AUTOMATIC TEXT SUMMARIZATION OF COVID-19 RESEARCH ARTICLES USING SPARSE ATTENTION BASED TRANSFORMER - BIGBIRD

NISHA DEVENDRA KUMAR
Student ID: 944554

Thesis Report

JUNE 2021

DEDICATION

This research paper is dedicated to all the coronavirus warriors who have fought the infection bravely with their friends and families.

I also dedicate this to my thesis supervisor and mentors, who have been a constant source of support in completing my research.

This research would not have possible without my family support. My sincere thanks to all of you

ACKNOWLEDGMENTS

My sincere thanks to UPGRAD and Liverpool John Moors University for giving this opportunity to enhance my capabilities in research and analyze coronavirus dataset and conclude my research in systematic approach.

I would like to thank my thesis supervisor, Mr. Saheb Chabbra, for his valuable and timely review and feedback. I would also like to thank DR. Manoj Jayabalan from Liverpool John Moores University for his continued support and guidance through weekly and one-on-one sessions throughout the duration of the program.

I also place on record, my sense of gratitude to one and all who, directly or indirectly, have lent their helping hand in this venture.

ABSTRACT

COVID-19 research dataset has a pivotal role in the current pandemic as clinical researchers are dependent on this increasingly growing corpus to refer and draw insights. Summarization of lengthy research documents such as COVID-19 is still a challenge in NLP domain as the standard Transformer architectures have quadratic dependency on the sequence length due to their full attention mechanism. The aim of this research work is to explore and build a method to generate context rich abstract summaries of research articles from CORD-19 data set by leveraging the potential of recently introduced sparse attention transformer i.e., BIGBIRD. BIGBIRD has the potential to handle input tokens up to 4096 which is almost 8 times higher than that of the existing transformers. The intent of the study is to explore the improvisation on CORD-19 data summarization task with BIGBIRD. In the process, the research aims to compare performance of BIGBIRD standalone architecture with that of BIGBIRD when ensembled with an extractive summarization layer. The intent here is to build a summarization strategy to improve the abstractive summary and, in the process, explore the possibility of leveraging an already existing extractive and abstractive summarization approach on BIGBIRD model for further improvisation. The intent here is to design an approach to innovate and achieve good performance metrics on COVID dataset which will eventually benefit the research community in the coronavirus crisis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BART……. Denoising Sequence-to-Sequence Pre-training for Natural Language Generation

BERT……. Bidirectional Encoder Representations from Transformers

BLEU……. Bilingual Evaluation Understudy

CORD-19…. Chronic Obstructive Respiratory Disease

COVID-19… Corona Virus Disease

ELECTRA…. Efficiently Learning an Encoder that Classifies Token Replacements Accurately

GPT……... Generative Pre-trained Transformer

LSTM……. Long Short-Term Memory

MLM……...Masked Language Modelling

MLPs……...Multilayer Perceptron

NLP……... Natural Language Processing

NLTK ……... Natural Language Tool Kit

PEGASUS…. Pre-training with Extracted Gap-sentences for Abstractive Summarization

RNNs……...Recurring Neural Network

RoBERTa….... Robustly Optimized BERT Pre-training Approach

ROUGE……. Recall-Oriented Understudy for Gisting Evaluation

UniLM…… Unified Language Modelling

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Automatic summarization (Ibrahim Altmami and El Bachir Menai, 2020) of lengthy documents using Machine Learning e.g., Legal data sets, Science / Medical Journals etc., has gained popularity since few years as the concept of chunking and filtering only the important information without losing the context from exhaustive factual documents not only saves time but also contributes to an improved understandability for target audience. Figure 1.1 gives an overview of summarization concept.

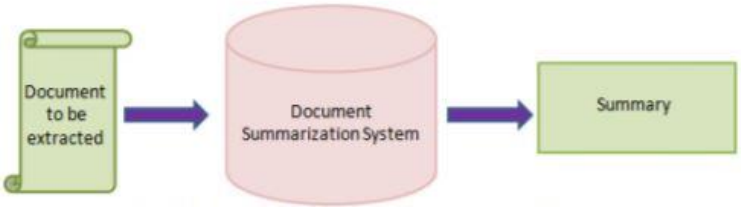

Figure 1.1 Concept of Summarization

Extractive Summarization (Moratanch and Chitrakala, 2017) and Abstractive Summarization (Moratanch and Chitrakala, 2016) are two existing approaches that has been implemented using Machine Learning and has found to be doing as extraction and abstraction concepts are subjective to domain at times, this always creates a scope of improvisation. Extractive summary retains the most important i.e., those sentences or phrases from the original text that are conveying the context of the original texts. In Abstract summarization new sentences are generated from the original text and at the same time preserving the context.

In recent past, with increasing use of various Deep Neural Network (Zhou et al., 2018) based architectures and Transformers (Vaswani et al., 2002) complex NLP problems can be solved. Transformers have revolutionized the way NLP problems can be approached. Text Summarization specific to various domain have gained popularity since a past few years. However, the nature of datasets varies with the nature of the domain. Getting the grasp of the entire content without having to go through the exhaustive content always fascinates the audience interested in a particular.

Text Summarization is an important feature in medical domain where there is a need for summarization long scientific journals, articles, and related texts for research purpose. In the wake of current pandemic leading Research Groups in collaboration with White House have prepared CORD-19 i.e., COVID-19 Open Research Dataset (Lu Wang et al., 2020) and made public for the research communities across the world. This data set contains over 4000,000 researched scholarly pandemic articles including over 150000 full length articles on coronaviruses including COVID-19 and SARS-CoV-2. This content can be leveraged by the Research communities across the world to synthesize new insights to battle the ongoing pandemic and help in insulating from the pandemics that may come up in the future. By leveraging Natural Language Processing and other AI techniques on this data, a lot of useful work can be done one of the ways to achieve this is by performing efficient text summarization. An efficient text summarization here will not only save time for the researchers in going through the exhaustive content but also leverage them to get hold of the context and help them draw new insights of each article which otherwise can be lost in exhaustive reading of the original content.

1.2 Problem Statement

The significance of summarization in covid research is elaborated in section 1.1. The existing summarization techniques (Tan et al., 2020) proposed for CORD-19 datasets usually involves transformers like GPT-2 (Liu et al., 2021), BART (Lewis et al., 2019), BERT (Devlin et al., 2019). These transformers have one major limitation that they cannot process input token sequences greater than 1024. This becomes a bottleneck in summarization of lengthy research documents such as that of CORD-19.

These strategies although being moderately efficient, do not rule out the possibility of missing out important contexts from the research articles as the context is usually spread across the entire document which can be lost due to conditioning strategy in existing summarization approaches. With this backdrop there has always been a need for a strategy which can accommodate more input token sequences to learn more context from the input documents and improve the summarization process.

With the recently introduced sparse based attention transformer- BIGBIRD (Zaheer et al., 2020), the possibility of improvising summarization tasks in research world has increased ss the input size increases by 8 times. BIGBIRD has capacity to process input

tokens up to 4096 which can be leveraged for research articles such as CORD-19 where 80% of the corpus have article length up to 5000 words. In this research the potential of BIGBIRD in automatic summarization of CORD-19 dataset is explored.

The possibility of improvisation will be further explored by leveraging the existing extractive and abstractive combination approach in summarization (Vladislav and Denis, 2020) using BIGBIRD. This will help to assess if this strategy is beneficial with BIGBIRD.

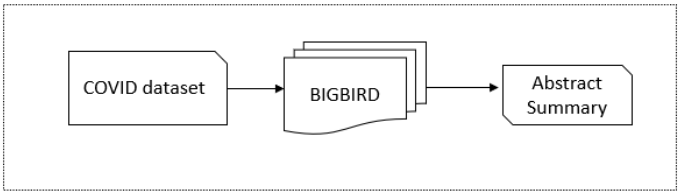

Figure 1.2 Summarization pipeline with BIGBIRD

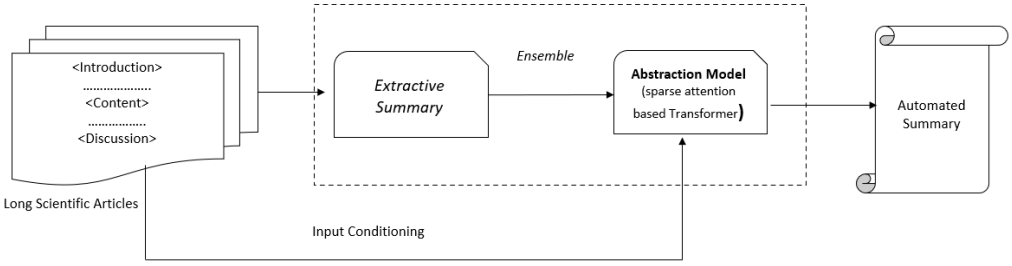

Figure 1.3 Summarization pipeline - BIGBIRD with Extraction Layer

Figure 1.2 and 1.3 gives an outline of the proposed architectures to address the problem statements.

1.3 Aim and Objectives

The main aim of this research is to propose a generate improved abstract summaries of COVID-19 data set (Scientific document) by leveraging BIGBIRD. In the wake of recent pandemic, the need for robust research from large volume of data within a short time span has become inevitable. Extractive and Abstractive summarization approaches through recent advances in Open AI and NLP, can leverage processing and retrieval of comprehensive information from healthcare domain within a short time frame.

This research is intended to help the medical research community to keep up with the rapidly growing coronavirus literature and draw insights in a short turnaround time to

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

- As BIGBIRD has capacity to process 4096 input tokens, train and evaluate the performance of BIGBIRD on CORD-19 summarization task
- Evaluate the effect of adding of extractive summarization layer before the final abstraction task in BIGBIRD
- Compare between BIGBIRD (standalone) and BIGBIRD with extraction ensemble to find the best approach for CORD-19 summarization task.
- Analyze the performance of BIGBIRD on three ranges of input document size i.e., short, medium and long
- Compare BIGBIRD with BART on input size that levels the playing field for comparison for COVID dataset.

1.4 Research Questions

The following research questions are suggested for each of the research objective as highlighted as follows.

- Can BIGBIRD address the gaps in the existing summarization strategies of long scientific documents such as COVID-19?
- Can adding extractive summarization layer before the BIGBIRD's abstractive summarization model enhance the performance of the summarization pipeline?
- Is the performance of BIGBIRD satisfactory for short, medium and long sized documents of CORD-19 dataset?
- Given an input size that levels the playing field for BART and BIGBIRD, is the performance of BIGBIRD better?

1.5 Scope of Study

The scope of the study is as follows: The study will explore the capabilities of sparse attention-based Transformer BIGBIRD (Zaheer et al., 2020) for automatic summarization of COVID -19 research articles. The research will explore the possibility of improvising text summarization of long scientific documents of COVID dataset by leveraging both extractive and abstractive summarization strategies together in BIGBIRD. This research will also find the suitability of BIGBIRD for summarization task for small to large sized documents. The research also aims to make comparisons to

discover if the strategy is performing better than the standard transformers with the conventional strategy used for long documents.

The below items are out of scope of this research.

Training and evaluation of research articles with lengths exceeding 5000 wordcount are not in scope of this research. This is because of the below two reasons: As the evaluation is centered around BIGBIRD and it accepts input token sequences up to 4096, evaluating articles with 5000 wordcount covers 75% of corpus. Also, the resources and time needed to train BIGBIRD is high so excluding excessively long documents makes the sample size manageable from training perspective. CORD-19 Summarization comparisons from other transformer models is not included in this research, the focus is only on comparison between performances of BIGBIRD under different ensembles. This is because of the below two reasons: Limitations of resources and time to train and evaluate other models for CORD-19 as BIGBIRD takes a lot of time to train.

1.6 Significance of the Study

The study aims to address the existing summarization gaps of CORD-19 articles which can greatly help the medical community in the coronavirus pandemic. This research is will help the clinical researchers in coping up with the rapidly growing coronavirus literature and draw insights in a short turnaround time to fight the pandemic.

1.7 Structure of the Study

The structure of the thesis is strategized as follows. Chapter 1 introduces the problem domain of the current study. It provides needed backdrop on existing trends in automatic text Summarization in NLP. It further throws the spotlight on the need of automatic summarization specific to healthcare domain and the significance COVID dataset summarization. The focus area of the study along with aim and objectives of the research that will address the problem statement is highlighted in section 1.2 and section 1.3. The research questions that the study intends to address is elaborated in section 1.4. Section 1.5 presents the in-scope and out of scope items followed by significance of study.

Chapter 2 presents the theoretical backdrop of the problem statement by systematically reviewing the latest summarization work done across the chapter. Section 2.2

walkthroughs the commonly used extractive and abstractive summary approaches followed by rare techniques such as graph (Moawad and Aref, 2012) and reinforcement learning (K et al., 2019) based in section 2.3 and 2.4. Section 2.5 introduces the revolutionary work in Transformers. Section 2.6 introduces the concept of sparse attention-based transformers. Section 2.7 puts the spotlight on ensemble strategies with both extraction and abstraction layers. Section 2.8 explores the evaluation metrics of summarization tasks. The summary of the reviews is discussed and concluded in sections 2.9 and 2.10.

Chapter 3 discusses the research design and research framework. Section 3.2 describes the data sources and data processing strategy. Section 3.3 walkthroughs the proposed summarization pipelines to be trained. Section 3.4 introduces to the performance evaluation metrics in the proposed framework. Comparison strategy between the trained models is discussed in section 3.5.

Chapter 4 elaborates the technical analysis done as part of research work. Sections 4.2 to section 4.5 walkthroughs the data collection strategy such as keyword retrievals from various sources that constitutes covid dataset and assess the challenges in the summarization task. Sections 4.5 and 4.6 walkthroughs the data analytics and processing steps conducted as part of study. Section 4.5 walkthroughs the model building phases of the proposed pipelines using BIGBIRD. Sections 4.6 and 4.7 explains the model evaluations and comparison done on the trained ensembles. Section 4.8 proposes the final model selection on the basis of evaluation metrics. Section 4.9 concludes with the summary.

