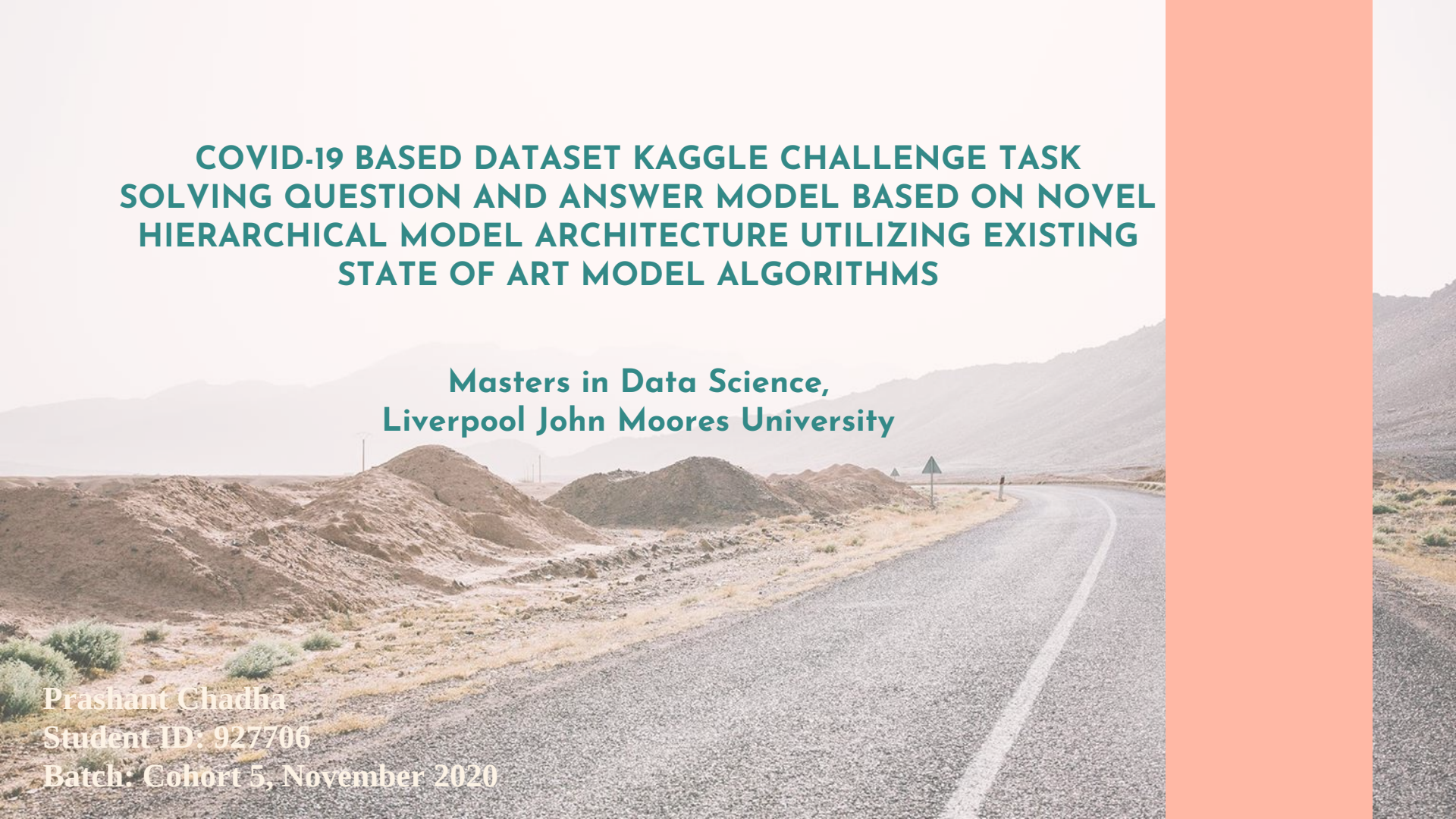
COVID-19 BASED DATASET KAGGLE CHALLENGE TASK SOLVING QUESTION AND ANSWER MODEL BASED ON NOVEL HIERARCHICAL MODEL ARCHITECTURE UTILIZING EXISTING STATE OF ART MODEL ALGORITHMS



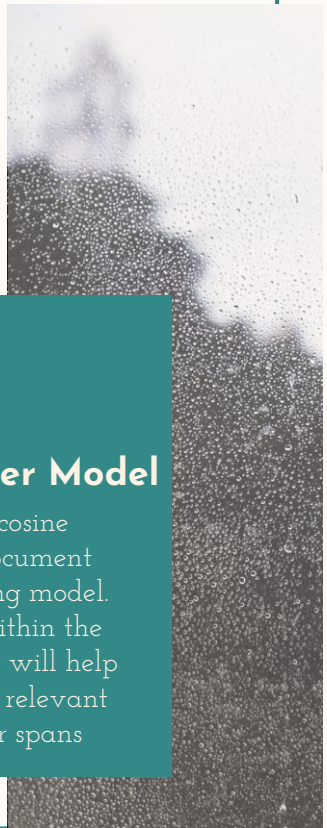
Masters in Data Science,
Liverpool John Moores University

Prashant Chadha
Student ID: 927706
Batch: Cohort 5, November 2020

Table of Contents

Introduction

- Corona Virus, first found in a patient in Wuhan-China at the end of 2019 and declared pandemic by WHO in January, has since spread in more than 180 countries across the globe and resulted in both, the loss of life and economy within these countries.
- Scientists across the globe have been studying the virus characteristics in order to suggest ways to reduce the spread, aid in development of therapeutics and vaccines.
- The Kaggle challenge is an initiative to encourage data scientist community to develop an algorithm that can aid scientists by learning from extensive literature and then giving relevant answers to the task query.
- The model designed should be capable of learning from exhaustive number of detailed documents and provide relevant answer span with respect to task query.
- Therefore, the decision to design a hierarchical model to answer the task query.

