

A Transformer-Based Q&A System for Analyzing the COVID-19 Literature: Application of BERT and BERTopic

Jonathan Bonilla
jonathan.bonilla@torontomu.ca,
Department of Computer Science
Toronto Metropolitan University
Toronto, Canada

Bojing Yao
bojing.yao@torontomu.ca,
Department of Computer Science
Toronto Metropolitan University
Toronto, Canada

Victor Ogunjobi
victor.ogunjobi@torontomu.ca,
Department of Engineering Innovation
Toronto Metropolitan University
Toronto, Canada

Abstract — This project aims to evaluate the feasibility of creating a Transformer-based Question-Answering (Q&A) model tailored to the COVID-19 Open Research Dataset (CORD-19). The team performed exploratory data analysis including TF-IDF and n-grams analysis, used BERTopic to identify key topics, generated extractive summaries using LexRank, and produced abstractive summaries using the BART model. Subsequently, a Transformer-based extractive Q&A model was trained using a Q&A dataset generated by applying a zero-shot Large-Language Model (LLM) on entities identified from the extractive summaries via spaCy’s part-of-speech (POS) tagging. Additionally, the project involved extensive experimentation with various models and hyperparameters to optimize Q&A model training. This project demonstrates innovative methodologies and promising results on repurposing a Transformer model on a new knowledge base through performing zero-shot question generation using a LLM.

I. INTRODUCTION

With the recent advancements in Artificial Intelligence (AI) models and the surplus of data available to us, our team has focused on developing a reliable solution that can adapt off-the-shelf AI models to new and specific subjects. For this project, we have selected the CORD-19 dataset, a comprehensive collection of research papers related to COVID-19 published prior to 2020. This dataset includes a range of specialized terminology and complex contexts that are often inaccessible to LLMs. Our aim with this project is to develop a reliable solution for creating a Q&A system that can be easily adapted to any knowledge base. Our contributions are as follows:

1. **Exploratory Data Analysis:** We conducted detailed exploratory data analyses, including TF-IDF and n-grams to understand the linguistic characteristics of the CORD-19 dataset.
2. **Topic Identification and Summary Generation:** Utilizing the BERTopic model, we identified key topics within the dataset, generating extractive summaries with LexRank. These were further refined into abstractive summaries using the BART model to enhance readability and utility.
3. **Development of Q&A Model:** We developed a responsive Q&A model using DistilBERT and BERT, which we then hosted on a public server with an interactive website. This setup allows real-time interaction and broadens the accessibility and application of our findings.

Through the combined efforts listed above, we can improve the effectiveness of adopting pretrained AI-driven systems for any complex or foreign knowledge bases.