Chapter 5 gives an in-depth analysis of evaluation results between the proposed models. Section 5.2 elaborates the model evaluation strategy and section 5.3 walkthroughs model evaluation outcomes. Section 5.4 compares BIGBIRD with an existing transformer BART. Section 5.5 churns out the meaningful insights followed by summary in section 5.6.

Chapter 6 concludes and draws recommendations from the research work. Section 6.2 discusses the outcome of the research and beneficiaries of the research work. Section 6.3 highlights the contribution to knowledge with this research. The chapter concludes with future recommendations in section 6.4.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Text summarization attempts to condense long texts without losing the context and at the same time preserving the important information. Initial works in areas of text summarization focused on extractive techniques aiming to retain the most important sentences from the documents.

Abstractive approach on the other hand involves summarizing the whole context in a more condensed structure drastically reducing the length of the content. Obviously abstractive summarization is a challenging task as generating creating abstractive demands command on the domain and as well as the natural language which is a tedious task for the machine.

Scientific document summarization is a special case of summarization as characteristics of scientific papers – length, writing styles, scientific terms and discourse structure demands an exclusive model consideration to maintain the context and at the same time retaining the accuracy of the topic. Researchers have engineered different approaches to address the challenges in Scientific document Summarization.

The metamorphosis of Long Short-Term Memory networks to attention mechanism combined with sequence-to-sequence framework was a pivotal in improvising language modelling tasks. Introduction of transformer architecture coupled with novel self-attention mechanism was a significant leap in language modeling task.

2.2 Extractive Summary Approaches

In a recent work extractive summarization as Text Matching (Zhong et al., 2020) is proposed. This is a novel summary framework which scores and extracts sentences one by one to form a summary, a strategy to formulate extractive summarization in form of semantic text matching problem. In this Siamese network structure and basic BERT(Devlin et al., 2019) have been combined to form Siamese-BERT architecture to compute the similarity between the source document and the candidate summary. In one of the models, Extractive summarization for lengthy structured content is attained by leveraging both local and global context from the entire document (Xiao and Carenini, 2020). This approach is inspired by natural topic-oriented structure of long documents which are created using human intelligence, where the binary conclusion of whether the

sentence should be part of the summary is dependent on the sentence itself, the entire document and the current topic. The representation of document is cascading of the last 'n' hidden states of the forward and backward RNNs, while the representation of topic segment is done by leveraging LSTM-Minus method.

REFRESH (Narayan et al., 2018) is a trained extractive summarization model for a globally optimized ROUGE metric and uses reinforcement learning.

NEUSUM (Zhou et al., 2018) is an extractive summarization system that has the capability of scoring and selecting sentences

2.3 Graph Based Summarization

There have been Graph Based Text summarization models that have advantages and shortcomings. Word based graph methodology (Le and Le, 2013) was good at maintaining syntactic constraints but produced grammatically incorrect sentences and didn't consider meaning of word or phrases which led to loss of context in the generated summary.

Improvisation in Graph Based Methodology was seen in semantic graph reduction model (Moawad and Aref, 2012) which initially creates rich semantic graph followed by semantic graph reduction that includes the domain ontology class instances which helps to capture the meaning of sentences and even paragraphs which finally yields a better result in text generation step with less data loss. Figure 2.1 outlines the semantic graph flow.

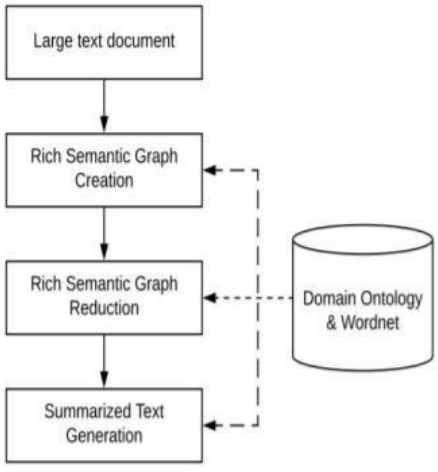

Figure 2.1 Workflow of semantic graph 1 model (Moawad and Aref, 2012)

In graph-based hybrid strategy to abstractive summarization applying Markov's Clustering (Sahoo et al., 2018), it was seen that it took into account sentence connections leading to sentence clustering followed by sentence positioning using sentence ranks and eventually sentence compression to produce effective summarization.

2.4 Reinforcement Learning for Text Summarization:

- Use of Reinforcement Learning has been explored for NLP tasks and based on this a text summarizer task combining Neural Networks and Reinforcement Learning (K et al., 2019) is proposed where Reinforcement Learning Algorithm is used to introduce feedback into the text summarization workflow which uses Encoder-Decoder module of Neural Networks. The comparison of the results with and without the feedback workflow added as shown in Figure 2.2, indicates that use of RL facilitates the system to make corrections via feedback and produce a more relevant summary with a better Rouge Score.

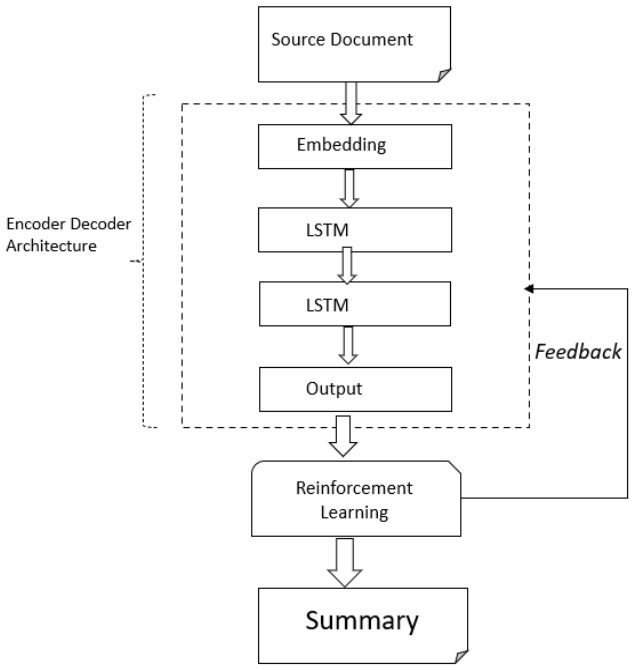

Figure 2.2 Reinforcement Learning for Text Summarization

- A hierarchical reinforcement learning model (Koupaee, 2018) approach that divides the task into a set of sub-tasks and optimization of individual sub tasks are done to optimize the abstractive summarization of texts. This approach experimented on WikiHow and CNN/Daily Mail data sets have given higher ROUGE scores.
- In one of the recent work in abstractive summarization the concept of discriminative adversarial search (Scialom et al., 2020) is proposed which uses Beam search which is de-facto algorithm used to decode generated sequences of text. Beam search has led to performance improvements of State of Art models Q&A Generation, Text Summarization and Neural Machine Translation

2.5 Transformers and Text Summarization

- The metamorphosis of Long Short-Term Memory networks to attention mechanism combined with sequence-to-sequence framework was a pivotal in improvising language modelling tasks. Introduction of transformer architecture coupled with novel self-attention mechanism was a significant leap in language modeling task. In Transformers architecture is Multi-head Attention Model i.e., self-attention is computed multiple times independently and in parallel and the outputs are concatenated followed by linear transformation.
- The key differentiator in Transformers is the application of a self-attention mechanism, which computes and evaluates the similarity scores for all pairs in an input sequence in parallel for individual tokens of the input sequence, completely bypassing the sequential dependency present in recurrent neural networks and thus outperforming previous sequential models.

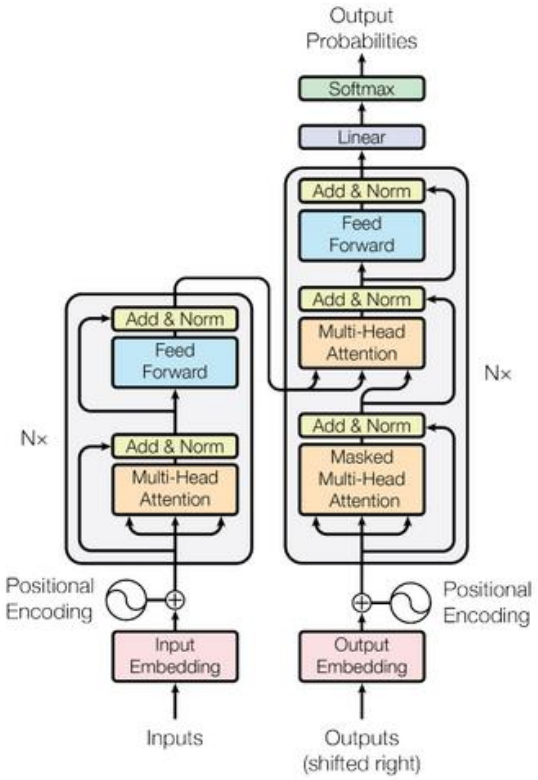

Figure 2.3 The Transformer – Model Architecture (Vaswani et al., 2002)

BERT (Devlin et al., 2019) implemented masked language modelling, which facilitated pre-training to learn interactions between left and right context words and enabling pre-trained deep bi-directional representations. Figure 2.3 shows that masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based on its context. However, since predictions are not made auto-regressively, effectiveness in generation tasks like abstract summarization effectiveness of BERT is reduced. BERTSUM (Liu, 2019) is a variant of BERT designed for Extractive Summarization which is achieved by modification of input sequence and embeddings of BERT. BERTSUM with Transformer is found to have achieved a very good performance on ROUGE metrics

T5 (Raffel et al., 2020) generalized the text-to-text framework to a variety of NLP tasks and showed the advantage of scaling up model size (to 11 billion parameters) and pre-training corpus, introducing C4, a massive text corpus derived from Common Crawl, which we also use in some of our models. T5 was pre-trained with randomly corrupted text spans of varying mask ratios and sizes of spans.
On a CNN/ Daily Mail Data set this has demonstrated a ROUGE-2-F score of 21.55.

Unidirectional language model such as GPT (Peters et al., 2018) can be used for text generation as tokens are predicted auto regressively as show in Figure 2.4, but due its limitation of conditioning of words on leftward context, it is incapable of learning bi-directional interactions.

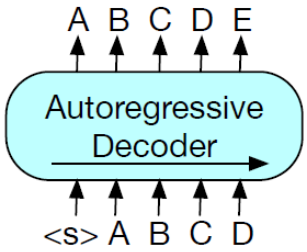

Figure 2.4 A schematic representation of GPT (Clark et al., 2020)

BART is a denoising sequence-to-sequence pre-training for Natural Language Generation, Translation, and Comprehension by (Mike Lewis et al., 2019). It performs pretraining of sequence-to -sequence models by denoising autoencoder as shown in Figure 2.5. Training of BART is done by corrupting text with an arbitrary noising function and learning a model is made to reconstruct the original text. BART is one of the best performing transformers as it generalizes BERT, GPT and many other most pre-training schemes.

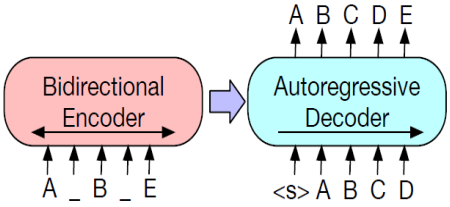

Figure 2.5 A schematic representation of BART (Lewis et al., 2019)

MASS (Song et al., 2019) proposed a model that involved masked sequence-to sequence generation to construe a sentence fragment from a given remaining part of the sentence that was randomly selected. UniLM (Dong et al., 2019) proposed a model that jointly trained on three types of modeling tasks: bidirectional (word-level mask followed by sentence prediction), sequence-to-sequence (word-level mask) prediction and unidirectional (both left to- right and right-to-left). Similar hybrid language model XLNet (Yang et al., 2019), is a generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all

permutations of the factorization order. In a recent work (Zhang et al., 2019) proposes pre-training with gap-Sentences for abstractive summarization. This model is popularly is known as PEGASUS model, in which the key sentences are removed or masked from an input source and are generated collectively as single output sequence from the remaining sentences as shown in Figure 2.6. This approach is similar to an extractive summary. In this model instead of continuous text spans it masks multiple whole sentences and chooses sentences based on importance as output. Its architecture is a standard Transformer encoder in which Both Gap Sentence Generation and Masked Language Model are applied simultaneously to achieve the effective abstractive summarization.

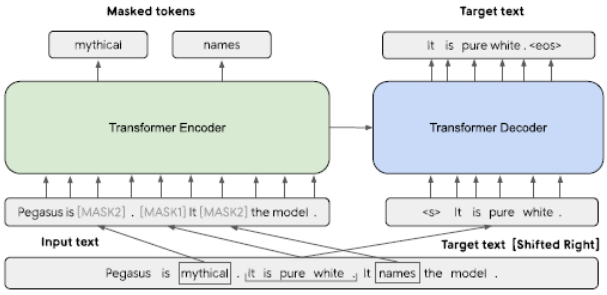

Figure 2.6 Pegasus Architecture (Zhang et al., 2019)

2.6 Longformers

Transformers models and their variants have found to revolutionize the way NLP problems such as summarization is perceived due to multi head self-attention mechanism, however they have a major drawback with specific to length of the source document. Transformer based Models fail to process documents where the length is long and this is majorly due to self-attention that scales quadratically with length of the sequence of the source input. The self-attention component in Transformer has time and memory complexity as O(n2) (where n is length of sequence input). Due to quadratic relationship of computational and memory requirements with input sequence length and existing hardware/ resource constraints transformers have limitation of input sequence length of 512 tokens, which limits its applicability for tasks that require longer contexts like Summarization. The Figure 2.7 shows graphical view of full attention mechanism.

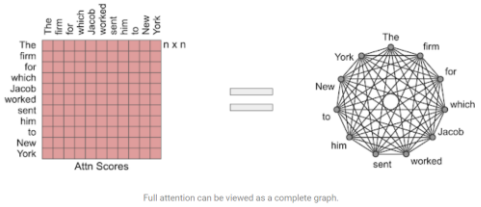

Figure 2.7 Full attention viewed as complete graph (ai.googleblog, n.d.)