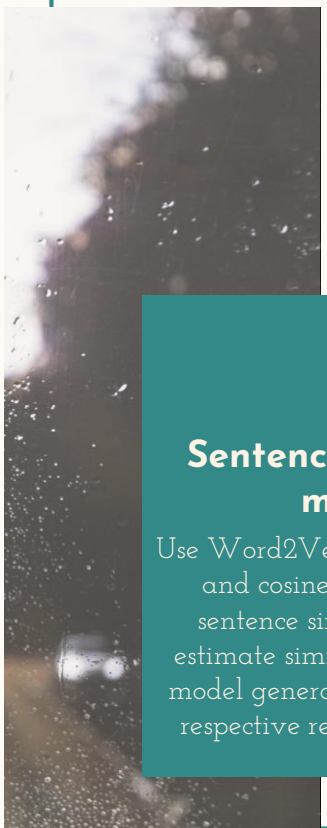


Document Ranker Model

Use TD-IDF and cosine similarity based document shortlisting or ranking model. This model layer within the proposed architecture will help in focussing on most relevant content for answer spans

Answer Span and Summarizer models

Use attention based pre-trained BERT model for answer spans generation and BERTSUM summarizer to summarize the spans



Sentence similarity model

Use Word2Vev word embeddings and cosine similarity based sentence similarity model to estimate similarity between the model generated summaries and respective reference summaries

Aim and Objectives

The main aim of this research is to develop a hierarchical network architecture capable of answering the Kaggle task query based on learning of content in extensive dataset.

The research objectives are formulated based on the aim of this study which is as follows:

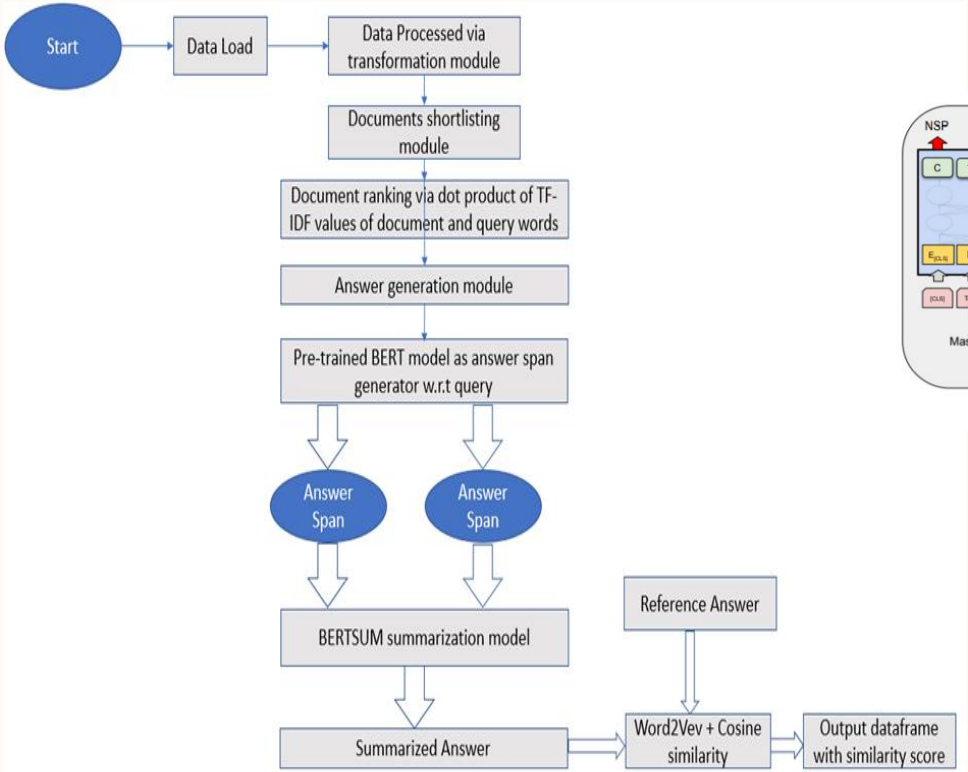
- To develop robust and efficient question and answer model capable of learning content from exhaustive list of research documents and answer the task questions.
- To suggest the best performing hierarchical model network with respect to answers generated by model network options being considered.
- To develop evaluation metric strategy most suited to judge the performance of model.
- To identify appropriate granularity level to be considered for input sequence .