II. MODEL OVERVIEW

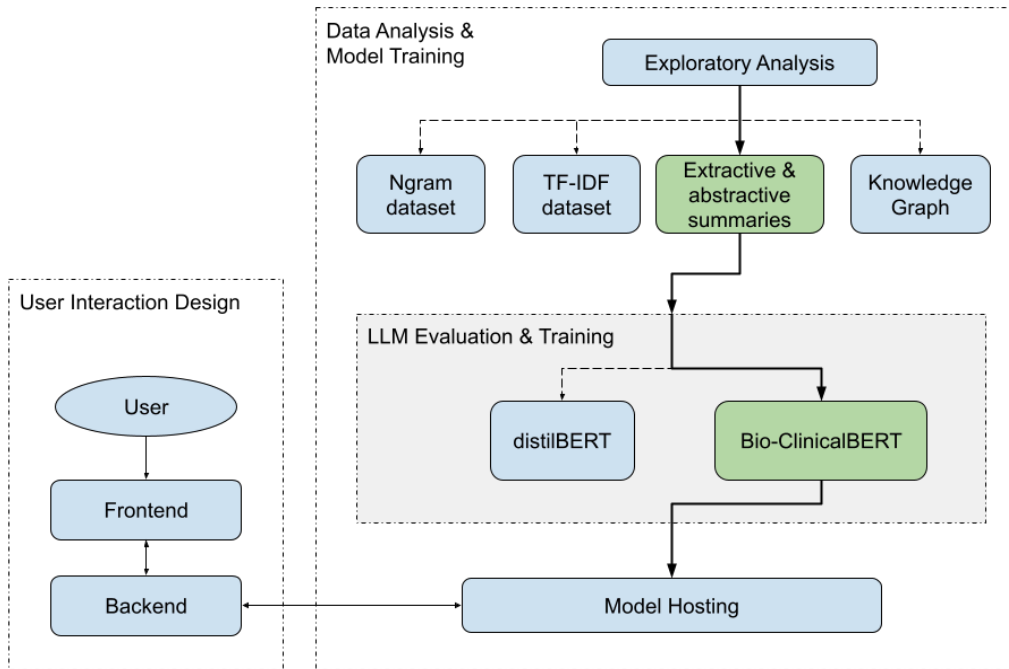


Figure 1 - Overview of System Structure

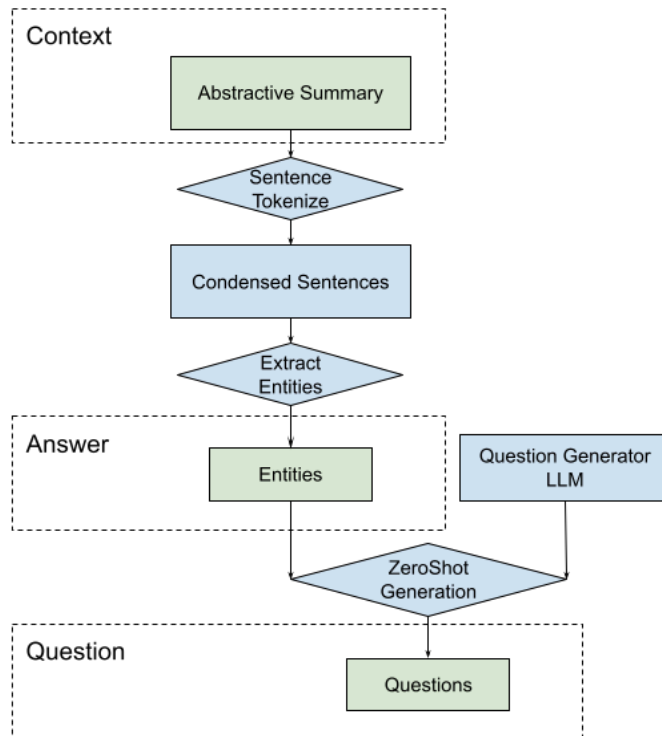


Figure 2 - Overview of QA Pair Generation

III. FORMAL DESCRIPTION

Initial steps were taken to generate TF-IDF scores based on the unigrams, bigrams and trigrams of this dataset. This is to better understand the key terms associated with the CORD-19 dataset and visualize the differences between the vocabularies in typical English sentences and the terminologies involved with the CORD-19 dataset. Initial analysis showcases a heavy focus on technical jargon used in this dataset compared to regular English documents. Overview of the system is showcased in Figure 1.

Topic Selection

The solution proposed by this project begins by obtaining a distribution of main topics among the dataset through BERTopic which is initiated with the KeyBERTInspired representation model to improve keyword extraction and improve topic relevance. A clean English dataset is obtained from the original CORD-19 dataset, and the paper_id attribute is used to distinguish each paper. The BERTopic package is used to produce clusters of the major topics discussed in the paper. The BERTopic model first vectorizes the documents before performing analysis on the documents through unigram, bigram and trigram analysis. Although the majority of the research papers were highly coupled with generic COVID-19 terms and research topics, 35 major topic clusters were still identified. A detailed overview of the topics can be found in the Appendix.

Summarization

After identifying the major topic clusters within the dataset, the most representative documents in the dataset for each topic were selected. To do this, we initially attempted to utilize BERTopic’s built-in method `get_representative_docs`. However, despite multiple attempts to run the BERTopic model with various parameters in hopes of getting it to identify the most representative documents, the model consistently failed to identify any. As a result, we decided to create our own implementation of the `get_representative_docs` function based on our analysis of the function’s source code. Our implementation works by using the document-topic probabilities matrix generated by BERTopic to find the indices of the three documents with the highest matching probabilities for each topic identified.

Once the representative documents were identified, for each topic we selected all sentences containing any of the keywords generated by BERTopic and stored them in a dictionary. This dictionary was subsequently used to generate an extractive summary with a maximum length of 3 sentences. These extractive summaries were then used as the input for generating the abstractive summaries, which were made with the BART model from the Transformers library. Due to BART’s restriction of accepting only 1024 tokens per input, we introduced a chunking method to divide our extractive summaries into 1024 token chunks. The abstractive summaries then served as the final summary for each topic.