Extended Transformer Construction (ETC) (Alberti et al., 2020) introduced a novel strategy for sparse attention, in which puts a limit on the computed pair of similarity scores based on structural information thereby reducing the quadratic dependency on input length to linear dependency and giving a superior performance for larger contexts. BIGBIRD (Zaheer et al., 2020) is a sparse attention mechanism (Figure 2.8) model is developed to address the quadratic complexity challenges of Transformers induced due to full self-attention mechanism. BIGBIRD utilizes MLM pretraining for base-sized models and for large sized models it is making use of summarization specific pretraining from Pegasus. The interesting feature here is that the sparse mechanism is implemented only at encoder size and this is primarily because the length of the sequence as output is quite small as compared to length of length of input sequence. The sparse mechanism at encoder level also helps to cover the input sequence entirely as the salient features in lengthy documents could be evenly distributed across the entire document.

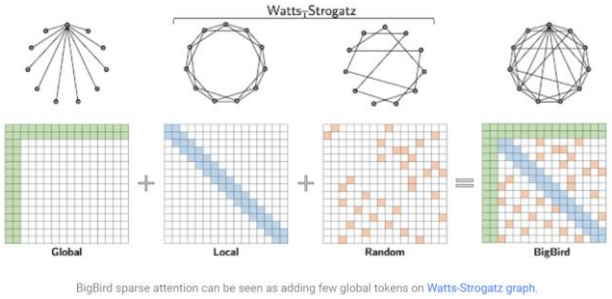

Figure 2.8 BIGBIRD Sparse Attention seen on graph (ai.googleblog, n.d.)

Longformers have been introduced to address the limitations of Transformers and at the same time retaining the benefits of attention mechanism. Longformers are pre-trained counterparts of Transformers with attention mechanism scaling linearly with length of the sequence of the source input which makes processing of longer documents feasible. In Longformers attention mechanism combines a local windowed attention with a task

motivated global attention which is a drop-in replacement for standard self-attention mechanism available in transformers.

Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) is variant of Longformer variant which supports generative sequence to sequence of lengthy documents. LED is based of BART's architecture with same number of layers with the difference that the position embedding is extended from 1K to 16K tokens to perform for longer input size. LED has been evaluated on long document summarization tasks such as scientific literature datasets and is found to outperform with a good Rouge Score.

REFORMER (Nikita Kitaev et al.,2020) proposes model to improve efficiency of Transformers by using reversible residual layers instead of standard layers and using attention with locality-sensitive hashing reducing the complexity to O(nlogn) where n is the length of input sequence.

2.7 Enhancing Abstraction using Extraction

In a recent work on neural document summarization using Transformer language models (Subramanian et al., 2019) author proposes combining extraction and abstraction strategies to come up with a more effective abstract summarization. The author has used encoder decoder architecture for Extractive summary where a sentence encoder is implemented using bi-directional LSTM and the decoder is implemented as autoregressive LSTM taking the sentence-level LSTM hidden state of the previously extracted sentence as input and predicting the next extracted sentence. For abstraction they have trained a Transfer Language Models with -220M parameters with 20 layers, 768 dimensional embeddings, 3072-dimensional position-wise MLPs and 12 attention heads. This model is trained on 4 components: Introduction, extracted sentences, abstract and rest of the paper. For creating abstracts of a long document, the trained model uses "Introduction" of the document as proxy to contain enough information for abstract along with the extracted content from extraction model. For a smaller document the introduction is the entire document. The combination has given a better rouge score.

In another similar work titled " Combination of abstractive and extractive approaches for summarization of long scientific texts" (Vladislav and Denis, 2020), the author proposes combining extractive and abstractive summary strategies with the usage of pre-trained Transformer Models. Extractive Model is trained as a classification to generating abstractive summary. They have experimented on 3 different architectures BERT, RoBERTa (Liu et al., 2019), and ELECTRA(Clark et al., 2020) for extractive

Summary. Word piece tokenizer is used where BERT (Devlin et al., 2019) and ELECTRA(Clark et al., 2020) is experimented. In the process they have added special tokens like [math], [graph], [table], [equation] extracted using regex. Rouge Score is calculated using extracted summary and Ground Truth Abstract. As an outcome of the experiment BERT exhibited the highest performance Metrics as compared to other models. Extractive Summarization output was further fed into abstractive summarization model. The Abstract Summarization various pre-trained autoregressive language models were experimented out of which GPT-2 and BART (Lewis et al., 2019) were found to perform better as compared to other models. Combining BERT with BART and conditioning was done on input combination of introduction and conclusion (from the original document) along with the extractive summary (derived from the extraction model). The experiment resulted in the best ROUGE (Lin and Rey, 2001) score.

In an attempt to combine extractive and abstractive summary another improvisation that has been experimented is Keywords-Guided Abstractive Sentence Summarization (Li et al., 2020) where he proposes extracting overlapping words between the input and the reference as the ground-truth keywords followed by Multi-task learning i.e., generating summary using the input sentence and the ground-truth keywords. Keywords are generated using the trained keywords extractor for the input sentence in the training set and then fine-tuning the sentence summarizer using the original sentence and the predicted keywords. During testing, first keywords are generated using the trained keywords extractor for the input sentence and then the summary is produced using the input sentence and the predicted keywords. Similar Keyword based extraction followed by abstraction is proposed in work related to COVID-19 Medical Research dataset's abstract text summarization (Tan et al., 2020). He proposes a model where initially source text is scanned to extract keywords using token classification tools such as part of speech tagging packages of NLTK, or part of speech tagging of fine-tuned BERT token classifier. The extracted keywords are categorized into nouns, verbs and noun and verbs. Subsequently the keywords are paired with the gold summary abstract and model is processed using GPT-2.

2.8 Evaluation Metrics

Various performance metrics have been defined to measure the various NLP tasks. BLEU Score (BiLingual Evaluation Understudy) (Papineni et al., 2001) for language

translation, SQuAD (Stanford Question Ansering Dataset) (Rajpurkar et al., 2015) for prediction of answers, GLUE (general Language Understanding Evaluation) (Wang et al., 2018) for collection of tasks and ROUGE (Lin and Rey, 2001) (Recall-Oriented Understudy for Gisting Evaluation) for text summarization.

ROUGE (Chin-Yew Lin and et al. 2020) compares machine generated summaries with reference summaries (human composed summaries) to determine the quality of summaries generated. It takes into account word pairs and n-gram sequences between the two summaries for comparison.

2.9 Discussion (key takeaways)

The study has covered the metamorphosis of Text Summarization in Machine learning encompassing techniques like graph-based approaches, leveraging Reinforcement Learning, Neural Networks, Transformers models and has finally put the spotlight on the gaps that are yet to be filled for long document summarization. Sparse-attention based models like BIGBIRD, LED (longtransformer) and Reformers have a lot of potential yet to be exploited in the field of text summarization for larger context domains. The evolution and metamorphosis of standard Transformer models has been discussed and along with the limitations of the same for larger documents has been explored in the study.

Literature review highlights the below focus areas that need more experimentation and analysis especially in the context of Text summarization of Long Documents such as scientific journals - (1) Exploring Sparse-attention based Transformer variants for long document summarization (2) Combining Extractive and Abstractive Summarization techniques to optimise the Abstractive Summarization output (3) Leveraging sparse attention based Transformers in combination strategy of extractive and abstractive strategy and evaluating the performance on basis of ROUGE Score.

2.10 Summary

This study reviewed a considerable number of literatures from conference proceedings, survey papers, thesis, journal articles and blogs to understand the evolution of Summarization in Machine Learning capacity. It has explored the latest strategies experimented and implemented to enhance extractive and abstractive summarization both in isolation as well as in combination. In the process of the review, we explored

the various standard Transformer Models for text summarization and we also studied the limitations of these models for summarization of larger context datasets such as Scientific journals which are important as the salient features of the subject are spread across the entire document. One of the interesting finds in the entire review was introduction of Sparse-attention based Transformer variants such as BIGBIRD and Longtransformer designed to take longer input sequences.

In the last section, the study has discussed sparse-attention based transformer variants which have been introduced recently have a lot potential for longer documents and this can be leveraged for a good ensemble technique to come up with an effective summarization model and this is the motivation for the current study.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

The study would like to explore how the latest sparse attention-based transformer variant such as BIGBIRD (Zaheer et al., 2020) can be leveraged for an effective text summarization of Covid dataset containing long scientific document corpus. The study will further explore the feasibility of improvisation with an ensemble model leveraging both extractive and abstractive techniques with the above-mentioned transformer variant. The study will include evaluation of performance through ROUGE (Lin and Rey, 2001) metrics.

Subsequent subsections will discuss all steps that were addressed in order to achieve the goal. The flow diagram in Figure 3.2 depicts planned sequence of activities in the modelling and evaluation phase. The flow diagram in Figure 3.1 depicts the overview of the two separate summarization pipelines proposed to be built for comparison.

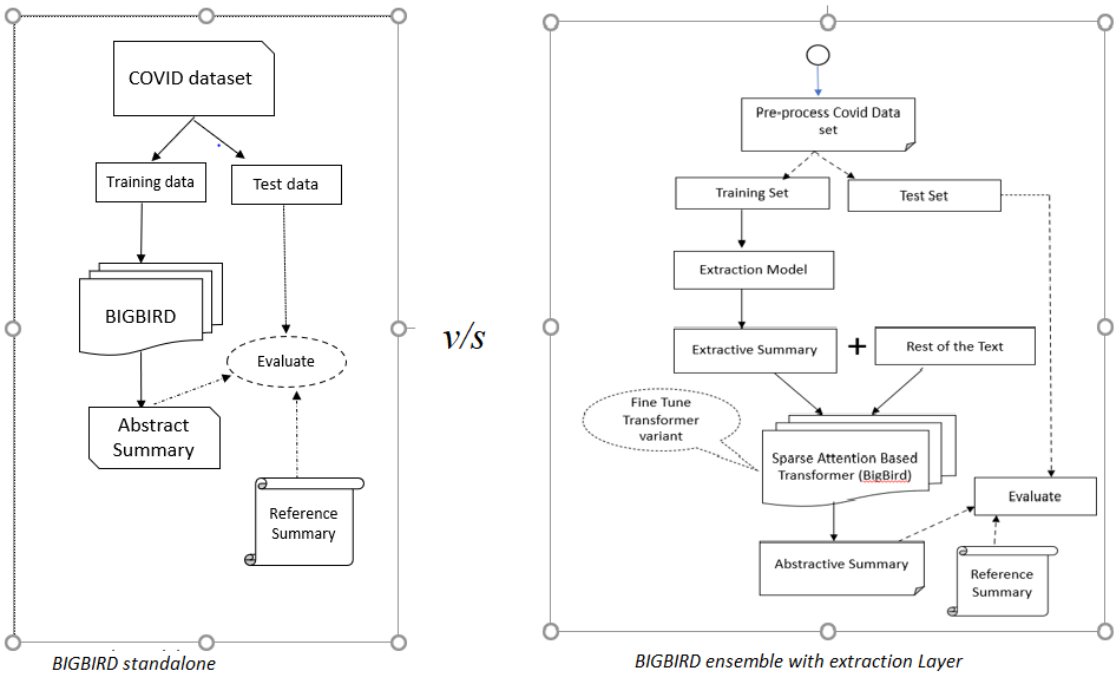

Figure 3.1 High Level Research Approach

The overall research work can be categorized in four broad set of activities

- Data Analysis & Pre-processing
- Training BIGBIRD on CORD-19 dataset
- Training BIGBIRD ensemble with extraction layer on CORD-19 dataset

- Performing evaluation on two summarization pipelines and draw comparisons and propose the best model.

The above specified research activities have been elaborated below.

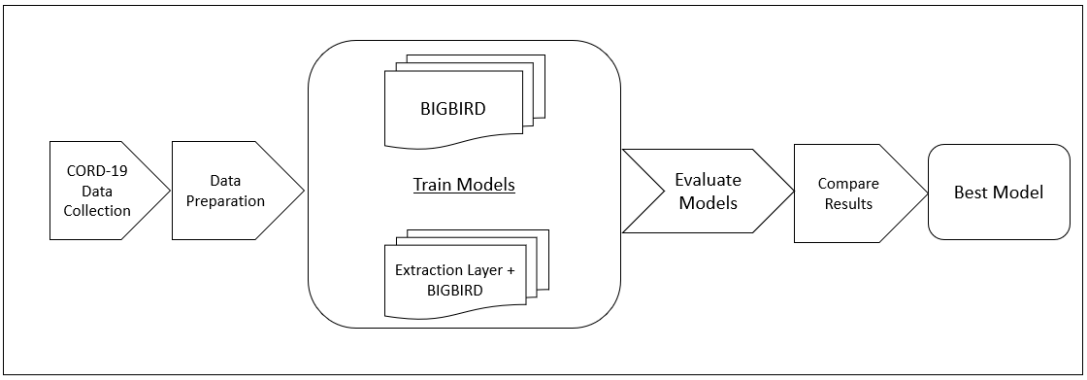

Figure 3.2 Research Framework

3.2 Data Analysis & Pre-processing

This section will provide an overview or probable steps to be performed to analyze, transform and process the existing data and prepare the training data

3.2.1 Dataset Description

Leading Research Groups in collaboration with White House have prepared CORD-19 i.e., COVID-19 Open Research Dataset (Lu Wang et al., 2020) and it has been made public for the research communities across the world. This data set contains over 4000,000 researched scholarly pandemic articles including over 150000 full length articles on coronaviruses including COVID-19 and SARS-CoV-2. This content can be leveraged by the Research communities across the world to synthesize new insights to battle the ongoing pandemic and help in insulating from the pandemics that may come up in the future. By leveraging Natural Language Processing and other AI techniques on this data, a lot of useful work can be done one of the ways to achieve this is by performing efficient Text Summarization. Figure 3.3 depicts the data collection workflow behind CORD-19 research corpus.