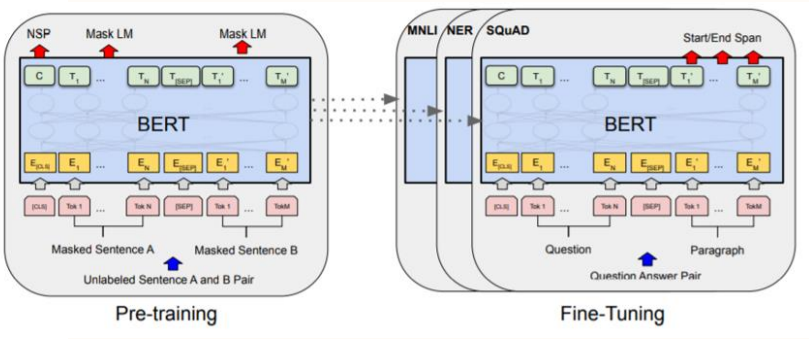
Word embeddings and Transformations

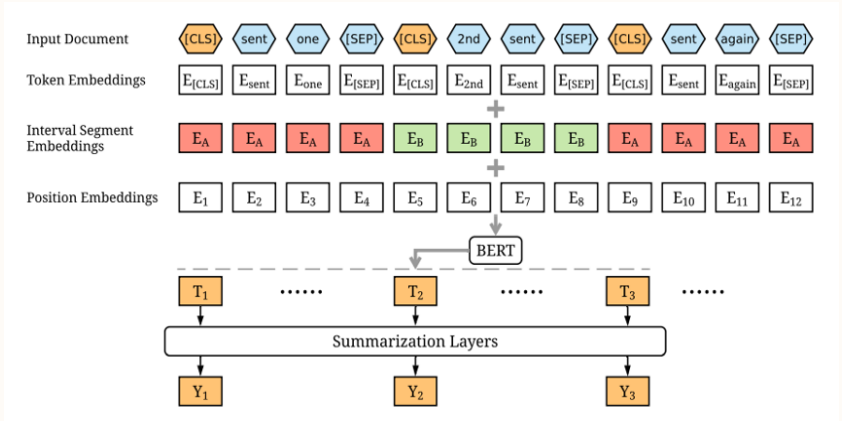
- The literature review was carried out with a purpose to first understand the word embeddings concept that plays big part in training models for NLP tasks.
- Word2Vec, consisting of CBOW and Skip-gram, outperformed existing models not only in terms of getting better evaluation metrics scores but were also faster to train and required less machine resources.
- The "Information retrieval" stage was developed using algorithms such as TF-IDF and Bag of Words (BoW) in some of the initial research work but the latest implementations have utilized search engines such as Anserini or state of the art and versatile BERT model.

Answer Span and Summary Models

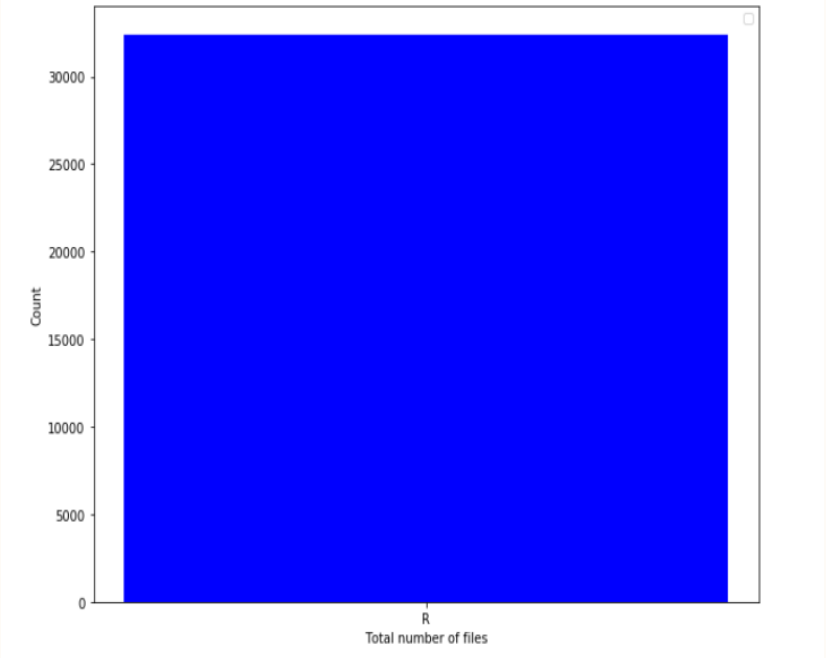
- CNN/RNN based models for answer span generation have showcased some of the best implementations of 'Attention' principle, achieving state of the art performances .
- Some of the limitations with CNN/RNN based models include model complexity and inconsistent performance in terms of learning semantic relationship within words separated over long distances.
- A novel model architecture that utilized stacked multiple feed-forward based layers in both the encoder and decoder to implement bi-directional 'Attention' inspired creation of BERT model.
- Most of the recent question and answer research work has implemented BERT in the "Information retrieval" and "Reader" stage and achieved benchmark setting results.
- BERTSUM combines BERT and summarization layers and the model has set benchmarks results in extractive summarization task.. It included novel sentence encoding techniques such as Internal segment encoding and performance optimization technique such as Trigram blocking.

