Information Retrieval using Semantic Similarity: Cosine, BERT & Clinical BERT

Akhil Nair, Aneesh Phatak, Shreyas Sadagopan, Sourabh Ghosh, Sankalp Singh

School of Information Studies, Syracuse University

Abstract

When a new virus is discovered and causes a pandemic, it is important for scientists to get information coming from all scientific sources that may help them combat the pandemic. The challenge, however, is that the number of scientific papers created is large and the papers are published very rapidly, making it nearly impossible for scientists to digest and understand what's important in this mass of data. We have developed and discussed a cosine similarity model, a deep learning model and a corpus specific scientific literature understanding system that can take in common terms and analyze a very large corpus of scientific papers and return highly relevant text excerpts from papers containing topical data relating to the common text inputted, allowing a single researcher to gather targeted information and quickly answer important questions about the new virus. We have developed and discussed an evaluation metric to identify the best performing IR algorithm in this paper.

1 Introduction

The COVID-19 pandemic has brought in a need for international efforts to understand, track and mitigate the disease, yielding a significant corpus of biomedical related publications for scientific research. The health practitioners and medical researchers will require specialized tools to keep up with the literature. COVID-19 has also 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 hone 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. In this paper, the technical goal is to create an Information Retrieval system and rapidly search through a corpus to find relevant information to address a particular information needed.

2 Research Questions

We aim to investigate the following research questions:

RQ1. How to identify the semantically relevant documents to a given query using NLP techniques?

For this, we have designed an automated query pipeline to search for relevant findings in the reference corpus. This would possibly speed up the search process and make an efficient retrieval-based system. This would also help doctors and medical professionals to arrive at highly relevant text excerpts by analyzing a large corpus of scientific papers.

RQ2. How to evaluate inter-model performances and assess important feature differences?

For this, we have implemented three machine learning models: 1) Feature engineering using TF-IDF weights with Cosine Similarity metric; 2) Transformer based BERT model with contextual

BERT Embeddings; 3) Clinical Bert Model Pre-Trained on the MED_STS Database.

RQ3. What are the major annotation rules while annotating the relevancy? What are the major methods the models use to compute the similarity?

3 Related Work

The versatility of natural language makes it difficult to define rule-based methods for determining semantic similarity measures. In order to address this issue, various semantic similarity methods have been proposed over the years (Chandrasekaran et al., 2020).

The techniques like Bag of Words (BoW), Term Frequency - Inverse Document Frequency (TF-IDF) were used to represent text, as real value vectors to aid calculation of semantic similarity. However, these techniques did not attribute to the fact that words have different meanings and different words can be used to represent a similar concept. For example, consider two sentences "John and David studied Maths and Science." and "John studied Maths and David studied Science." Though these two sentences have exactly the same words, but they do not convey the same meaning. To address these drawbacks of the lexical measures various semantic similarity techniques have been proposed over the past three decades.

Let us briefly look at the advantages and disadvantages of the 4 most broadly used methods i.e., knowledge-based, corpus-based, deep neural network-based and a hybrid method which uses a combination of other methods. Knowledge-based methods takes into consideration the actual meaning of text however, they are not adaptable across different domains and languages. Corpus-based methods have a statistical background and can be implemented across languages, but they do not take into consideration the actual meaning of the text. Deep neural network-based methods show better performance, but they require high computational resources and lack interpretability. Hybrid methods are formed to take advantage of the benefits from different methods compensating the shortcomings of each other. Each method has its advantages and disadvantages, and it is difficult to choose one best model, however, most recent hybrid methods have shown promising results over other independent models. Hence, we intend to leverage multiple knowledge-based algorithms and come up with the best available algorithm to calculate the similarity score on our corpus.

4 Research Design and Methodology

To answer our research questions, we developed a state-of-the-art intelligent framework to extract the most scientifically relevant publications regarding COVID-19's origin and evolution and analyze and summarize information at sentence level by combining multiple advanced NLP techniques (BOW, TF-IDF, sentence embedding, BERT) and deep learning techniques. Our research was fragmented and broken down into multiple simpler modules inspired from the prior clinical-NLP works such as (Lee CH et al). We employed special processing and manipulation techniques required for medical databases (Sarab AlMuhaideba et al).

4.1 Data Understanding and Pre-Processing

The data contains of 20000 journals with their abstract, body text and the authors. We processed the data to clean the text with the following cleaning process and then tokenized both abstract and body text with functionality in NLTK package (nltk.word_tokenize) (Wencheng et al.). We converted and wrangled the data from json files to pandas format. We dropped unrelated columns like URL, Author details etc. We removed all articles that had fewer than 1000 words and also removed all non-English articles.

We replaced brackets, contractions and punctuations. We converted all sentences to lower case and performed stemming after lemmatization. We also removed punctuations and stop words. Our models have been built using the 'Abstract' as the Input text.

4.2 Data Explorations

We visualized the text corpus that we created after pre-processing to get insights on the most frequently used words, the bigrams and the trigrams. This process helped us to understand the intricacies in the data set and manipulate the variables accordingly. The top occurring words as expected are related to corona virus and pandemics.