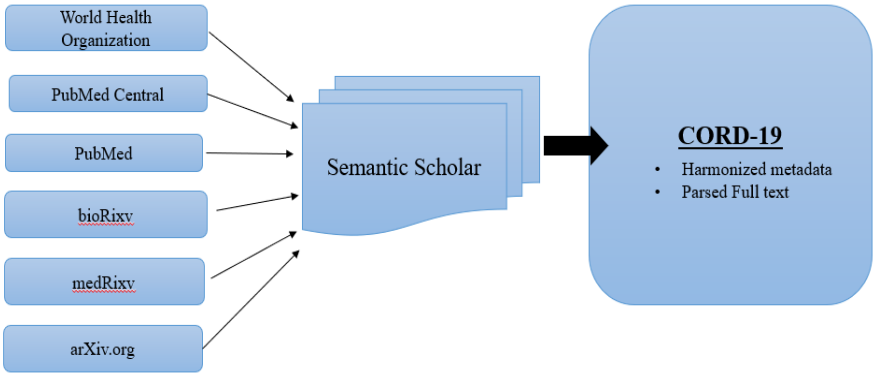

Figure 3.3 Sources of CORD-19 Research data

The article text holds full text of a paper, section name is a list of papers sections, sections contain full text of paper divided into sections. That means, we could identify which text is from the introduction section, which is from the abstract section or conclusion. Table 3.1 gives and word count statistics of CORD-19 research corpus indicating that average length of the articles is 4000 words and abstract summaries have average length of 200 words.

Table 3.1 Data statistics of CORD-19 dataset

Data Set	Total Scientific Articles	Average no. of words per articles	Reference Summaries (Average no. of words)
CORD-19	490904	4000	200

Table 3.2 Research Article Size Distribution

	50%	75%	max
Article (wordcount)	3440	5165	254446
Abstract(wordcount)	157	234	5091

The data statistics in Table 3.2 shows that 75% of the total articles in the document have length up to 5000 words and abstract length as 234 words. Hence it is clear that COVID dataset is dominated by lengthy articles.

3.2.2 Data Pre-processing

- The data from PDFs and other paper documents are parsed to provide structured text which is preserved in JSON schema and converted into readable csv file format.
- The important columns that have been extracted are paper_id, title, authors, affiliations, abstract, text and bibliography.
- As part of data cleaning too short papers and too long papers, papers without abstract were excluded.
- For the scope of the training, only those records with article length up to 5000 words were included as part of training strategy. This was due to the fact that articles with word length 5000 words constituted 75% of the total corpus and BIGBIRD's capacity is 4096 input tokens. So, the strategy is catering to majority of the records in the corpus and is aligned with the capacity of BIGBIRD.
- Removal of irrelevant characters was done as part of data cleaning.
- For the purpose of training only articles and abstract attributes were retained.
- Sections which are subtopics in the articles were extracted and concatenated as single string. This value was added as a derived attribute "sections" in the training data.
- The data was split into training, validation and test sets.

3.3 Training proposed Models

The two summarization pipelines workflows shown in figure 3.1 were trained as below.

3.3.1 Training BIGBIRD

BIGBIRD has greater capacity to process input tokens as compared to other transformers. As the data corpus used in the training data has articles length up to 5000 words and BIGBIRD's capacity is 4096 input tokens, BIGBIRD was trained on the final training data directly. The model was trained on two attributes i.e., articles and sections. Figure 3.2 depicts the workflow of the pipeline trained under this approach.

3.3.2 Training BIGBIRD Ensemble with Extraction Layer

The ensemble planned as part of this strategy consists of an Extraction Layer followed by BIGBIRD's abstractive summarization layer. Figure 3.2 depicts the flowchart of the ensemble. As the length of the input articles in the training samples is up to 5000 words

with average number of records falling in the range of 3000 words, an extractive strategy that can cover the entire document was needed.

The approach proposed for extraction layer is a combination of BERT (Devlin et al., 2019) and K-medoid clustering (Miller, 2019). A pre-trained BERT model is used for sentence embedding. Each sentence in the article is transformed into 768 high dimensional representation. K-medoid clustering analysis is done on the transformed high dimensional representations. The entire flow results in cluster centers which represents the semantic centers of the analyzed text. These semantic centers when collated together can constitute the extractive summary of the articles. The extractive summary is thus constructed using the cluster centers. The BERT model used in the architecture is DistilBERT (Sanh et al., 2019) from Huggingface Transformer. The extraction % targeted in the architecture is 40-50% of the entire document.
The final layer in the ensemble is the abstractive summarization layer using BIGBIRD. The data preparation for this model includes selection of Introduction and Results sections from the articles. Hence for the purpose of training only those records from training data are considered which have Introduction and Results sections. The motive behind selecting Introduction and Results section is based on the general assumption and observation that Introduction and Results sections contain the main context of research articles. Extracted summary from the extraction layer is taken and combined with Introduction and Results sections. The final training dataset consists of "Introduction + Extracted summary + Results"

3.4 Model Evaluation

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Rey, 2001) is one of the widely used performance metric used for evaluating performance of Text Summarization tasks in NLP. ROUGE compares machine generated summaries with reference summaries (ground truth) to determine the quality of summaries generated. It takes into account word pairs and n-gram sequences between the two summaries for comparison. The various ROUGE measures available are ROUGE-N, ROUGE-L, ROUGE-W and ROUGE -S. *ROUGE Recall is the* coverage of machine generated summary with respect to reference summary *F*-Score version of ROUGE is the harmonic mean of ROUGE Recall and ROUGE

Precision. For evaluation purpose in the existing work F-Score version of ROUGE is evaluated.

The final summaries generated by the two summarization pipelines is compared to the ground truth summaries and ROUGE scores are computed. The ROUGE-1 and ROUGE-2 scores obtained in both the pipelines are compared to find the best summarization strategy. The length of the input article is one of the properties of the CORD-19 dataset that can influence the quality of summarization. For the purpose of evaluation in this research, we have defined three size ranges for the data set as below.

- Short articles (word count<1500)
- Medium articles (1500<word count<4000)
- Long articles (wordcount > 4000)

The evaluation of each trained model is done across the above three ranges of data length to find patterns (if any) of performance deterioration in BIGBIRD ensembles due to the length factor. For a fair comparison the evaluation of the two summarization models is done the same test data.

3.5 Comparison between the trained models

The evaluation results under various scenarios such as ROUGE scores across various size of the input articles, execution time, quality of the abstracts generated were compared between the two trained models to suggest the best model.

3.6 Summary

This study intends to cover the two high level objectives. The first part is to explore the potential of the state-of-art sparse attention-based transformer- BIGBIRD for text summarization of long scientific documents i.e., COVID data set taken in the existing study. The first part of section describes the dataset statistics, pre-processing strategies and limitations of COVID dataset. The section also highlights the limitations of the standard Transformer models to handle lengthy input texts for summarization due to quadratic dependency of self-attention mechanism in the architecture and existing resource constraints. Further it also elaborates the potential of the latest sparse attention-based transformers to process longer input sequences which is 8x as compared to standard transformers.

The section discussed the detailed BIGBIRD ensemble with extraction layer strategy which is part II of the study (Figure 3.1). The extractive summarization strategy is one

of the inputs to be used to train the final model in the ensemble. Significance of ROUGE score for finalizing the extractive summary model has been briefly elaborated.

The section also highlights the advantage of BIGBIRD to handle 8x longer input sequences as compared to the standard transformers, which can be leveraged in this existing study to include more text for conditioning and training, which was a limitation in the standard transformers.

The final section elaborates the quantitative and qualitative evaluation of the final model based on ROUGE scores. ROUGE score comparison at various levels starting from selection of best extractive summarization model to the selection of final conditioning scenario in final ensemble model will play the vital role. The comparison of the performance of the final model will also be done with that of the standard transformers. F-score version of ROUGE will be used in all the evaluations as performance metrics.

CHAPTER 4: ANALYSIS

4.1 Introduction

COVID-19 pandemic has resulted in an unprecedented research worldwide and has created a need for accelerated research work to get optimum results in short turnaround time for faster damage control along with future preparedness to avert similar crisis. CORD-19 Open Research Dataset (Lu Wang et al., 2020) is a continuously growing corpus of coronavirus research work. The corpus primarily integrates papers and preprints from several sources. The paper in the dataset is the base unit of published knowledge, whereas

a preprint is an unpublished yet publicly available counterpart of a paper. The primary objective of CORD-19 dataset is to establish a centralized platform where biomedical experts, computing community and policy makers can work together to churn out effective treatments, prevention and management policies for COVID-19.

4.2 Sources of papers

CORD-19 is curated with rich collection of metadata and structured full text papers to equip the computing community to explore and develop text mining and Information retrieval systems.

The sources of Papers in CORD-19 dataset are as below:

- PubMed Central (PMC)
- PubMed
- World Health Organization's Covid-19 Database,4
- preprint servers of bioRxiv, medRxiv, and arXiv

The metadata of the papers collected from different sources is harmonized and deduplicated using clustering. Extraction of full text and bibliographies from each PDF is done using PDF parsing pipeline and is converted to JSON files. Subsequently post processing is done to clean up the links between inline citations and bibliography entries.

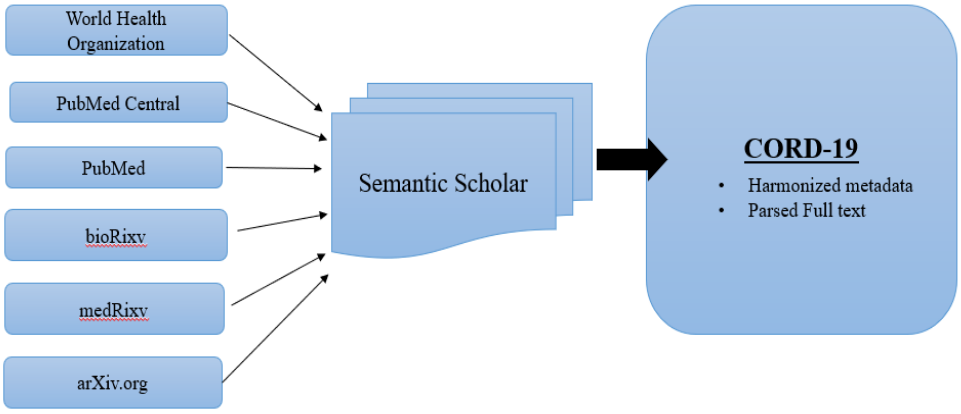

Figure 4.1 Sources of CORD-19 Research articles

Each paper has bibliographic metadata as title, authors, publication venue; unique identifiers such as a DOI, PubMed Central ID, PubMed. The papers retrieved as part of The PMC Public Health Emergency Covid-19 initiative contain COVID-19 related keywords. The keywords used are depicted in Table 4.1.

Table 4.1 Research Paper Retrieval Keywords

Paper Retrieval - Keywords
"COVID" OR "COVID-19" OR "Coronavirus" OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome"

4.3 Research Objectives of CORD-19 dataset

A spike in CORD-19 publication is observed in 2020 in response to COVID pandemic. CORD-19 aims to connect the machine learning community with biomedical domain experts to identify effective treatments and management policies for COVID-19. The objective is to extract useful information from historical coronavirus literature, synthesize knowledge from literature in short span of time and address questions related to symptoms, mortality rates, infection, identification of drugs for repurposing and interaction with other diseases. A number of papers on COVID-19 is being published every day and effective automated methodologies are needed to extract, analyse and synthesize the vast growing content.

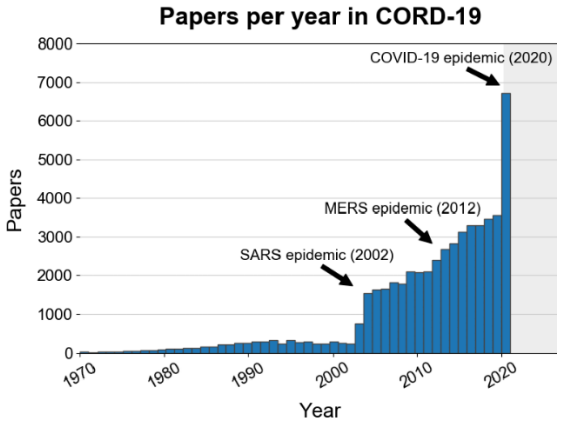

Figure 4.2 Spike trends in COVID research data (Lu Wang et al., 2020)

4.4 Importance of Automated Text Summarization in CORD-19

In the current pandemic one of the major challenges for biomedical domain experts and researchers is to be updated with recent papers, extract relevant information and synthesize knowledge from the corpus of CORD-19 data. Time is the key differentiator in the entire process considering the nature of the crisis. A short turnaround time is required in the entire workflow as reading thousands of articles to filter the relevant content can be exhaustive, at times non-productive and prone to oversights. An automatic text summarization of CORD-19 datasets can assist the researchers and clinicians to expand the purview of scientific exploration by reaching large corpuses, filtering and placing their focus on the relevant content. The entire strategy can help in minimizing redundancy, reduce human oversights and infuse agility in the entire process flow of research work.

4.5 Challenges in Existing Text Summarization strategies

The text summarization strategies explored for CORD-19 dataset as of now have one common limitation with respect to the length of input sequences. The best performing summarization pipelines that include transformers such as BART, GPT-2, T-5 can only process input sequences up to 1024 length. This limits the quality of the summarization as most of the content have to be cropped. A number of alternate strategies such as including only Introduction and Conclusion sections of the scientific articles to generate summaries or combining extraction followed by abstraction have been proposed. The strategies proposed so far work within the limited input length constraints. Longer input sequences is one of key features of CORD-19 datasets. Table 4.2 and Table 4.3 depicts the statistical analysis size distribution of the research articles.