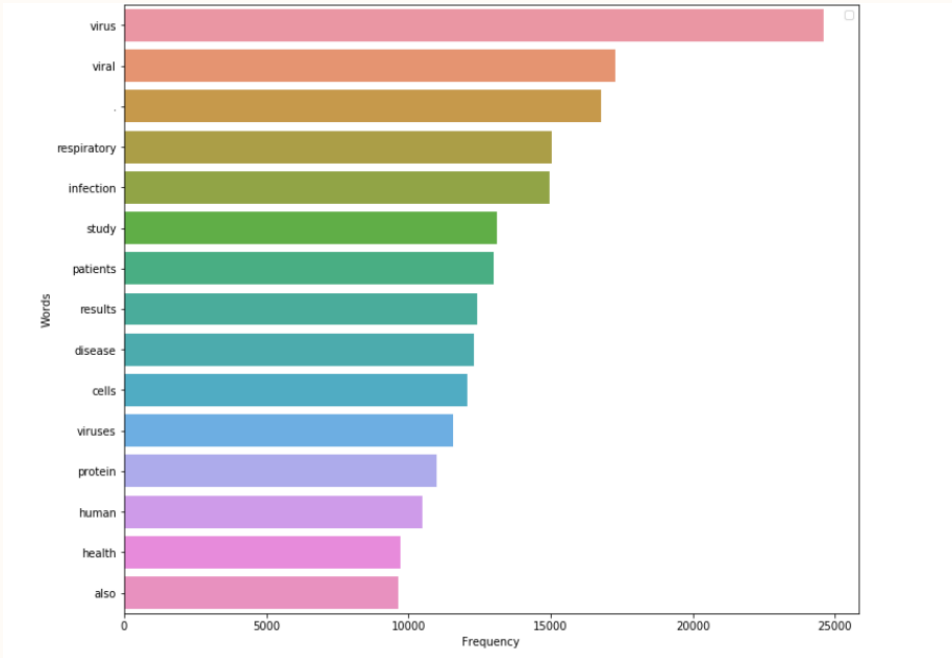




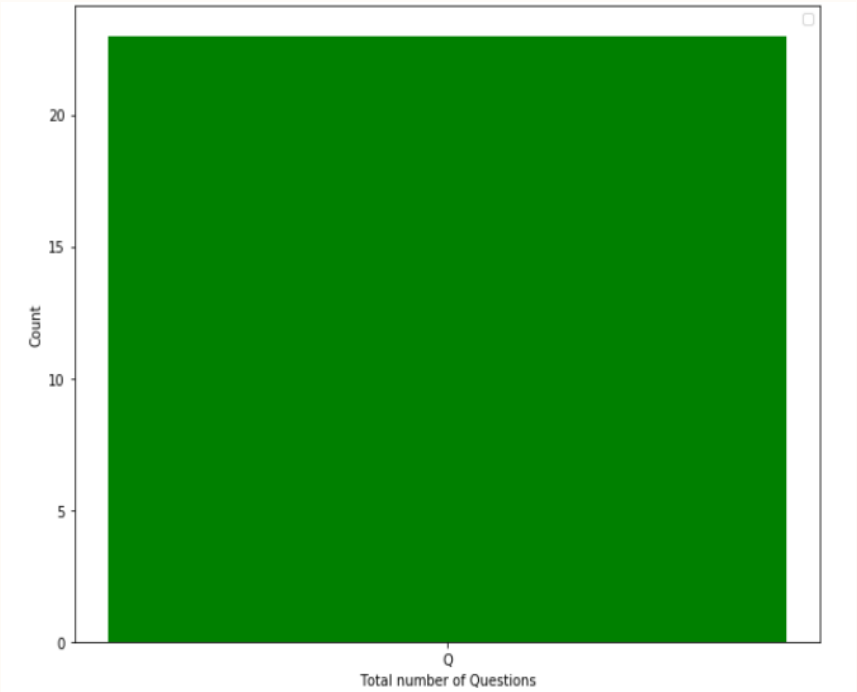


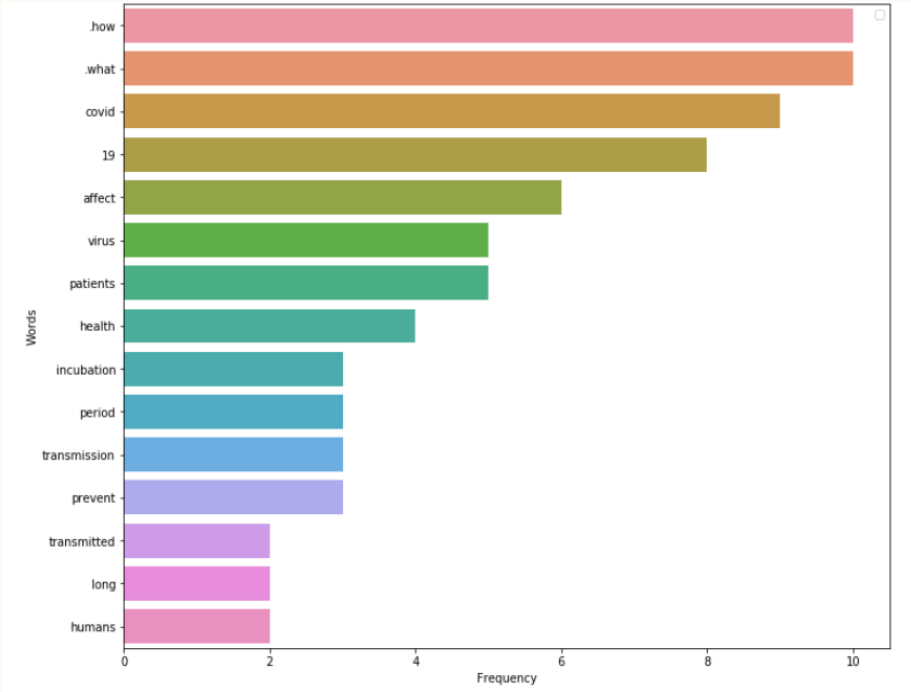
Proposed Model Architecture



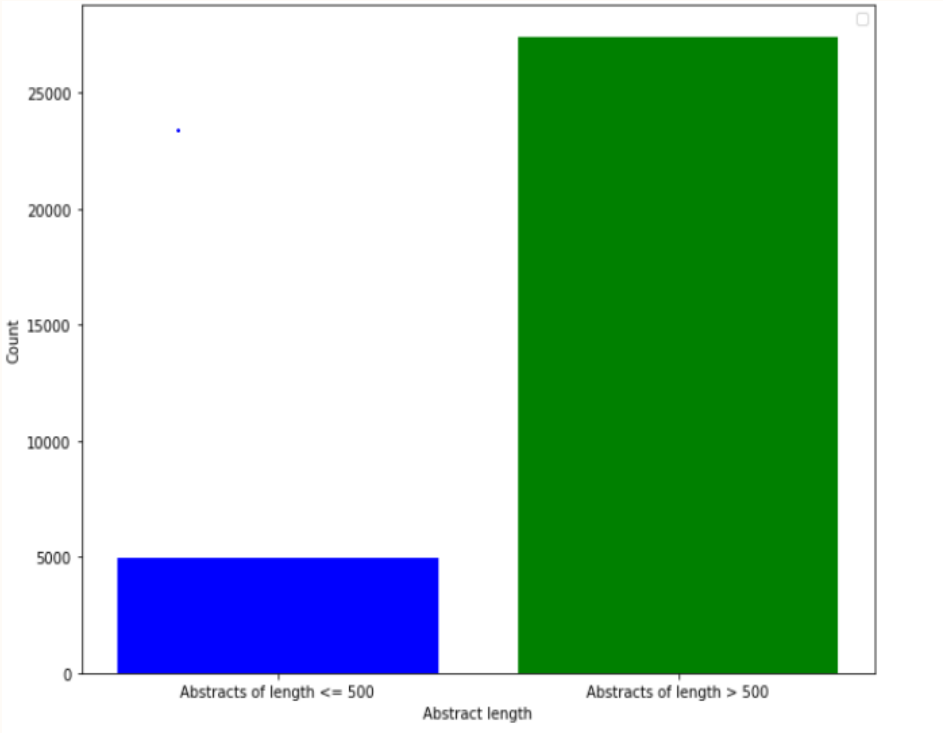


- The total number of files (abstracts) considered in the final input dataset is 34283 as shown in the left plot.
- The Words-Frequency plot shows that some of the most commonly mentioned words in the model answer span based summaries include 'infection', 'transmission', 'patients', 'health', 'incubation', 'period', 'risk', 'heart', 'covid-19'.



