PROJECT PROPOSAL – INFORMATION RETRIEVAL USING SEMANTIC SIMILARITY – BERT & CLINICALBERT – USE CASE COVID-19

Introduction and Motivation:

COVID-19 has resulted in an explosion of online questions related to the pandemic. While there is a rich collection of content on the web in the form of publications and articles, medical researchers would like to home in on the most relevant pieces of text in conjunction with their scope of research. Traditional search engines do a limited job in this context but generating a set of most relevant outputs through pieces of texts and the publications they come from as answers to queries would even further speed up the process of gathering data most pertinent to the interests of the researcher. We intend to answer the following question through this project:

1. What are the underlying topics within the medical case report in relation to COVID-19?
2. What are the similarities and dissimilarities between COVID-19 and the earlier pandemics?
3. How to speed up the pace of medical research and find relevant research publications without browsing all the existing publications for successful collaboration and innovation?

Literature Review Summary:

Semantic similarity uses the idea of likeliness of meaning of terms in document to calculate the distance between those terms. One of the prior researches done in this domain at Lakehead University, Canada suggests that semantic similarity methods can be broadly classified into four categories: Knowledge-based, Corpus-based, Deep neural network-based and a Hybrid method which uses a combination of the other methods. We intend to leverage multiple knowledge-based algorithms and come up with the best available algorithm to calculate the similarity score on our corpus.

Proposed Research Designs and Evaluation Techniques:

Stage 1: Data Assimilation and Understanding the Requirements Stage 2: Data Pre-Processing & Cleaning Stage 3: Feature Engineering and manipulations Stage 4: Similarity Implementation: BERT and Clinical BERT Stage 5: Model Evaluations and Improvement Stage 6: Prediction Tool Deployment	**Data Description** CORD-19 by Allen AI is a resource of over 300,000 scholarly articles. Total Rows: 318137 Columns: `paper_id, title, Abstract, authors, body_text` Data Set link: https://allenai.org/data/cord-19

Evaluation Metrics:

- Compute semantic similarity scores
- Compare and list the best scores based on a sentence level query
- Comparing Clinical Bert, Bert models with other standard techniques such as cosine similarity and BLEU
- Selecting model with best results based on prediction accuracy
- Research on other evaluation metrics to judge performance measure of information retrieval methodology

Appendix:
The outputs will be the Top 10 Articles most similar to the Query. Intended output:

```
Query: What is known about transmission, incubation, and environmental stab
ility of coronavirus
1. Article: Virus ecology: a gap between detection and prediction
Score: 1.1326019326045822

2. Article: Six weeks into the 2019 coronavirus disease (COVID-19) outbreak
- it is time to consider strategies to impede the emergence of new zoonotic
infections
Score: 1.055644503277209

3. Article: A Scenario-Based Evaluation of the Middle East Respiratory Synd
rome Coronavirus and the Hajj
Score: 1.032464095919743
```

Future Scope:

- Group articles based on research design – like – clinical trials, observational studies etc
- Organize content for the information retrieval into categories
- Leverage the LitCovid Data Set to validate the categories

References:

- https://arxiv.org/pdf/2006.09595.pdf - COVID-19 Information Retrieval with Semantic Search
- https://arxiv.org/pdf/2004.13820.pdf - Evolution of Semantic Similarity
- https://arxiv.org/pdf/1812.01808.pdf - Enriching Article Recommendation with Phrase Awareness
- https://www.ncbi.nlm.nih.gov/research/coronavirus/ - LitCovid