Figure1: Word cloud to visualize the word occurrences present in our Covid corpus

4.3 Data Engineering and Manipulations

To build our models, we engineered the following features and restricted the data for further analysis. We created word and sentence tokens and created word vectors using tf-idf vectorizer. We calculated the tf-idf features and investigated the top features to get an insight on our corpus. We created sentence embeddings using sentence transformers. After these steps, we had all the necessary features, and the data was in the format required for our models.

4.4 Building the Predictive Models

Query Set. We identified a set of 5 questions relevant to Covid-19 as our query base. We ran our model for each of the below mentioned 5 queries. (https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html)

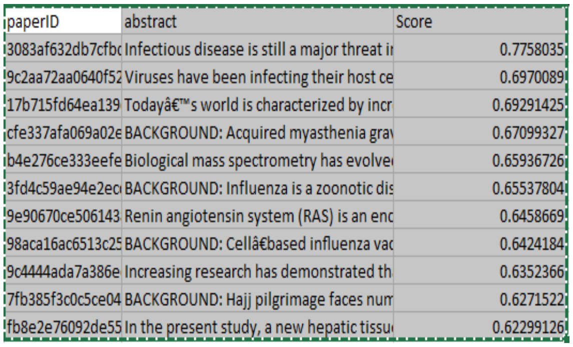
Q1. What is known about transmission, incubation, and environmental stability of coronavirus?
Q2. What do we know about coronavirus risk factors?
Q3. What do we know about coronavirus genetics, origin, and evolution?
Q4. What do we know about vaccines and therapeutics for coronavirus?
Q5. What has been published about coronavirus medical care?

Model 1. (Zhixiang (Eddie) Xu et al, Yanshan Wang et al.)
We implemented a state-of-the-art intelligent search engine based on the advanced machine learning and text mining techniques. In the step, we used the tf-idf and Bag of words features to compute the cosine similarity between the given query and the corpus to retrieve the most relevant articles.

Model 2. (Yuhan Su et al., Zongcheng et al., Ronghui et al.)
In the second model, we developed a semantically inspired BERT based search engine with a view to extend the framework's ability to extend information extraction at the sentence level. Another key purpose was to dig out further insights into our refined queries, which represents our specific scientific interest in this area. Word embeddings give us context dependent vectorized representation of all the Sentences. We used the transformer-based BERT Model to build an IR model.

Model 3. (Debasmita Das et al., Yuxia Wang et al., Jinhyuk Lee et al.)
Finally, we implemented a fine-tuned, Clinical BERT-based semantic text similarity model which leverages PyTorch transformers and is a ready to use algorithm which has been pre-trained on the MED-STS corpus. Since this has been trained on a relevant corpus, this has been

established as a benchmark to evaluate the performances of the other user defined models.



paperID	abstract	Score
3083af632db7cfbc	Infectious disease is still a major threat ir	0.7758035
9c2aa72aa0640f52	Viruses have been infecting their host ce	0.6970089
17b715fd64ea139	Todayâ€™s world is characterized by incr	0.69291425
cfe337afa069a02e	BACKGROUND: Acquired myasthenia grav	0.67099327
b4e276ce333eefe	Biological mass spectrometry has evolve	0.65936726
3fd4c59ae94e2ec	BACKGROUND: Influenza is a zoonotic dis	0.65537804
9e90670ce506143	Renin angiotensin system (RAS) is an enc	0.6458669
98aca16ac6513c25	BACKGROUND: Cellâ€based influenza vac	0.6424184
9c4444ada7a386e	Increasing research has demonstrated th	0.6352366
7fb385f3c0c5ce04	BACKGROUND: Hajj pilgrimage faces num	0.6271522
fb8e2e76092de55	In the present study, a new hepatic tissu	0.62299126

Table1: Sample Output from the Clinical Bert Model

4.5 Developing the Evaluation Metrics

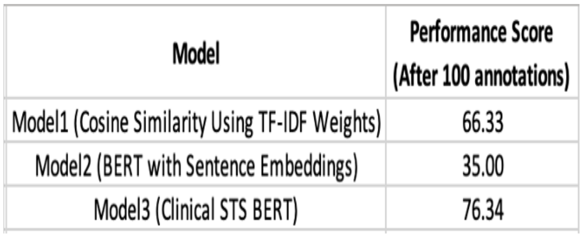
Annotations. In order to compare the above models, since all the models are unsupervised, there are no common evaluation strategies. We have developed a common metric by annotating the relevance of the returned results for each of the queries for all the models.

Strategy. Manual annotations based on the closeness of the returned results for a given query is determined using two levels (Relevant or Non-Relevant). We performed annotations for 5 queries for the top 20 results. So, a sum total of 100 annotations per model were labeled. All 5 authors acted as annotators and medical experts. The best model is the model with the best relevancy score. We have developed a comprehensive and a reproducible ruleset which can be implemented on any given dataset. The proposed complexity factors include discrimination, boundary delimitation, expressiveness of the language, degree of ambiguity and the weight of the context.