Q&A Model

To develop our Q&A model, we first tokenized each topic summary into discrete sentences to then be processed by spaCy for entity extraction. Utilizing the `bart_squad_qg_hl` model from Hugging Face, we then generated questions using a zero-shot approach with the extracted entities serving as the answers and the original summaries providing context.

During the first iteration of our Q&A model, we initially chose DistilBERT for its efficiency on limited hardware and also its reduced complexity. However, DistilBERT did not yield satisfactory results, prompting us to search for alternative models for our next iteration. We ultimately selected BioClinical-BERT as it was particularly well-suited to our dataset due its specialization in biomedical research. This second iteration of the model demonstrated better and more accurate results.



<p>Algorithm: Extractive Summary Generation</p> <p>Input: Representative documents, topic_keywords</p> <p>Output: Extractive summaries</p> <ol style="list-style-type: none"> 1. Load BERTopic model 2. For each topic <ol style="list-style-type: none"> a. Identify and extract sentences with topic keywords from representative documents. b. Apply LexRank to generate an extractive summary from these sentences. 3. Return extractive summaries for each topic. 
<p>Algorithm: Abstractive Summary Generation</p> <p>Input: Extractive summaries</p> <p>Output: Abstractive summaries</p> <ol style="list-style-type: none"> 1. For each extractive summary <ol style="list-style-type: none"> a. Chunk extractive summary into 1024 token chunks b. For each chunk <ol style="list-style-type: none"> i. Use the BART model to generate an abstractive summary. c. Concatenate chunk summaries together into one abstractive summary for each topic. 2. Return abstractive summaries corresponding to each topic. 
<p>Algorithm: Q&A Model Creation</p> <p>Input: Extractive summaries</p> <p>Output: Question-Answer pairs</p> <ol style="list-style-type: none"> 1. Load bart-squad-qg-hl BART model and spaCy NLP model. 2. Extract sentences from summaries and identify entities using spaCy. 3. Generate questions using BART based on identified entities. 4. Pair questions with their corresponding entities as answers. 5. Store and export question-answer pairs to a CSV for use and analysis.

Table 1. Algorithm box detailing summary generation and Q&A model creation

Website

The user interaction component for this project includes a frontend and a backend. A simple user interface is built with NextJS, which captures user queries and a backend that processes these queries. The backend identifies the most relevant abstractive summary as context, sends it with the query to our HuggingFace-hosted model, and then displays the model's response on the frontend.

HuggingFace is used to host the model due to its easy-to-use interface and its versatility to scale. HuggingFace also works with the Transformers library for a pipeline that incorporates model hosting directly after model training. This setup ensures the model is always up to date, and the project group can also leverage the various scaling capabilities provided by HuggingFace for any opportunity to publish the project to production environment.

On the backend, Express and Node.js are used to build a flexible and efficient server. The server is hosted through Render, a cloud platform that includes automatic scaling and efficient deployment process. The frontend is constructed through NextJS for simple responsive design, and hosted on Vercel, which is a platform optimized for frontend deployment.

This user interaction design provides a robust and scalable solution for our application. This setup allows us to focus on developing features and improving our service, confident in the infrastructure that supports our application.

Differences

The main ways our project deviates from typical implementations include developing a custom function to select representational documents from our dataset. Additionally, we adopted BioClinical-BERT, a model tailored for biomedical documents. To address BART's token limitations, we implemented a method for dividing summaries into manageable chunks so that no information was lost in the generation of abstract summaries. Furthermore, we combined extractive and abstractive summarization methods to speed up the summarization process. Finally, we incorporated zero-shot question generation allowing us to generate questions without prior task-specific training.

IV. LITERATURE REVIEW

For our literature review, we will reference both foundational and recent studies to provide context for our approach in text analysis and question-answering systems.

1. **BERT [1]**: Devlin et al. introduced a new way of pre-training language representations that allows BERT to understand deep bidirectional contexts within text, making it powerful for a wide range of tasks including question answering. This has significantly influenced NLP applications by improving the understanding of context in text. The deep contextual understanding enabled by BERT is foundational to our project, allowing our model to interpret the complex language used in CORD-19 papers.
2. **BERTopic [2]**: Grootendorst developed BERTopic to utilize BERT embeddings for flexible and efficient topic modeling, which is essential for processing and summarizing large datasets like CORD-19. BERTopic's advanced topic modeling technique helps our system effectively categorize and summarize the dense and complex CORD-19 literature, making the data easier to navigate and parse.
3. **Navigating the Landscape of Large Language Models [3]**: Weng's comprehensive review examines the scalability and adaptability of large language models, highlighting their use in specialized fields like biomedical literature. It provides important insights into their practical applications and limitations, essential for their effective implementation.