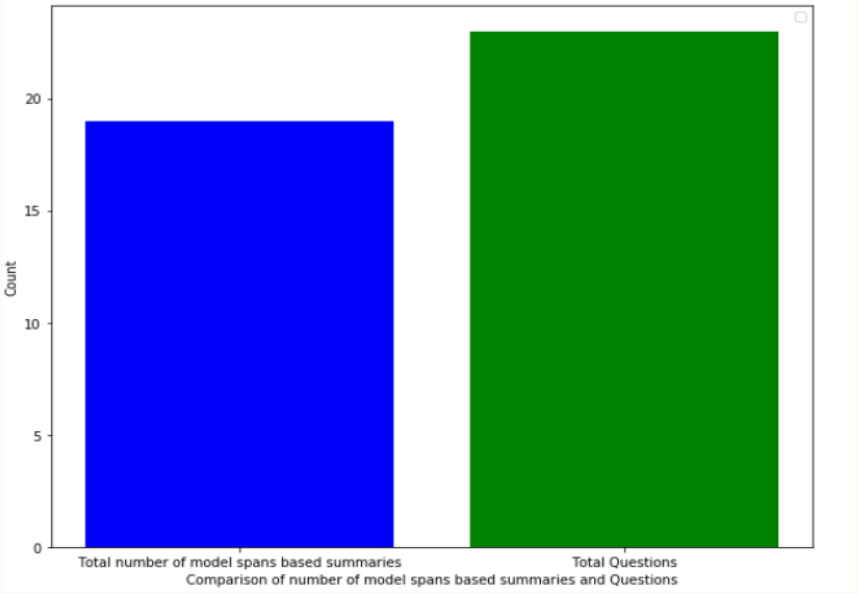


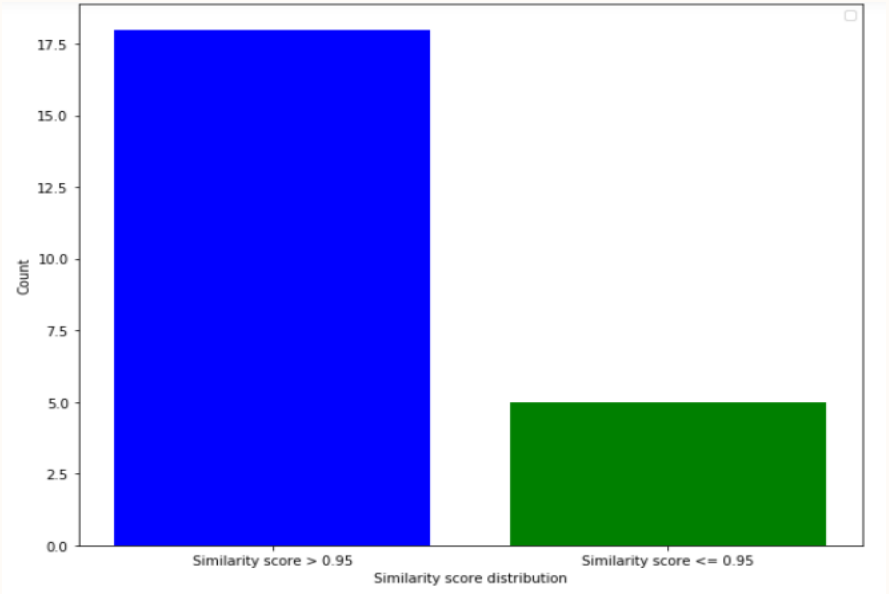
- The total number of questions considered in the final input dataset is 23 as shown in the left plot.
- The Words-Frequency plot shows that some of the most commonly mentioned words in the model answer span based summaries include 'how', 'what', 'covid-19', 'affect', 'virus', 'patients', 'health', 'transmission', 'incubation', 'patients', 'transmission'.