Table 4.2 Input data size distribution

Article Length (words)	<1024	<5000	>5000
% Of Records	9%	80%	20%

Table 4.3 Quantile distribution of Article Size

	Median	80%-ile
Articles Length (wordcount)	3440	5160

Table 4.2 indicates that only 9 % of total corpus have article length up to 1024 words, whereas 80% fall in range within 5000 words and 20% have length greater than 5000 words. In table 4.3 The median (50%) is found to be around 3500 words. The above analysis highlights the gap in the existing Summarization strategies and is suggestive of the fact that the probability of important contexts getting "lost in summarization" is high due to input length constraints of existing strategies.

4.5 Dataset Description

CORD-19 Open Research Dataset (Lu Wang et al., 2020) containing over 29,000 scholarly articles about coronavirus family and COVID-19 is obtained from Kaggle (Kaggle CORD-19 challenge, n.d.). The data is in JSON format, which was further processed and converted into csv format. The final converted CSV files were used and were categorized as shown in table 4.4.

Table 4.4 Data Source detailed view

Source	Records
Commercial use subset	9524 entries with 9 columns
Non-commercial use subset	2490 entries with 9 columns
PMC custom license subset	26505 entries with 9 columns
bioRxiv/medRxiv subset	1625 entries with 9 columns

Data set from four sources contribute to 40,144 records with 10 attributes. The primary attributes present in the records are paper_id, title, authors, affiliations, articles, abstract and bibliography. The attributes that are vital for the purpose of this research are articles and abstract. Articles are scientific journals, research papers & preprints related to coronavirus family and COVID-19. Abstracts are the ground truth summaries of each article present in the dataset.

The word count statistics in table 4.5 describes length distribution of the dataset. The *median* and the 75% quantile of the article, and abstract (summary) clearly indicates that 75% of the article have wordcount <=5165 and abstract length <=234.

Table 4.5 Article and Abstract length quantile distribution

	50%	75%	max
Article (wordcount)	3440	5165	254446
Abstract (wordcount)	157	234	5091

The data from the above processed csv files have the same attributes and were combined into a single csv file. Figure 4.3 is the snapshot of the attributes in the final processed dataset.

paper_id	title	authors	affiliations	abstract	articles	bibliography	raw_authors	raw_bibliography
572a7a9b3e92b960d92d9755979eb94c448bb5	Immune Parameters of Dry Cows Fed Mannan Oligo...	S T Franklin, M C Newman, K E Newman, K I Meek	S T Franklin (University of Kentucky, 40546-02...	281	INTRODUCTION\n\nThe periparturient period is a...	Immune response of pregnant heifers and cows t...	[{'first': 'S', 'middle': ['T'], 'last': 'Fran...	{'BIBREF0': {'ref_id': 'b0', 'title': 'Immune ...
b790e8366da63c4f5e2d64fa7bbd5673b93063c	Discontinuous Transcription or RNA Processing ...	Beate Schwer, Paolo Vista, Jan C Vos, Hendrik ...	Beate Schwer, Paolo Vista, Jan C Vos, Hendrik ...	1	Discontinuous\n\nTranscription or RNA Processi...	Poly (riboadenylic acid) preferentially inhibi...	[{'first': 'Beate', 'middle': [], 'last': 'Sch...	{'BIBREF0': {'ref_id': 'b0', 'title': 'Poly (r...
4f204ce5a1a4d752dc9ea7525082d225caed8b3	NaN	NaN	NaN	1	Letter to the Editor\n\nThe non contact handhe...	Novel coronavirus is putting the whole world o...	[]	{'BIBREF0': {'ref_id': 'b0', 'title': 'Novel c...

Figure 4.3 Attributes of baselined dataset

4.4 Data Preparation for Research Analysis

The data was sourced, then checked for null values, invalid values, word count analysis and study the structures of the different types of articles. The interpretation of each data preparation is explained in different section below.

4.4.1 Handling missing values

In the data cleaning process, it was found that around 27% of total records had no ground truth summaries(abstracts). Figure 4.4 shows the null value distribution across the attributes. Abstract is the target label of the dataset; hence these records have no contribution in the training process hence these rows were dropped. Attributes like title, authors, affiliations also have null values but these columns will be treated separately in the subsequent steps.

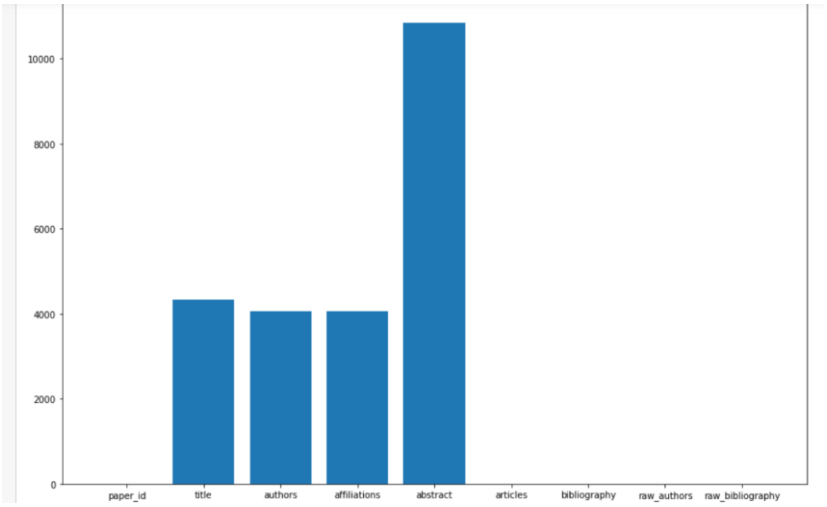

Figure 4.4 Missing Values in CORD-19 dataset

4.4.2 Elimination of Variables

The dataset has 9 attributes out of which articles and abstract are the only attributes that will be used in the training process. The other attributes (affiliations, bibliography, title, raw_authors, raw bibliography) were dropped.

4.4.3 Handling articles of foreign languages

A number of records in the dataset are articles in foreign languages . These records are out of scope for our research so these records have been eliminated from our final dataset. The python package used to identify the foreign language texts used here is

"langdetect". Figure 4.5 is a snapshot of sample foreign language article from the the dataset.

```
foreign_language=cord19.loc[16816]
print(foreign_language)
```

کلیدی: های واژه برونشی ت مراز پلی ای زنجیره واکنش آزمون عراق، گوشتی، جوجه پرندگان، عفونی

Figure 4.5 Foreign Language text sample in COVID dataset

4.4.4 Handling Outliers and scoping input length range

Articles that were too long or too short were removed from the final dataset. Similarly abstracts that were too long or too short were also excluded from the final dataset. Statistical Analysis of the dataset depicted in Table 4.6 revealed that 80% of the records had article length reaching maximum up to 5500 words and 80% of abstracts had length reaching maximum up to 270 words. For the research records with article length reaching up to 5500 were considered, the remaining records were excluded from final dataset. This range selection not only covered 80% of the corpus but also was in tandem with our strategy to use Sparse Based attention-based transformer- BIGBIRD, which can be handle input sequences up to 4096. Figure 4.6 is snapshot of size distribution representation before the clean up process.

Table 4.6 Data length Median view

	Quantile	
	50%	80%
Articles	3595	5589
Abstract	200	273

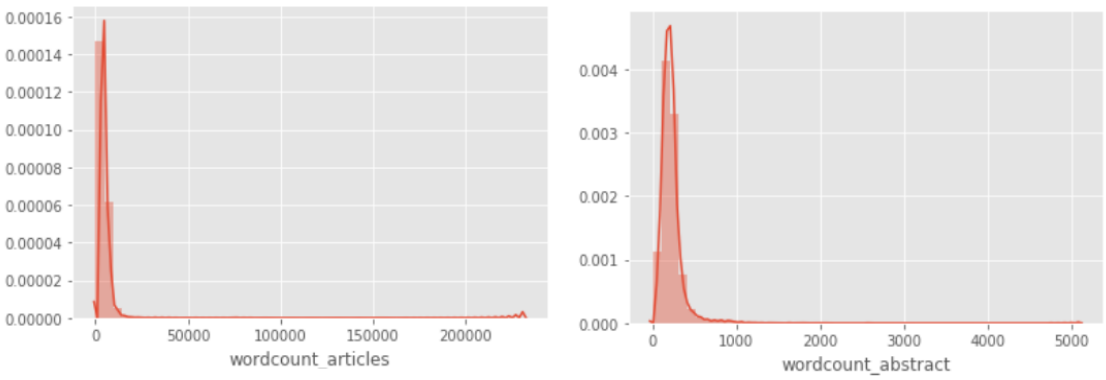

Figure 4.6 Articles and Abstract length distribution

4.4.5 Text Cleaning of Input data

Text Clean-up was done to remove special characters, character encodings, citations, hyperlinks, extra spaces, new lines and other related items. The intent here is to get rid of redundant information from the input text to so as to retain only the relevant content for summarization. This strategy significantly contributed in reduction of word counts and thereby optimizing the input length. The final processed dataset consists of around 17,000 records with 50% of the records with input length between 3000-5500 words, which is a rich corpus to train the model. The length of the ground truth summaries of more than 50% of the records are in range between 150-280 words. Figure 4.7 shows the size distribution of articles post clean up.

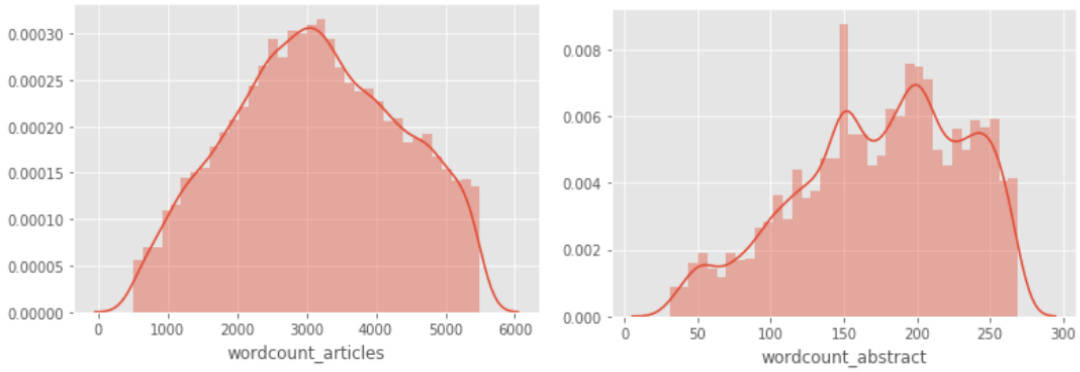

Figure 4.7 Articles and Abstract length distribution post clean up

4.4.6 Derived Attributes

CORD-19 dataset has various subsections in the articles. These subsections are variable in nature and is dependent on the topics covered in each article. On analyzing the input articles, it was observed that Introduction, Discussion/ Results are the common subsections present in majority of the articles. However, 35% of the articles do not have "Introduction" as subsection and 25% of the articles have the "Results/ Discussion" sections missing. Subsections are one of the supportive attributes that helps to segregate the topics being discussed in the journals, so "sections" is a derived attribute in the final training dataset. A python code using regex is written that scans the input articles and identifies the section from each article and concatenates the final outcome. The concatenated string is added as a value in "sections" attribute of the training dataset. Figure 4.8 is the sample snapshot of "sections" attribute for a few articles.

Sections extracted from Articles
Introduction, Experimental apparatus,Computational analysis, Results and discussion, Conclusion
Introduction , Patients, PCR based test for respiratory virus detection, Quantitative real time PCR technique for A H1N Statistical analysis, Respiratory virus infection rates and distribution, Patient data and clinical diagnosis, Discussion
Introduction ,Sequence Retrieval and Analysis, Potential Designing of siRNA and miRNA, Similarity Search and Target Alignment GC Calculation and Secondary Structure Prediction, Calculation of RNA, RNA Interaction Through Thermodynamics, Calculation of Heat Capacity and Concentration Plot of Duplex, Prediction of Secondary Structure , Results and Discussion, Conclusion

Representation of derived sections in final dataset

Sections
Introduction Experimental apparatus Computational analysis Results and discussion Conclusion
Introduction Patients PCR based test for respiratory virus detection Quantitative real time PCR technique for A H1N Statistical analysis Respiratory virus infection rates and distribution Patient data and clinical diagnosis Discussion
Introduction Sequence Retrieval and Analysis Potential Designing of siRNA and miRNA Similarity Search and Target Alignment GC Calculation and Secondary Structure Prediction Calculation of RNA RNA Interaction Through Thermodynamics Calculation of Heat Capacity and Concentration Plot of Duplex Prediction of Secondary Structure Results and Discussion Conclusion

Figure 4.8 Derived attributes in training data

4.4.7 Final Dataset post-processing

The final dataset used for research post processing includes three attributes - articles, sections and abstract. Articles and sections are input labels where articles contain the research content on coronavirus family and COVID-19 whereas sections are sub are indicators of content structuring using various sub-topics. Abstract is the target label containing human generated summaries. Figure 4.9 is the snapshot of sample records post processing.

	articles	sections	abstract
0	INTRODUCTION\n\nThe periparturient period is a...	INTRODUCTION Animals and Animal Procedures Sam...	Abstract\n\nThe objective of this study was to...
1	Introduction\n\nMullis et al. developed the po...	Introduction Experimental apparatus Computatio...	Abstract\n\nThis research reports the design, ...
2	Introduction\n\nHospital emergency departments...	Introduction Patients PCR based test for respi...	Abstract\n\nTo characterize respiratory virus ...
3	Introduction\n\nAminopeptidase N hydrolyses ...	Introduction Methods Cross linking of proteins...	Abstract\n\nBovine renal brush-border membrane...
4	INTRODUCTION\n\nCoronavirus gene expression in...	INTRODUCTION Subcellular localization of the c...	Abstract\n\nThe coronavirus 3C-like proteinase...