These citations inform the technical foundation of our project and also align closely with our goal to create an adaptable and efficient question-answering system tailored to the complexities of the CORD-19 dataset.

V. RESULT ANALYSIS & COMPARISON

Comparison is performed between a baseline model and the trained model. DistilBERT is chosen as the baseline model and the fine tuned model is a variation of BERT base uncased that is finetuned on biomedical data. As showcased in Table 2, the qualitative analysis is performed by feeding the model the same question and the context in which the models can extract a correct answer from. The difference in the phrase extracted is compared and the quality of the final model is inferred. As can be seen from the table, the answers extracted by the chosen model are more accurate than the baseline model. In cases where both the baseline model and the chosen model's answers are not accurate, the chosen model's output is slightly more relevant to the question instead of being a random phrase from the context. Judging by the baseline model's output, the baseline model is not capable of understanding the technical vocabularies in the dataset and cannot produce meaningful responses. The baseline model is also not capable of identifying key phrases in the context, as its outputs to many different questions are the same. This comparison verifies that the chosen model is well adopted for the CORD-19 dataset.

Question	Expected	Baseline_model (distilbert)	finetuned_model
Who announced the Novel Coronavirus SARS-CoV-2 COVID-19 as a pandemic outbreak?	The World Health Organization	COVID-19 in Central Italy. The main stressors	World Health
What are not always available in many medical institutions?	basic protective equipment	COVID-19 in Central Italy. The main stressors	medical institutions

What are the main stressors?	prolonged periods of work in isolation, high workloads, compassion fatigue, and a lack of time for physical activity, meditation, or relaxation	COVID-19 as a pandemic	main stressors
What is the most common cause of death in dogs?	Cancer	pandemic. college	Cancer
How many dogs are newly diagnosed with cancer per year?	46 million	pandemic. college	46 million dogs

Table 2. Comparison of outputs between baseline model and fine-tuned model

VI. LIMITATIONS

There are four key limitations of this project, including the lack of computational resources for the Q&A model training, the lack of a human-evaluator for the Zero-Shot question generation, the potential for information loss in the summarization process, and the need for a more sophisticated method of checking query-context similarity. As the model training is performed locally on a non-CUDA system, the processing time is much longer than expected, resulting in a limited number of trials performed and models evaluated. This causes the chosen model to have a higher loss than desired, and therefore decreased accuracy in generating a correct answer. The lowered number of models evaluated prior to training might result in the project group not having access to a better model that is more adept at handling biomedical information. Moreover, as our systems were limited in their ability to process large amounts of textual information for abstractive summary generation, some information may not be captured accurately in our extractive summaries. This can lead to inaccurate answers as the contexts that are provided may not be sufficient for answering more complex queries. Finally, our current system employs a very simple cosine similarity check for user queries which can result in irrelevant contexts being chosen due to the similarity of their vector representations, rather than their semantic similarity.

As this project aims to demonstrate the feasibility of a solution and a prototype Q&A platform, these limitations are not critical, however, they may be overcome for any future work in order to produce a better model. Recommendations for improvement include a systematic way of evaluating off-the-shelf models, preparing a CUDA enabled cloud environment with better hardware for model training, generating abstractive summaries on documents themselves rather than on extractive summaries, and implementing a more sophisticated approach to checking similarity between user queries and potential contexts.

VII. CONCLUSION

This project produces a feasible solution of developing a Question-Answering system tailored for the CORD-19 dataset. Through using a pipeline of various methods that include exploratory data analysis, topic identification and summary generation, zero shot question generation, this project demonstrates adaptability that can be used on any knowledge base. Qualitative comparison with a baseline Q&A system shows that this project's innovative methods yield improved performances, particularly in understanding complex technical terms.

Despite these promising results, there are still areas in need of improvements such as limited training time and a lack of variety of models evaluated, and the need for a refined summary generation process. Future implementations should consider more powerful machines and refine the summarization strategies.

To conclude, this project produced an innovative step towards specialized question-answering systems, integrating various analytical techniques as well as large language models, and takes one step closer towards building adaptive and accurate question-answering systems that cater to any specialized datasets.