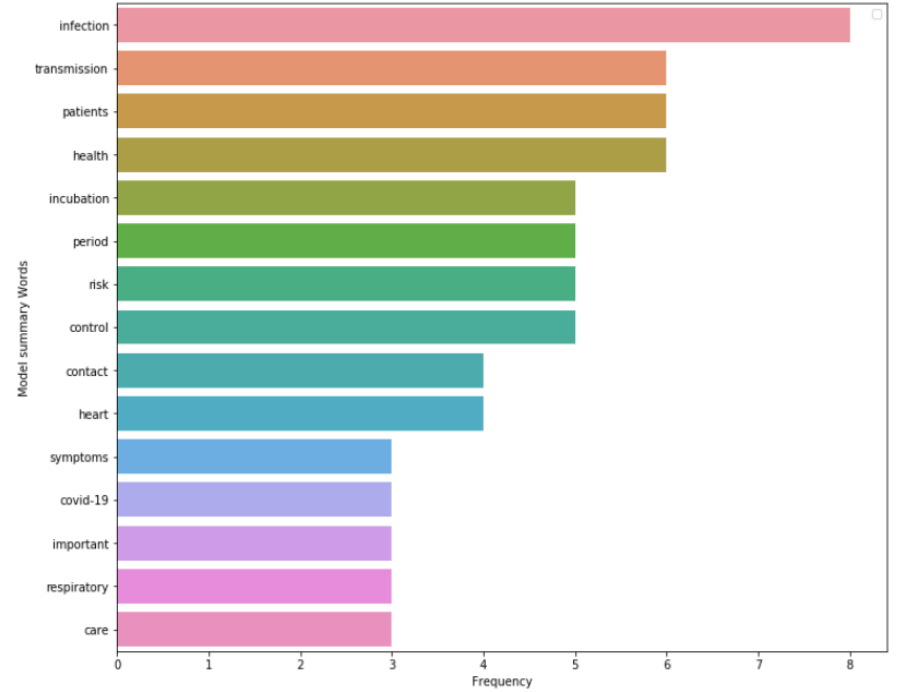
Column	Description
Sha	File id extracted from “paper id” dictionary within respective files
Title	Title of the respective files as mentioned in the 'metadata' file
Abstract	The ‘abstract’ mentioned within the 'metadata' file against the respective file records

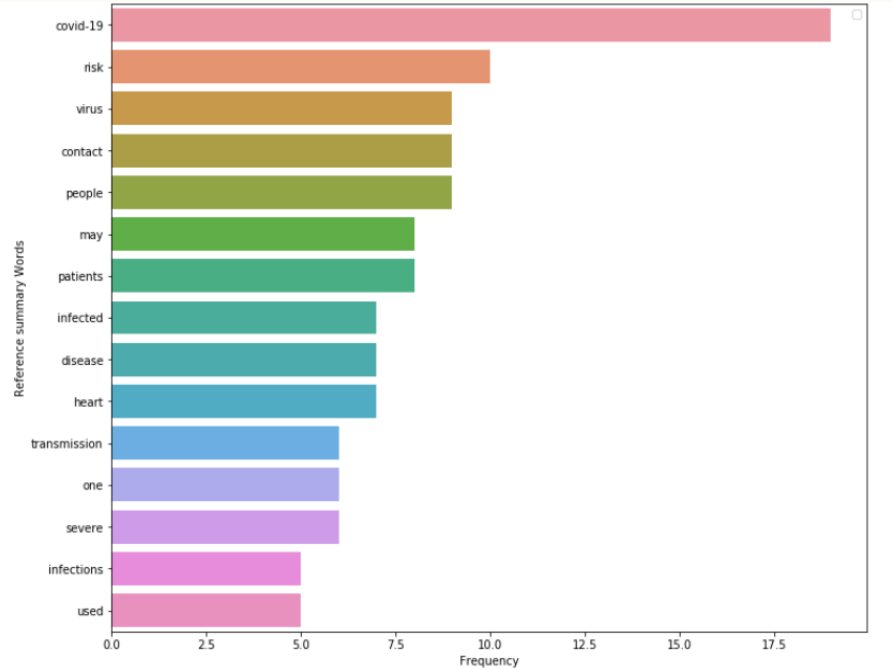
Column	Description
Task	The task mentioned in the Kaggle competition
Questions	The questions with respective to the task considered
Reference Answers	The reference answers with respect to the questions collected from cited websites