Figure 4.9 Training Data Set

4.4.8 Exploratory Data Analysis

A bivariate analysis was attempted to establish a relationship between article length and abstract length. However, a negligible correlation was discovered.

	wordcount_articles	wordcount_abstract
wordcount_articles	1.000000	0.199648
wordcount_abstract	0.199648	1.000000

Figure 4.10 Weak Correlation between article and abstract sizes

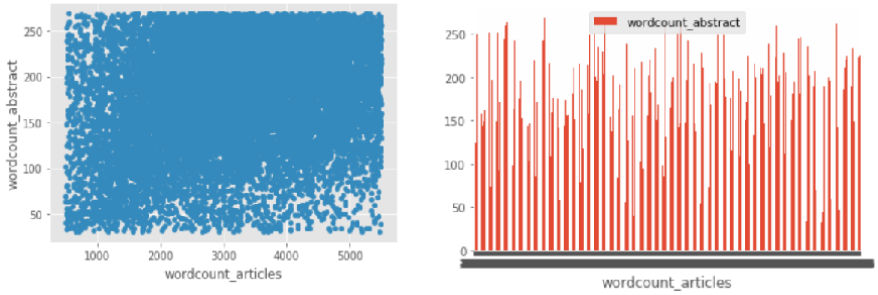

Figure 4.11 Weak correlation bar graph representation

Figure 4.10 and 4.11 shows that there is no correlation between article length and abstract length. The final dataset includes articles of maximum length reaching up to 5500 words and abstract length reaching up to 270 words. 75% of the total dataset are well structured as they have Introduction, Results and Discussion sections. 25% of the dataset do not contain these basic sections. For the research task the priority is given to articles with Introduction, Results and Discussion sections as these datasets will enable us to draw a comparison with Full attention-based Transformers such as BART due to input size limitation of 1024 sequences.

The quantile distribution of final dataset can be summarized in the figure 4.12.

	Quantile				
	min	25%	50%	75%	max
Article	501	2241	3111	4060	5499
Abstract	50	137	181	220	269

Figure 4.12 Data size quantile distribution in the cleaned dataset

4.5 Building Summarization Model

As part of the research two summarization pipelines were built and comparison was done to select the best ensemble.

4.5.1 Summarization pipeline with only BIGBIRDPEGASUS

As part of this workflow depicted in figure 4.13, the final COVID dataset is trained on BIGBIRD for summarization task. BIGBIRD is capable of processing input sequences up to 4096 and the records in the final COVID dataset are also around 5000 words, BIGBIRD was leveraged to train the model on COVID dataset.

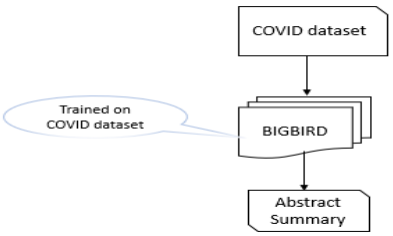

Figure 4.13 BIGBIRD standalone summarization pipeline

4.5.1.1 Training BIGBIRD

This model is trained on official checkpoint *google/bigbird-pegasus-large-pubmed.* This is already a BigBird's fine-tuned model on PubMed dataset. PubMed data consists of scientific papers in the field of medicine. The advantage of using this checkpoint is that model already understands the lengthy scientific language from the medical journals which is better than initialising randomly from scratch. The model was trained with below steps:

- Split the data into train, validation and test sets
- Load official checkpoint *google/bigbird-pegasus-large-pubmed* (bigbirdpegasus-large-pubmed, n.d.)
- Setup tokenizer using BIGBIRD tokenizer for converting text into numbers which the model can take as input.
- Set up "collate_fn" for tokenization and train the model with the dataset.
- The training of BIGBIRD is carried out in Paperspace equipped with 1x Quadro P6000 GPU. A total of 5 epochs are performed. The training dataset consists of 15000 training samples The training parameters include the learning rate 5e-5, with batch size =2 and gradient accumulation of 2 steps. The model is trained in fp32 (single precision floating point) mode.
- The challenge here is that BIGBIRD takes a lot space and time to train. The model took 4 days to train for 5 epochs.

The model was built using the below hyperparameters:

- attention_type = "block_sparse"
- model_type: "bigbird_pegasus"
- block_size = 64
- learning_rate=5e-5

- num_train_steps = 10000
- dropout: 0.1
- activation_function: "gelu_new"
- tokenizer: "BigbirdPegasusTokenizer"
- length_penalty": 0.8
- max_position_embeddings: 4096
- encoder_layers: 16
- decoder_layers": 16
- num_train_epochs=5
- per_device_train_batch_size=2
- per_device_eval_batch_size=2
- gradient_accumulation_steps=2

4.5.1.2 Generating Predicted Summaries

To generate predicted abstract summaries an evaluation function is created which tokenizes each article up to a maximum length of 4096 tokens and
beam search is used to generate the predicted abstract of the articles. The parameters used in beam search are:

- num_beams=5
- length_penalty=0.8
- max_length=256

In the last step the predicted abstract tokens are decoded and the resulting predicted abstract string is saved in the batch.

4.5.1.3 Evaluating the Trained BIGBIRD

The final trained model is used to evaluate the Test dataset using ROUGE-1, ROUGE-2, ROUGE-L metrics. The evaluation time taken by BIGBIRD is also higher. The evaluation was done on 200-600 datasets. The objective here was to find the highest ROUGE scores obtained by the model.
The rouge metric from Huggingface implementation is used. This is a wrapper around Google Research reimplementation of ROUGE. This calculates average rouge scores for a list of hypotheses and references. The outcome of the evaluation is the aggregated rouge score for each rouge parameters which is returned as low,

mid and high. For the purpose of research evaluation, the high score is used as reference. Figure 4.14 is the snapshot of sample outcome of evaluation on 100 samples.

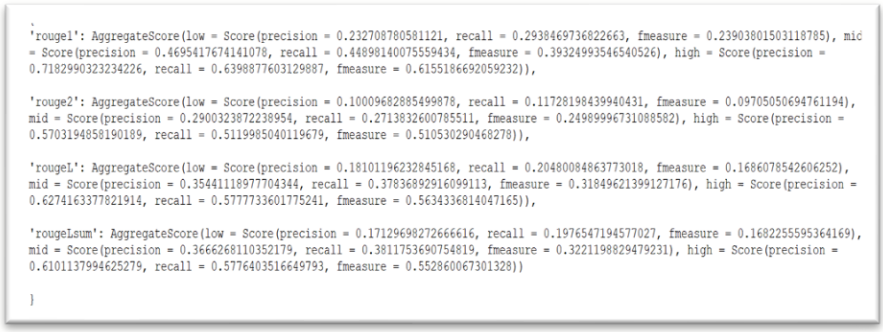

Figure 4.14 Sample ROUGE aggregated score

4.5.2 Summarization with Extraction Layer and BIGBIRD Abstraction

For training this pipeline only those records that have Introduction and Results/Discussions sections are taken. Extraction summary is performed on these datasets. The resultant training data for the final summarization layer with BIGBIRD has the below structure:

- Introduction + Extractive Summary + Results/ Discussions

The summarization pipeline can be outlined in figure 4.15.

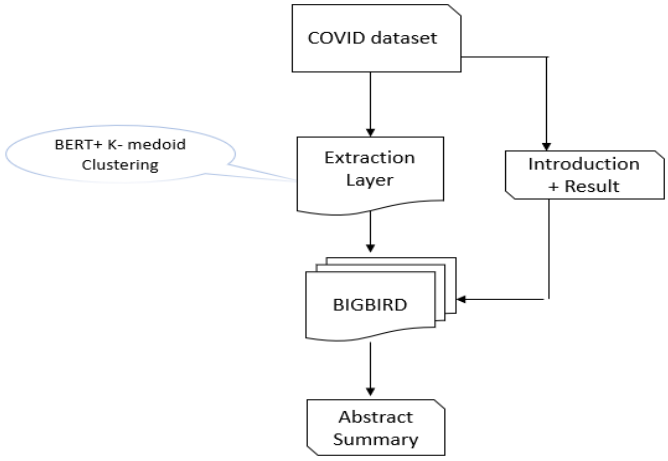

Figure 4.15 BIGBIRD with extraction layer

4.5.2.1 Preparing the Extractive Summary Layer

As the length of the input articles is up to 5000 words with average number of records falling in the range of 3000 words, an extractive strategy that can cover the entire document was needed.

The approach implemented here is a combination of BERT and K-medoid clustering. A pre-trained BERT model was used for sentence embedding. Each sentence in the article was transformed into 768 high dimensional representation. K-medoid clustering analysis is done on the transformed representations. The entire flow results in cluster centers which represents the semantic centers of the analyzed text. The extractive summary is constructed using the cluster centers. The BERT model used in the architecture is DistilBERT (Sanh et al., 2019) from Huggingface Transformer. The extraction % targeted in the architecture is 40-50% of the entire document.

4.5.2.2 Training the Abstraction layer with BIGBIRD

The extractive summary obtained is combined with Introduction and Result sections of the training dataset. The final training dataset has the structure of the input text as:

- Introduction + Extractive Summary + Results/ Discussions

The data is split into train, validation and test sets. This model is trained on official checkpoint *google/bigbird-pegasus-large-pubmed.* The training strategy for BIGBIRD is same as that of followed is same as explained in “Summarization pipeline with only BIGBIRDPEGASUS”.

4.5.2.3 Generating Predicted Summaries

To generate predicted abstract summaries an evaluation function is created which tokenizes each article up to a maximum length of 4096 tokens and
beam search is used to generate the predicted abstract of the articles. The parameters used in beam search are:

- num_beams=5
- length_penalty=0.8
- max_length=256

In the last step the predicted abstract tokens are decoded and the resulting predicted abstract string is saved in the batch.

4.5.2.4 Evaluating the Trained Ensemble

The final trained ensembles were evaluated on the Test dataset using ROUGE-1, ROUGE-2, ROUGE-L metrics separtely. The evaluation was done on 200-600 datasets. The objective here was to find the highest ROUGE scores obtained by the model. The details of the evaluation is detailed in chapter 5.

4.6 Performance comparison with BART

To evaluate the performance of BIGBIRD with respect to other models such as BART, a common baseline had to established as BIGBIRD's huge architecture can process 4096 token sequences and BART can only process 1024 token sequences. To assess the comparison of BIGBIRD with BART on fair grounds the below steps were performed: The Training sets with only those records which had Introduction and Results section were taken into account. BIGBIRD model was trained with only Introduction and Results section of the COVID training dataset. Training of BART model was done with the same dataset

	Quantile	
	50%	75%
Introdiction + Results	1292	1748

Figure 4.16 Word Count statistics for introduction and Results

Inclusion of only the "Introduction" and "Results" sections as part of training strategy is based on general observation that Introduction and Results convey the main context of the entire document. This strategy helped in drastically reduced the length of the input documents which helped in giving a fair ground to both BIGBIRD and BART for comparison. However, it was observed that since the length of the COVID dataset is huge, inclusion of only the Introduction and Results section still has word counts higher than 1000 for most of the records as shown in figure 4.16.

The results were evaluated on the test dataset containing the Introduction and Results sections and comparisons between the outcome was done.

4.7 Comparing final results across the trained Models

The results across all the trained models were compared and performance of BIGBIRD was evaluated on test data using ROUGE-1, ROUGE-2 and ROUGE-L metrics. Precision, Recall and F-1 measures of rouge scores are evaluated.

4.8 Final Model selection

The ensemble with highest ROUGE (Lin and Rey, 2001) score is finalized. ROUGE-1, ROUGE-2 are primarily considered for evaluation. F-1 measure of ROUGE-1 and ROUGE-2 is the primary performance metrics considered for evaluation. It was observed that better performance was obtained in the summarization pipeline with BIGBIRD standalone. The details of the evaluation and selection is covered in chapter 5.

4.9 Summary

The summarization pipelines were built successfully and evaluated separately on a common test dataset using ROUGE metrics. The two models were compared and BIGBIRD without the extraction layer was found to be efficient than the other model. Additionally, performance of excessively long documents is evaluated on the final model using extractive summary as input. BIGBIRD is found to be highly efficient for long document summarization because of its capability to process input token up to 4096. The challenge here is huge memory requirement and longer training duration. Hence training time required by BIGBIRD is higher than the other transformers.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

The aim of the research is to explore the capabilities of sparse based attention transformer BIGBIRD (Zaheer et al., 2020) in Automatic text summarization of CORD-19 dataset (Lu Wang et al., 2020) . BIGBIRD has the capability to reduce quadratic memory requirement by full attention to linear requirements and process input sequences up to 4096 tokens, which makes it a good candidate for summarization of lengthy research documents. The existing transformers cannot process longer input sequences which has an impact on quality of summarization especially in biomedical domain where the length of document is usually large and important contexts can be lost due to input size constraints. As BIGBIRD can process more inputs, it can learn more contexts and improve the quality of summarization. In this research, the possibility of leveraging BIGBIRD in the coronavirus pandemic to solve the challenge of CORD-19 research data summarization is explored. The research analyses BIGBIRD based summarization ensembles and performance is evaluated and compared across different ensembles and the best ensemble is suggested.

A python code is developed to understand the CORD-19 dataset features and build BIGBIRD CORD-19 Summarization model. The same code is extended to create ensembles with both extraction and abstraction layers and performance comparisons are done. Further comparison with BART is also attempted. This section describes the results of research analysis.

5.2 Model Evaluation Strategy

As part of research analysis, two summarization pipelines with BIGBIRD were trained and evaluated and comparisons were drawn to assess the performance of BIGBIRD on COVID dataset.

5.2.1 Defining the Evaluation strategy

The length of the articles in CORD-19 dataset is an important feature as it influences the quality of summarization.

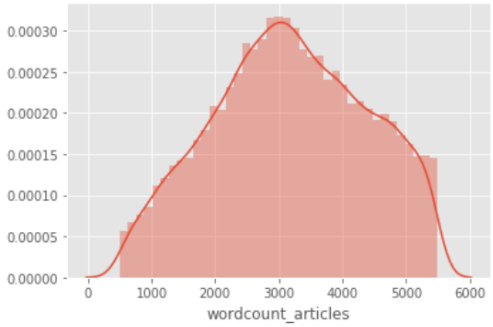

Figure 5.1 Size Distribution of Input Articles

This feature is variable in nature as the size of the articles is spread across the entire dataset (figure 5.1) in the range within 500-word count to 5000+ word count, the evaluation was done on below the below size range:

- Short articles (word count<1500)
- Medium articles (1500<word count<4000)
- Long articles (wordcount > 4000)

To set a common baseline for comparison between the two summarization pipelines same test data is used. Evaluation is done on the same test data and results are compared based on rouge scores obtained in each pipeline.