VIII. REFERENCES

- [1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [2] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- [3] Weng, B. (2024). Navigating the Landscape of Large Language Models: A Comprehensive Review and Analysis of Paradigms and Fine-Tuning Strategies. *arXiv preprint arXiv:2404.09022*.

IX. APPENDIX A

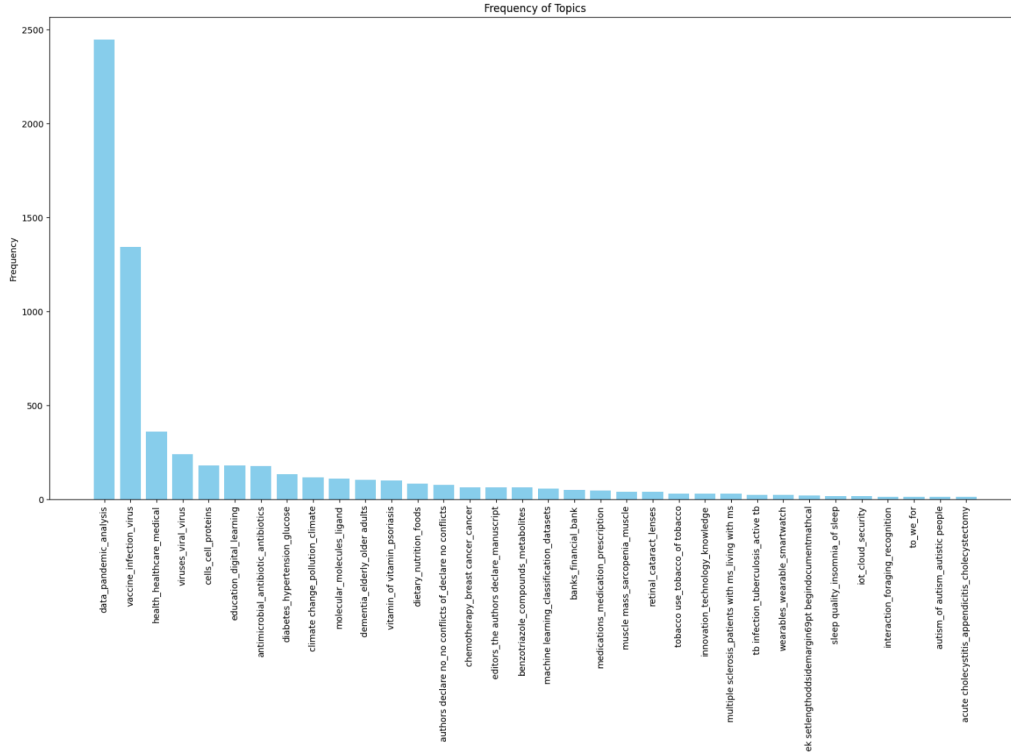


Figure A1 - Barchart of BERTopic output topic distributions, Frequency of topic vs topic labels

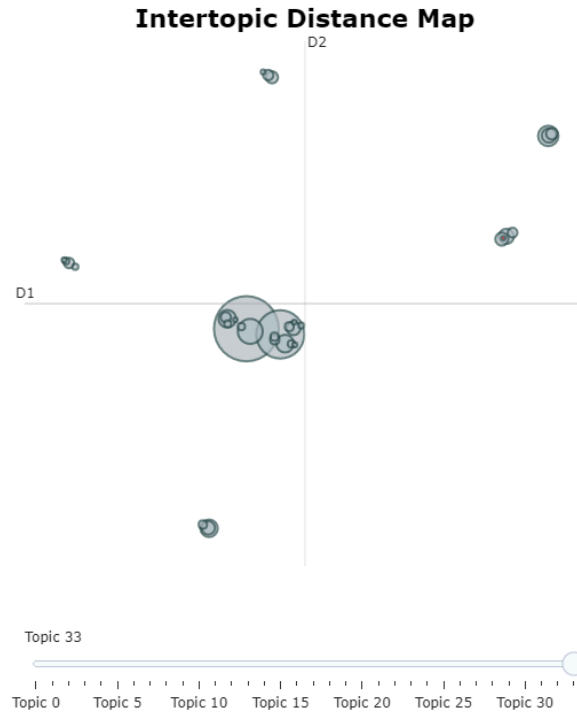


Figure A2 - Graph of distance between topics