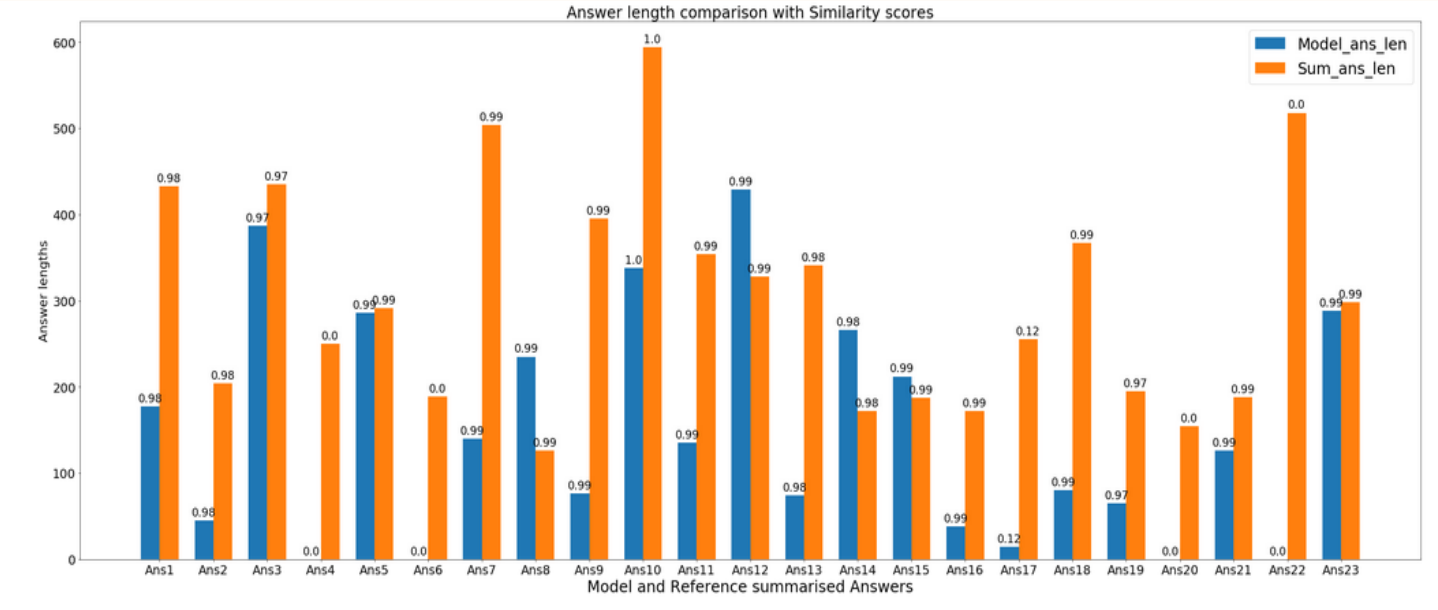


- The proposed model generated answers for 19 questions out of the 23 that were part of the dataset
- The number of questions for which similarity score between model span-based summaries and reference answer-based summaries achieved was above 0.95 was 18.

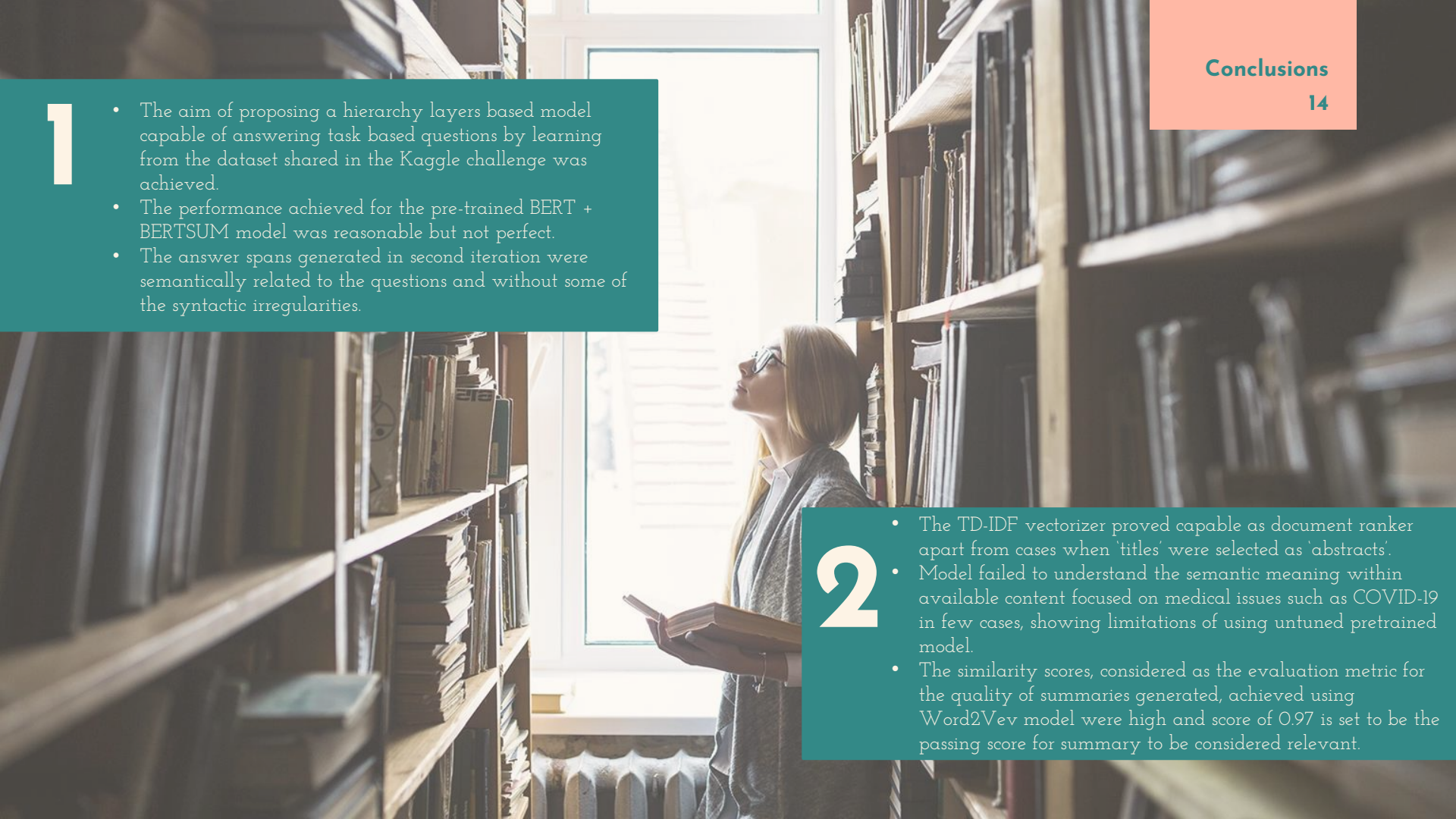




- The first plot shows that some of the most mentioned words in the model answer span-based summaries include 'infection', 'transmission', 'patients', 'health', 'incubation', 'period', 'risk', 'heart', 'covid-19'.
- The second shows that some of the most commonly mentioned words in the reference answer based summaries include 'covid-19', 'risk', 'virus', 'infected', 'transmission', 'patients', 'heart', 'contact', 'people'.



- The Model answer span-Reference answer based summaries comparison plot shows that size difference between the two answers does not necessarily impact the similarity scores till both answers are big enough to capture important information, making them similar semantically.
- The similarity scores are '0' for all the cases where answer span based summaries are not generated and the score is low for an instance (Ans17) when the answer span based summary is too small and does not have enough information to match the reference answer summary.



Conclusions

1

- The aim of proposing a hierarchy layers based model capable of answering task based questions by learning from the dataset shared in the Kaggle challenge was achieved.
- The performance achieved for the pre-trained BERT + BERTSUM model was reasonable but not perfect.
- The answer spans generated in second iteration were semantically related to the questions and without some of the syntactic irregularities.