5.2.2 Metric for evaluation

The rouge metric library from Huggingface implementation is used to evaluate the performance of the model. This is a wrapper around Google Research reimplementation of ROUGE. This library calculates average rouge scores for a list of hypotheses and references. Aggregated scores of ROUGE-1, ROUGE-2 are ROUGE-L for a list of hypotheses are given as outcome. The aggregated score is structured as low, mid and high. In this research, the high score section from the aggregated score is used as reference to identify the peak performance point of the model.

Figure 5.2 is the snapshot of sample outcome of evaluation on 50 samples.

```
{'rouge1': AggregateScore(low=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586),
                          mid=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586),
                          high=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586)),
 'rouge2': AggregateScore(low=Score(precision=0.9319727891156463, recall=0.7965116279069767, fmeasure=0.8589341692789968),
                          mid=Score(precision=0.9319727891156463, recall=0.7965116279069767, fmeasure=0.8589341692789968),
                          high=Score(precision=0.9319727891156463, recall=0.7965116279069767, fmeasure=0.8589341692789968)),
 'rougeL': AggregateScore(low=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586),
                          mid=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586),
                          high=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586)),
 'rougeLsum': AggregateScore(low=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586),
                             mid=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586),
                             high=Score(precision=0.9797297297297297, recall=0.838150289017341, fmeasure=0.9034267912772586))}
```

Figure 5.2 Sample ROUGE Score aggregated output

5.3 Evaluation of Models

BIGBIRD standalone and BIGBIRD ensemble with extraction layer were evaluated on the same dataset on ROUGE scores and comparisons were drawn.

5.3.1 Evaluation of summarization pipeline with BIGBIRD (standalone)

In the summarization pipeline involving only BIGBIRD, the model is trained using official checkpoint "*google/bigbird-pegasus-large-pubmed"* on CORD-19 dataset for 5 epochs and the model was evaluated on test dataset using ROUGE-1, ROUGE-2 and ROUGE-L metrics. Below are the model validation results extracted from aggregated score obtained from rouge metric library of Huggingface. The highest aggregated score obtained on the test for various size ranges are shown in table 5.1, table 5.2 and table 5.3:

- Short Articles (word count<1500)

Table 5.1 Evaluation Results- Short Articles

	Precision	Recall	f-measure
ROUGE-1	71.83%	63.99%	61.55%
ROUGE-2	57.03%	51.19%	51.05%
ROUGE-L	62.74%	57.77%	56.34%

- Mid-sized articles (1500<word count<4000)

Table 5.2 Evaluation Results- Medium Articles

	Precision	Recall	f-measure
ROUGE-1	56.41%	46.21%	50.76%
ROUGE-2	31.33%	25.61%	28.18%
ROUGE-L	38.88%	31.81%	35.00%

- Lengthy articles (wordcount > 4000)

Table 5.3 Evaluation Results- Long Articles

	Precision	Recall	f-measure
ROUGE-1	59.90%	49.40%	54.14%
ROUGE-2	26.21%	21.60%	23.68%
ROUGE-L	35.26%	29.08%	31.88%

A few samples from the summaries generated by BIGBIRD were manually evaluated and the results were satisfactory. The Abstract generated showed satisfactory coverage of the context of the article in majority of the samples evaluated. In few samples it was seen that the generated summary was longer than the ground truth and had better coverage of the context. Figure 5.3, 5.4 and 5.5 are the snapshots of generated summary and ground truth summary pairs from the sample.

Abstract Generated by BIGBIRD (Standalone)	Ground Truth Summary
The spike protein of severe acute respiratory syndrome coronavirus is responsible for mediating membrane fusion between the virus and the host cell. In this study, we investigated the importance of a conserved domain flanked by cysteines 822 and 833 in the function of membrane fusion and virus entry. The domain was found to be highly conserved within the virus family and was found to be essential for the successful function of membrane fusion and virus entry. In the context of virus particle, we observed a drastic loss in fusogenic ability in the mutants of the domain flanked by cysteines 822 and 833, in comparison to the wild type spike protein, in cell cell fusion assays and in the generation of pseudotyped virions. In contrast, we observed no effects on either surface expression or receptor binding, but observed a drastic loss in fusogenic ability in the mutants of the domain flanked by cysteines 822 and 833, in comparison to the wild type spike protein, in cell cell fusion assays and in the generation of pseudotyped virions. We conclude that the domain flanked by cysteines 822 and 833 is essential for the successful function of membrane fusion and virus entry.	The S2 domain of the coronavirus spike (S) protein is known to be responsible for mediating membrane fusion. In addition to a well-recognized cleavage site at the S1-S2 boundary, a second proteolytic cleavage site has been identified in the severe acute respiratory syndrome coronavirus (SARS-CoV) S2 domain (R797). C-terminal to this S2 cleavage site is a conserved region flanked by cysteine residues C822 and C833. Here, we investigated the importance of this well conserved region for SARS-CoV S-mediated fusion activation. We show that the residues between C822-C833 are well conserved across all coronaviruses. Mutagenic analysis of SARS-CoV S, combined with cell-cell fusion and pseudotyped virion infectivity assays, showed a critical role for the core-conserved residues C822, D830, L831, and C833. Based on available predictive models, we propose that the conserved domain flanked by cysteines 822 and 833 forms a loop structure that interacts with components of the SARS-CoV S trimer to control the activation of membrane fusion.

Figure 5.3 Comparison of BIGBIRD summaries with Ground Truth Summaries

Abstract Generated by BIGBIRD (Standalone)	Ground Truth Summary
The crazy paving sign is an area of scattered or diffuse ground glass attenuation with superimposed interlobular septal thickening and intralobular lines on thin section computed tomography images of the lungs. The crazy paving sign can be caused by a variety of diseases, including infectious, neoplastic, idiopathic, inhalational, and hemorrhagic disorders. In this report, we describe a 69 year old woman with parainfluenza virus pneumonia who had a crazy paving sign on high resolution computed tomography.	The crazy-paving sign is the appearance of a smooth linear pattern superimposed on an area of ground-glass opacity on thin-section computed tomography (CT). A 69-year-old woman was admitted to our hospital for treatment of pneumonia. Thoracic CT showed a crazy-paving sign in the right lung field on admission. She received ceftriaxone and clarithromycin, and the symptoms and infiltration shadow promptly disappeared. Serologic testing revealed a greater than 4-fold increase in the IgG titer for parainfluenza virus I. To our knowledge, there is no previous report of the crazy-paving sign in associated with viral pneumonia in a non-immunocompromised host or with parainfluenza pneumonia.

Figure 5.4 Comparison of BIGBIRD summaries with Ground Truth Summaries

Abstract Generated by BIGBIRD (Standalone)	Ground Truth Summary
The aim of this study was to evaluate the utility of a multiplex polymerase chain reaction assay for detection of 15 common bacterial, fungal and viral gastrointestinal pathogens in formalin fixed, paraffin embedded ileum tissue specimens from infants with acquired enterocolitis. In total, 37 paraffin embedded ileum tissue samples were analyzed for 15 common bacterial, fungal and viral gastrointestinal pathogens by using a multiplex PCR assay. The ileum samples were from 22 patients with acquired enterocolitis, 14 patients with non acquired enterocolitis and one patient with clinical confirmed gastrointestinal disease. The samples were tested positive for human beta actinV and for 15 bacterial toxins including norovirus, rotavirus, adenovirus, and enterovirus. The samples from non acquired enterocolitis were negative for any of the following pathogens norovirus, rotavirus, adenovirus, and enterovirus. In contrast, none of the ileum samples from patients with acquired enterocolitis were positive for any of the 15 bacterial pathogens. In conclusion, this is the largest collection of ileum tissue samples tested for gastrointestinal pathogens by this method. The assay should be tested in cases of acquired enterocolitis that occurred simultaneously with epidemic bacterial or viral intestinal	Necrotizing enterocolitis (NEC) is the one of the most common gastrointestinal emergencies in premature infants and has been linked with viral antigens for as much as 40% of cases in single center cohorts. We examined 28 tissue sections from surgically resected ileum from 27 preterm infants with NEC from 2 separate institutions for 15 common bacterial, viral, and parasitic gastrointestinal pathogens using multiplex RT-PCR amplification and suspension array detection methods. We did not detect infectious enteritis pathogens in any of the NEC tissues and conclude that gastrointestinal pathogens are a rare cause of NEC.

Figure 5.5 Comparison of BIGBIRD summaries with Ground Truth Summaries

5.3.2 Evaluation of Ensemble with Extraction Layer followed by BIGBIRD

As part of the ensemble, model is trained on the below sections together:

- Introduction
- Extractive Summary
- Results section

The Extractive summarization is the first layer which is implemented with BERT and K-medoid clustering. The extraction % implemented in the architecture is 40-50% of the entire document. The Extracted summary along with Introduction and Result sections becomes the input to BIGBIRD model. The model is trained on 5 epochs. Predictions are done on the trained model and evaluation is done using ROUGE scores. Below are the model validation results extracted from the aggregated score which is obtained from rouge metric library of Huggingface. The highest aggregated scores obtained on the test data for the 3 categories of document size are shown in table 5.4, table 5.5 and table 5.6:

- Short Articles (word count<1500)

Table 5.4 Evaluation Results- Short Articles

	Precision	Recall	f-measure
ROUGE-1	48.86%	43.43%	30.86%
ROUGE-2	24.13%	21.51%	14.63%
ROUGE-L	32.97%	33.90%	21.26%

- Mid-sized articles (1500<word count<4000)

Table 5.5 Evaluation Results- Medium length Articles

	Precision	Recall	f-measure
ROUGE-1	53.33%	49.24%	39.76%
ROUGE-2	21.68%	18.32%	19.17%
ROUGE-L	32.15%	34.84%	28.46%

- Lengthy articles (wordcount > 4000)

  Table 5.6 Evaluation Results- Lengthy Articles

	Precision	Recall	f-measure
ROUGE-1	66.66%	50.60%	54.27%
ROUGE-2	23.15%	20.00%	21.46%
ROUGE-L	41.33%	29.08%	31.19%

A few samples from the summaries generated by ensemble with Extraction Layer and BIGBIRD were manually evaluated and the results were not as good as that of BIGBIRD standalone model. The coverage of the context as compared to that of BIGBIRD was less in most of the cases. Figure 5.6 and 5.7 are the snapshots of generated summary and ground truth summary pairs from the sample.

Abstract Generated by Extraction Layer + BIGBIRD	Ground Truth Summary
The crazy paving sign is an area of scattered or diffuse ground glass attenuation with superimposed interlobular septal thickening and intralobular lines. The crazy paving sign has a variety of causes, including infectious, neoplastic, idiopathic, inhalational, and hemorrhagic disorders. In this report, we describe the crazy paving sign in a 69 year old woman with parainfluenza virus pneumonia in a non immunocompromised host.	The crazy-paving sign is the appearance of a smooth linear pattern superimposed on an area of ground-glass opacity on thin-section computed tomography (CT). A 69-year-old woman was admitted to our hospital for treatment of pneumonia. Thoracic CT showed a crazy-paving sign in the right lung field on admission. She received ceftriaxone and clarithromycin, and the symptoms and infiltration shadow promptly disappeared. Serologic testing revealed a greater than 4-fold increase in the IgG titer for parainfluenza virus I. To our knowledge, there is no previous report of the crazy-paving sign in associated with viral pneumonia in a non-immunocompromised host or with parainfluenza pneumonia.

Figure 5.6 Comparison of BIGBIRD summaries with Ground Truth Summaries

Abstract Generated by Extraction Layer + BIGBIRD	Ground Truth Summary
Abstract A child may develop an opaque pulmonary consolidation with unusually round shape, which raises the concern of a tumor in the chest, causing anxiety to the pediatrician and parents. We report the case of a child with so called round pneumonia, whose initial chest computed tomography findings mimicked those of a lung mass, but subsequently showed complete response to oral antibiotics alone. The findings and recommendations in the literature are also reviewed.	Round pneumonia" or "spherical pneumonia" is a well-characterized clinical entity that seems to be less addressed by pediatricians in Taiwan. We herein report the case of a 7year-old boy who presented with prolonged fever, cough, and chest X-rays showing a welldemarcated round mass measuring 5.9 Â 5.6 Â 4.3 cm in the left lower lung field, findings which were typical for round pneumonia. The urinary pneumococcal antigen test was positive, and serum anti-Mycoplasma pneumoniae antibody titer measurement using a microparticle agglutination method was 1:160 (þ). After oral administration of antibiotics including azithromycin and amoxicillin/clavulanate, which was subsequently replaced by ceftibuten due to moderate diarrhea, the fever subsided 2 days later and the round patch had completely resolved on the 18th day after the diagnosis. Recent evidence suggests treating classical round pneumonia with antibiotics first and waiving unwarranted advanced imaging studies, while alternative etiologies such as abscesses, tuberculosis, nonbacterial infections, congenital malformations, or neoplasms should still be considered in patients with atypical features or poor treatment response.

Figure 5.7 Comparison of BIGBIRD summaries with Ground Truth Summaries

5.3.3 Comparison Between the Summarization Pipelines

For the purpose of comparison ROUGE-1 and ROUGE-2 metrics is used as reference. F-measure is the primary reference to evaluate the performance between the two summarization pipelines i.e., BIGBIRD and ensemble with Extractive Layer and BIGBIRD. The comparative view is represented in the table 5.7.

