visualization [0]

1. Download Libraries and Pretrained models
2. Data Extraction through Clinical Trials API
3. Data Cleansing and Transformation
4. Data Embedding through the model stsb_mpnet_base_v2 (details below)
5. Data Embedding through the model stsb-roberta-base-v2 (benchmarking below).
6. Data Filtering based on the patients' input
7. Compute inclusion criteria similarity with the patients' input.
8. Results visualization on Google Colab.

We have Used Sentence Transformers which is a Python framework for state-of-the-art sentence, text, and image embeddings. The initial work is described in the paper Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [1]
The Sentence Transformers we have used are

- stsb-roberta-base-v2
- stsb_mpnet_base_v2(benchmarking below)

These are sentence-transformers models: It maps sentences & paragraphs to a 768-dimensional dense vector space and can be used for tasks like clustering or semantic search.

Benchmarking SentenceTransformer Pretrained model [2]

Model Name	Base Model	Pooling	Training Data	STSb Performance (Higher = Better)	Speed (Sent. / Sec on V100 GPU)
stsb-mpnet-base-v2	mpnet-base	Mean Pooling	NLI+STSb	88,57	2800
stsb-roberta-base-v2	roberta-base	Mean Pooling	NLI+STSb	87,21	2300
stsb-distilroberta-base-v2	distilroberta-base	Mean Pooling	NLI+STSb	86,41	4000

nli-mpnet-base-v2	mpnet-base	Mean Pooling	NLI	86,53	2800
stsb-roberta-large	roberta-base	Mean Pooling	NLI+STSb	86,39	830
nli-roberta-base-v2	roberta-base	Mean Pooling	NLI	85,54	2300
stsb-roberta-base	roberta-base	Mean Pooling	NLI+STSb	85,44	2300

More ground to cover for a pretrained AI models suited for Clinical Trials data[3]

BERT models were pre-trained on the joint general-domain corpus of English Wikipedia and BooksCorpus, with the total of 3.3B tokens. Two domain-specific version of BERT are of interest for our task: BioBERT [45], pre-trained on a large biomedical corpus of PubMed abstracts and PMC full-text articles comprising 18B tokens, added to the initial BERT training data; and SciBERT [46], pre-trained on a corpus of scientific texts with the total of 3.1B tokens, in addition to the initial BERT training corpus.

BioBERT has only cased model, with a few versions with different pre-training data (PubMed abstracts only, PMC full-text articles only, or both). We used the model pre-trained on both datasets. SciBERT provides both cased and uncased models and has two versions of vocabulary: BaseVocab (the initial BERT general-domain vocabulary) and SciVocab (the vocabulary built on the scientific corpus). The uncased model with SciVocab is recommended by the authors, as this models showed the best performance in their experiments. We tested both cased and uncased models with SciVocab.

The Bio_ClinicalBERT model was trained on all notes from MIMIC III, a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA. For more details on MIMIC, see here. All notes from the NOTEEVENTS table were included (~880M words).[4] This model is based on this research paper [5]

Data Engineering

The new data engineering pipelines have the following updates

1. After extracting inclusion criteria from eligibility criteria we convert the paragraphs into a list of separated inclusion criteria, then stack each clinical trial numbered inclusion criteria on top of each other (repetitive process aggregated into one data frame).

2. The same steps were applied on Exclusion Criteria, then exported into Excel format, where each data frame has the clinical trial number, the numbered Exclusion criteria, and the raw Exclusion criteria (the same for inclusion criteria).

3. The inclusion criteria are also fed to an automatic pipeline to normalize words and collection of words based on a dictionary provided by Clinical Net.

4. Inclusion and exclusion criteria were also separated into n-grams of words to determine which are the most frequent n-grams of words(before that text was cleansed and normalized).

5. The filtering and preprocessing pipeline are based on the patient input and the features of the clinical trial: HealthyVolunteer, Age, Gender, Phase, LocationStatus(LocationCountry, LocationCity, LocationFacility), patient willingness to travel destination(Country, City), and Condition.

P.S.

1. The data engineering pipelines notebooks in the Github repository do not have a specific order, they're still in the research phase, and the same for modeling notebooks.

2. Once we identify the final components for the AI system that will be implemented, there we'll be a specific repository that we'll have the python files ready for the production phase hosted on Machine Learning cloud services such as AML studio or Vertex AI.

Conclusion

The first approach of using Semantic Text Similarity on the whole patient input compared to the Clinical Trials' data inclusion criteria paragraphs is less performant than having a filtering pipeline based on the clinical trials data where specific questions driven from the dataset are asked to the patients to recommend them specific clinical trials.

There are still more grounds to cover, where we need to build a pipeline to filter out inclusion criteria that have been answered by the client and ask him about the remaining inclusion criteria that there was no predetermined question for them. Exclusion criteria are also very rich in pieces of information that can be used to drive questions for them to filter to clinical trials. In addition, Clinical Net feedback about the n-grams for both inclusion and exclusion criteria will be essential to the next steps.

Finally, the use of Cloud Services such as Microsoft Azure Machine LEarning of Google Vertex AI will be mandatory, since the source code is public on Github for now, because we're committing files from Google Colab to Github directly, and for that, the github repository need to stay public, which is risky. Moreover, using Google Colab is inefficient, since at each working session we need to restart all the operations, since Colab does not save working sessions including the data, and these free tools are not suitable for the production environment.

Citations

[0]
Title = Models
Url = https://github.com/MWFK/NLP-Semantic-Similarity/tree/main/ClinicalTrials/Models

[1]
Title = "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks",
Url = "http://arxiv.org/abs/1908.10084"

[2]
Title = SentenceTransformer Pretrained Models
Url =
https://docs.google.com/spreadsheets/d/14QplCdTCDwEmTqrn1LH4yrbKvdogK4oQvYO1K1aPR5M/edit#gid=0

[3]
Title = Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations
Url = https://www.sciencedirect.com/science/article/pii/S2590177X19300575

[4]
Title = ClinicalBERT - Bio + Clinical BERT Model
Url =
https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT#clinicalbert---bio--clinical-bert-model

[5]
Title = Publicly Available Clinical BERT Embeddings
Url = https://arxiv.org/abs/1904.03323

[6]
Title = Data Engineering
Url =
https://github.com/MWFK/NLP-Semantic-Similarity/tree/main/ClinicalTrials/Data%20Engineering