Annotation Ruleset. We extracted subject, verb & object clause from the sentence to identify the primary entities and the presence of "is-a" relationship**.** We checked if the predicted document covers maximum topics that optimally represents the input query**.** Our annotations were made on rules and keywords such as - severity, fatality**,** children, adult, old population at risk, health workers at risk**,** symptoms, critical, patients, epidemic, hospitalized**,** mortality, vulnerability, prevention, control respiratory infections, bronchitis, gastrointestinal.

5. Results and Discussion

We have calculated the performance based on the relevancy scores achieved by each of the models after 100 manual annotations. Clinical STS BERT model performed the best with a relevancy score of 76.34%. Cosine similarity using TF-IDF weights model performed with a relevancy score of 66.33%**.** BERT model with sentence embeddings performed with a relevancy score of 35%**.**



Model	Performance Score (After 100 annotations)
Model1 (Cosine Similarity Using TF-IDF Weights)	66.33
Model2 (BERT with Sentence Embeddings)	35.00
Model3 (Clinical STS BERT)	76.34

Table2: Performance relevancy scores for the three models implemented after 100 annotations

The Relevancy score was computed as a difference between the most similar results returned by the model and agreed by human annotators as relevant answer for a given input query based on our annotation rules. Model 1 using tf-idf features performed with better relevance since it was able to capture word level importance of the results mapping with query. Model 2 uses word embeddings as feature vector and therefore failed to capture entity level information such as "COVID", "Epidemiology" etc. and performed with less relevant answers to the queries. It also captured information from other viruses and pandemics. Model 3 pre trained on Clinical and health corpus performed with most relevance as it was able to capture similarities between health-related context.

6. Conclusion and Future Work

Based on our assumptions and model evaluations, we were able to answer our research questions for "Information Retrieval using Semantic Similarity - Cosine, Bert and Clinical Bert". We were able to perform model comparison using relevancy scores tested on same queries across all models (here the relevant truth was assumed to be human annotations). Using the metrics defined for one of our research questions, we can state that resultant excerpts for queries using a Clinical Bert will arrive at highly relevant text excerpt, with regards to COVID with a relevancy score of 76.34%. We were also able to formulate a comprehensive reproducible annotation rule set during our analysis.

In the future, we plan to look into the topic variance between COVID and prior pandemics and develop a LDA based topic model. We also intend to extend and test the model evaluation for queries capturing information for other viruses**.** We will also try to categorize our results into various categories similar to that of the LitCovid corpus in the future.

References

Zhixiang (Eddie) Xu. An alternative text representation to TF-IDF and Bag-of-Words ∗

Yuhan Su, 1 Hongxin Xiang, 1 Haotian Xie, 2 Yong Yu, 1 Shiyan Dong, 3 Zhaogang Yang,☒ 3 and Na Zhao☒ 1. Application of BERT to Enable Gene Classification Based on Clinical Evidence

Debasmita Das, Yatin Katyal, Janu Verma, Shashank Dubey, AakashDeep Singh, Kushagra Agarwal, Sourojit Bhaduri, RajeshKumar. Information Retrieval and Extraction on COVID-19 Clinical Articles Using Graph Community Detection and Bio-BERT Embeddings

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining

Zongcheng Ji, PhD,1 Qiang Wei, MS,1 and Hua Xu, PhD1. BERT-based Ranking for Biomedical Entity Normalization

Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, Shanfeng Zhu. BERTMeSH: Deep Contextual Representation Learning for Large-scale High-performance MeSH Indexing with Full Text

Yuxia Wang, Fei Liu, Karin Verspoor, Timothy Baldwin. Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity

Yanshan Wang,a,* Stephen Wu,b Dingcheng Li,a Saeed Mehrabi,a and Hongfang Liua,*. A Part-Of-Speech Term Weighting Scheme for Biomedical Information Retrieval

Sarab AlMuhaideba, Mohamed El Bachir Menaib aDepartment of Computer Science, Prince Sultan University, 66833 Riyadh 11586, SA. An individualized preprocessing for medical data classification

Wencheng Sun, 1 Zhiping Cai,☒ 1 Yangyang Li, 2 Fang Liu, 3 Shengqun Fang, 1 and Guoyan Wang 4 . Data Processing and Text Mining Technologies on Electronic Medical Records: A Review

Donald A Szlosek, MPHi and Jonathan Ferrettii . Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems

Liuqing Li#, Jack Geissinger#, William A. Ingram, Edward A. Fox*. Teaching Natural Language Processing through Big Data Text Summarization with Problem-Based Learning

Irena Spasic1, PhD ; Goran Nenadic2, PhD. Clinical Text Data in Machine Learning: Systematic Review

Dhivya Chandrasekaran, Vijay Mago. Evolution of Sematic Similarity – A Survey

COVID-19 Information Retrieval with Semantic Search

Enriching Article Recommendation with Phrase Awareness

LitCovid.
https://www.ncbi.nlm.nih.gov/research/coronavirus/

Dataset Link.
https://allenai.org/data/cord-19