2

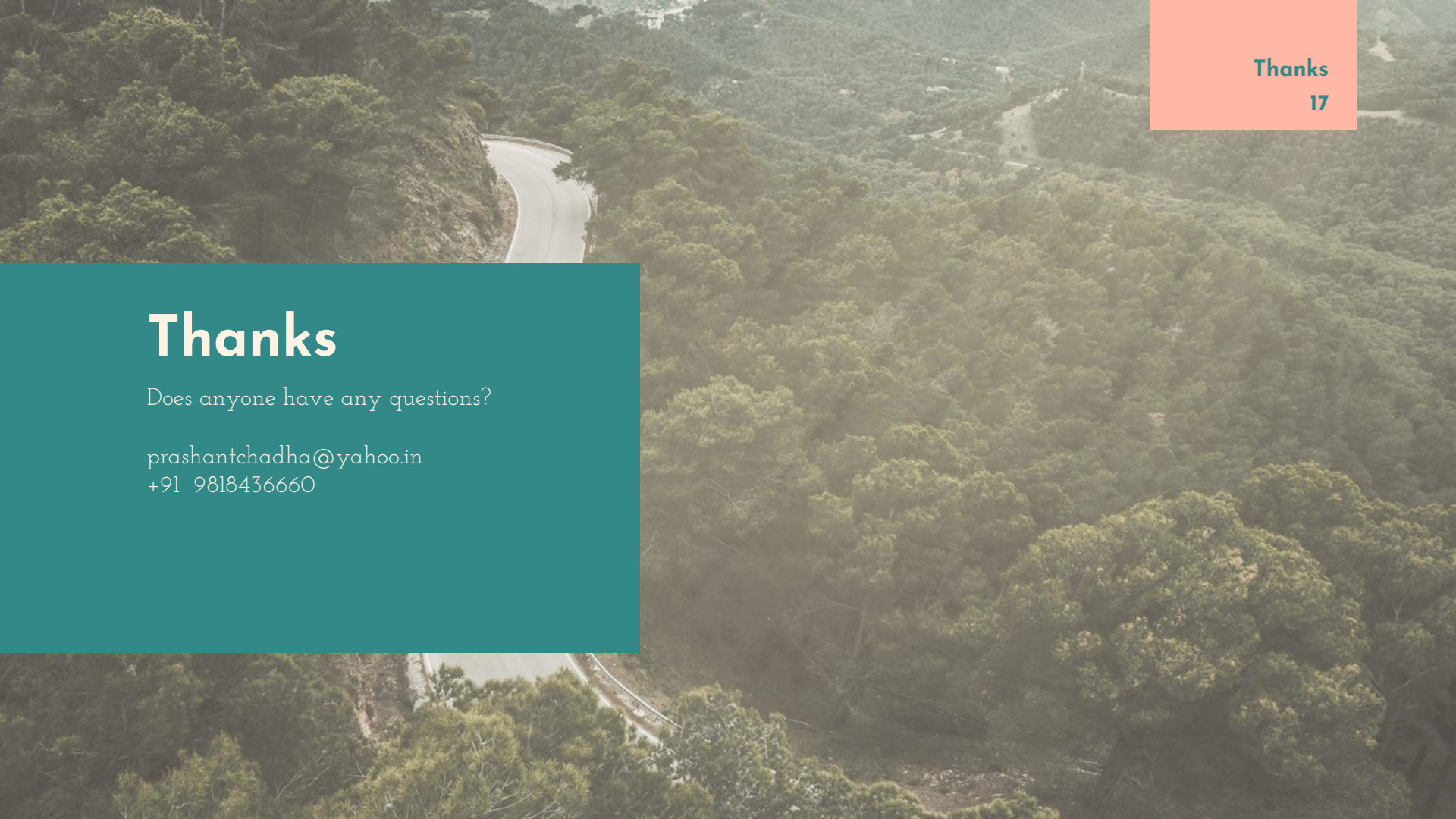
- The TD-IDF vectorizer proved capable as document ranker apart from cases when 'titles' were selected as 'abstracts'.
- Model failed to understand the semantic meaning within available content focused on medical issues such as COVID-19 in few cases, showing limitations of using untuned pretrained model.
- The similarity scores, considered as the evaluation metric for the quality of summaries generated, achieved using Word2Vev model were high and score of 0.97 is set to be the passing score for summary to be considered relevant.

Contribution to Research

- Proposed model is an effective tool for research community to extract relevant information from large dataset.
- Use of TD-IDF and cosine similarity as a document ranker model layer highlights the versatility of TD-IDF algorithm.
- Use of BERTSUM summarization to extract relevant information from pretrained, but untuned, BERT model generated answer spans instead of considering just highest confidence score based single answer span proved to be effective strategy of avoiding loss of important information and a good use case for BERTSUM based summarization.
- The optimisation code used for BERTSUM was a useful update to the model that can be used in other such tasks.
- Use of Word2Vec and cosine similarity model as an evaluation metric for unsupervised question-answer model performance proved to be an effective strategy and good use case for Word2Vec model.

Future Works

- Fine-tune the BERT answer generation model with dataset consisting of more questions and respective answers. The use of the proposed model architecture on other datasets focussed on question-answer task.
- Evaluate the performance of the proposed model architecture considering top-15 documents instead of top 10 that were considered in this project
- Use of the research body in cases of missing abstracts for both the document ranker and answer span models.
- The implementation of different models such as Anserini search engine as document ranker model and compare the performance on the same dataset.
- The implementation of BART as summarization model and compare the performance.



Thanks

Does anyone have any questions?

prashantchadha@yahoo.in
+91 9818436660