Table 5.7. BIGBIRD Standalone and BIGBIRD Ensemble with Extraction Layer Comparison View

Ensembles	Document Size	ROUGE-1			ROUGE-2		
		Precision	Recall	f-measure	Precision	Recall	f-measure
BIGBIRD	Short	71.83%	63.99%	61.55%	57.03%	51.19%	51.05%
Ensemble with Extraction Layer + BIGBIRD		48.86%	43.43%	30.86%	24.13%	21.51%	14.63%
BIGBIRD	Medium	56.41%	46.21%	50.76%	31.33%	25.61%	28.18%
Ensemble with Extraction Layer + BIGBIRD		53.33%	49.24%	39.76%	21.68%	18.32%	19.17%
BIGBIRD	Long	59.90%	49.40%	54.14%	26.21%	21.60%	23.68%
Ensemble with Extraction Layer + BIGBIRD		66.66%	50.60%	54.27%	23.15%	20.00%	21.46%

5.3.3.1 Interpreting Comparison Results

The comparison of ROUGE -1 and ROUGE-2 scores between the two models for the same dataset shows that BIGBIRD standalone has outperformed the BIGBIRD extractive ensemble for short and medium sized CORD-19 articles. The F-measure of ROUGE-1 and ROUGE-2 is high which shows that BIGBIRD was able to absorb more contexts from the inputs documents as compared to the ensemble.

However, it is interesting to see that the performance of the BIGBIRD with extraction layer improves as the document size increases beyond 4000 words. The ROUGE-2 score of BIGBIRD standalone is still higher than that of Ensemble with the extraction layer.

5.3.3.2 Comparison of Summaries generated

Below are the few snapshots of content comparison between the summaries generated by BIGBIRD standalone and ensembled BIGBIRD with extraction layer. The Extraction layer shows better results in few cases for longer documents as shown in figure 5.8, figure 5.9 and figure 5.10

Abstract Generated by BIGBIRD Standalone	Abstract Generated (Extraction Layer+ BIGBIRD)
We report the case of a 7year old boy who presented with prolonged fever, cough, and chest rays showing a welldemarcated round mass measuring 5.9 5.6 4.3 cm in the left lower lung field, findings which were typical for round pneumonia. The urinary pneumococcal antigen test was positive, and serum anti SARScoplasma pneumoniae antibody titer measurement using a microparticle agglutination method was 1 160. The oral administration of antibiotics including azithromycin and, which was subsequently replaced by ceftibuten due to moderate diarrhea, the fever subsided 2 days later and the round patch had completely resolved on the 18th day after the diagnosis. This evidence suggests treating classical round pneumonia with antibiotics first and waiving unwarranted advanced imaging studies, while alternative etiologies such as abscesses, tuberculosis, nonbacterial infections, congenital malformations, or neoplasms should still be considered in patients with atypical features or poor treatment response.	We report the case of a 7year old boy who presented with prolonged fever, cough, and chest rays showing a welldemarcated round mass measuring 5.9 5.6 4.3 cm in the left lower lung field, findings which were typical for round pneumonia. The urinary pneumococcal antigen test was positive, and serum antiVcoplasma pneumoniae antibody titer measurement using a microparticle agglutination method was 1 160. The patient was treated with a combination of antibiotics including azithromycin and, which was subsequently replaced by ceftibuten due to moderate diarrhea, the fever subsided 2 days later and the round patch had completely resolved on the 18th day after the diagnosis. The typical clinical presentation and radiographic appearance, misdiagnosis of round pneumonia is unlikely, and unwarranted additional imaging should be avoided.

Figure 5.8 Comparison of Summary from the two BIGBIRD ensembles

Abstract Generated by BIGBIRD Standalone	Abstract Generated (Extraction Layer+ BIGBIRD)	Input length (words)
A hydrophobically modified chitosan derivative, 3 trimethylammonium chitosan chloride, and its hydrophobically modified counterpart, 3 trimethylammonium chitosan chloride, were evaluated for their antiviral activity against murine hepatitis virus, human coronavirus, and human metapneumovirus. We found that the minimal inhibitory concentration value of 3 trimethylammonium chitosan chloride and 3 trimethylammonium chitosan chloride was 10, and 50, respectively, for murine hepatitis virus and human coronavirus. In addition, 3 trimethylammonium chitosan chloride and 3 trimethylammonium chitosan chloride inhibited the replication of human coronavirus and human metapneumovirus in vitro, respectively. In addition, 3 trimethylammonium chitosan chloride and 3 trimethylammonium chitosan chloride inhibited the replication of human coronavirus and human metapneumovirus in vitro, respectively. The antiviral activity of 3 trimethylammonium chitosan chloride and 3 trimethylammonium chitosan chloride against murine hepatitis virus suggests that these compounds may represent a novel class of antiviral compounds for the treatment of a wider spectrum of coronaviral diseases.	The recent emergence of two new human coronaviruses has brought the whole family of coronaviruses back to the limelight. In this paper, a cationically modified chitosan derivative and its hydrophobically modified derivative are shown to be potent inhibitors of IVVV63 replication. The polymers also showed a prominent activity against murine hepatitis virus, suggesting that developed compounds may represent a novel class of antiviral compounds for the treatment of a wider spectrum of coronaviral diseases.	4910

Figure 5.9 Comparison of Summary from the two BIGBIRD ensembles

Abstract Generated by BIGBIRD Standalone	Abstract Generated (Extraction Layer+ BIGBIRD)	Input length (words)
A ribosomal frameshifting mechanism has been described in the avian coronavirus infectious bronchitis virus, which encodes components of the virus specific reverse transcriptase polymerase. In vitro transcription and translation, we demonstrated that a ribosomal frameshift signal is produced as a result of a highly efficient ribosomal frameshift that suppresses the I termination codon. In this study, we define the minimal amount of information from the I2 overlap that is sufficient to induce frameshifting at a high efficiency, and we present evidence that efficient frameshifting depends on the formation, by these sequences, of a tertiary structure in the form of a pseudoknot.	The ribosomal frameshifting signal of the avian coronavirus infectious bronchitis virus has been identified by a combination of site directed mutagenesis and amino acid sequence analysis frameshifting. The site at which frameshifting occurs has been defined by a combination of site directed mutagenesis and amino acid sequence analysis frameshifting. The essential region of the frameshift signal consists of approximately 80 nucleotides downstream of the slippery sequence, and mutational analysis of this region has identified a pseudoknot in the form of a pseudoknot. In the case of mutants where the first predicted stem was allowed to form but not the second, frameshifting was markedly reduced, although at a much lower level. In the case of mutants where the first predicted stem was allowed to form but not the second, frameshifting was markedly reduced, although at a much lower level. In the case of mutants where the first predicted stem was allowed to form but not the second, frameshifting was markedly reduced, although at a much lower level. In the case of mutants where the first predicted stem was allowed to form but not the second, frameshifting was markedly reduced, although at a much lower level. In the case of mutants where the first predicted stem was allowed to form but not the second, frameshifting was markedly reduced.	5176

Figure 5.10 Comparison of Summary from the two BIGBIRD ensembles

5.4 Comparison with other Transformers

As the last lap of comparison, BIGBIRD was compared with one of the best performing transformers i.e., BART. As there is a huge gap in processing capacity of input tokens between BART and BIGBIRD, to establish a common ground for comparison, both BART and BIGBIRD were trained on only two sections i.e., Introduction and Results of CORD-19 dataset.

The models were evaluated based on ROUGE scores and comparisons were drawn. As BIGBIRD takes a lot of time to train, BIGBIRD was trained for only 1 epoch and BART was trained till 3 epochs. Below is the ROUGE score comparison for BART and BIGBIRD on a common dataset. As seen in the table 5.8, BIGBIRD scores higher than BART. The f-measure is higher for both ROUGE-1 and ROUGE-2.

Table 5.8 BIGBIRD and BART comparison view on conditioned dataset

MODEL	Trained on Sections	ROUGE-1			ROUGE-2		
		Precision	Recall	F1	Precision	Recall	F1
BIGBIRD	Introduction + Result	46.74%	40.76%	41.11%	18.37%	16.00%	16.10%
BART	Introduction + Result	50.13%	39.33%	40.91%	18.31%	14.11%	14.80%

5.5 Meaningful Insights

- BIGBIRD performs better for CORD-19 dataset than most of the other models.
- The ensemble strategy involving extractive summary followed by Abstractive summarization of BIGBIRD is not useful in all scenarios. This strategy can be

useful for very long documents i.e., when document size exceeds 4500 wordcount.

- The BIGBIRD with Extractive Summary layer performs poorly for COVID documents that have less than 4000 wordcount.
- BIGBIRD alone has the capability perform better in most of scenarios.

5.6 Summary

The evaluation of BIGBIRD on CORD-19 dataset proves the suitability of the model to address the summarization challenges in biomedical domain. This also gives an insight about BIGBIRD's capability to perform independently without an ensemble strategy. BIGBIRD when trained to a greater number of epochs can perform even be

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

Abstractive summarization has always been a challenge for longer documents in biomedical domain. CORD-19 dataset is an example where the research articles are lengthy and information is spread across the entire document. The existing strategies do not address the gaps of such domain specific corpus as they cannot process large input sequences and there is always a possibility of losing the important context in summarization. This research confirms the hypothesis that BIGBIRD stands out from the existing models in summarization of research articles because of its sparse based attention mechanism and its capacity to process input sequences. This section walkthroughs the findings and conclusions of experiments done with the BIGBIRD ensembles followed by future recommendations.

6.2 Discussion and Conclusions

The experiments detailed in the chapters 4 and 5 have confirmed the hypothesis that BIGBIRD is an efficient model for summarization of COVID-19 research articles. The first part of the experiment had BIGBIRD as the only model in the summarization pipeline. The highest f- measures of ROUGE-1 and ROUGE-2 obtained were 61.55% and 51.05% respectively. These scores are strong indicators that the quality of summary generated by the model is high and is better than the previous work done with GPT-2 (Tan et al., 2020). To further confirm the findings, the contents of the ground truth and model generated summaries were compared manually. The comparison further backed up the ROUGE score results as the context coverage was almost same as that of the ground truth. Moreover, in few records it was observed that generated summary had better context coverage as compared to the ground truth. It is also observed that the generated summaries are not up to the mark for research articles that are dominated by numerical and statistical contents.

The second part of the experiment where BIGBIRD was ensembled with an extraction layer, the results were not as good as that of the first experiment. The f-measures of ROUGE-1 and ROUGE-2 decreased for the same test data. However, there was pattern in the declining trend. For documents with word length below 4000, f-measures of ROUGE-1 and ROUGE-2 decreased to 39.76% and 19.17% respectively. The ROUGE

scores started improving when document length exceeded 4000-word length and as the document size reached 5000-word length it became comparable to that of BIGBIRD model from first experiment. This is opposite to the first experiment where the model outperforms with short documents yielding maximum rouge scores and rouge scores starts dropping as the article length increases above 4000 words. However, the decreased score is still higher than that of second model.

The conclusion of the research work done is that BIGBIRD without the extraction layer performs better for scientific research articles such as CORD-19 dataset. This model can perform well for research articles of size up to 5000 words and conditioning with extractive layer is limiting the model's capacity to learn context. As 80% of articles in the CORD-19 corpus fall within the 5000-word count, BIGBIRD can make a significant difference in the research community.

6.3 Contribution to knowledge

This research is intended to help the medical research community to keep up with the rapidly growing coronavirus literature and draw insights in a short turnaround time to fight the pandemic. This research has proved the potential of BIGBIRD for difficult NLP tasks such as summarization especially in research community. This study has proved that BIGBIRD can perform independently and still outperform the contemporary transformers in the biomedical domain. This also sets the stage for leveraging BIGBIRD's capacity for other NLP tasks in the current coronavirus pandemic such as Q&A and information retrieval system.

6.4 Future Recommendations

The BIGBIRD approach can be boosted further if more intensive computation resources are available. The existing implementation is limited due to computation power. BIGBIRD when trained to a greater number of epochs can further improve the performance

The evaluation of abstract summary requires more exploration as ROUGE scores do not always reflect the quality of generated summaries. In a few samples it was observed that the quality of summary was better and more precise than ground truth summary, still the ROUGE score was not able to justify it. Ground truth summaries have limitations and need not always be the base for comparison especially for models such as

BIGBIRD, which can process larger contexts from the document and may generate a more succinct summary.

The BIGBIRD ensemble with extraction layer can be leveraged for documents that are excessively long. The extraction layer can reduce the document to 5000 words and this can be used as input to BIGBIRD summarization model. More experiments are needed to ascertain the performance.

In the end, it is hoped that the text summarization approach evaluated in this research can help medical community in their research activities to come up with faster insights to not only fight the ongoing pandemic but also in prevention of another one in future.

REFERENCES

Alberti, C., Cvicek, V., Ainslie, J., Onta, S., Fisher, Z., Pham, P., Ravula, A., Sanghai, S., Wang, Q. and Yang, L., (2020) ETC: Encoding Long and Structured Inputs in Transformers ˜.

Anon (n.d.) ai.googleblog.

Anon (n.d.) bigbirdpegasus-large-pubmed. [online] Available at: https://huggingface.co/google/bigbird-pegasus-large-pubmed.

Anon (n.d.) Kaggle CORD-19 challenge. [online] Available at: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

Beltagy, I., Peters, M.E. and Cohan, A., (2020) *Longformer: The Long-Document Transformer*. *arXiv*.

Clark, K., Luong, M.T., Le, Q. V. and Manning, C.D., (2020) *Electra: Pre-training text encoders as discriminators rather than generators*. *arXiv*.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp.4171–4186.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M. and Hon, H.W., (2019) Unified Language Model Pre-training for Natural Language Understanding and Generation. *arXiv*, NeurIPS.

Ibrahim Altmami, N. and El Bachir Menai, M., (2020) *Automatic summarization of scientific articles: A survey*. *Journal of King Saud University - Computer and Information Sciences*, Available at: https://doi.org/10.1016/j.jksuci.2020.04.020.

K, B., P C, R. and Murali, R., (2019) Automatic Text Summarizing System Using Reinforcement Learning Technique. *SSRN Electronic Journal*.

Koupaee, M., (2018) Abstractive Text Summarization Using Hierarchical Reinforcement Learning. pp.1920–1949.

Le, H.T. and Le, T.M., (2013) An approach to abstractive text summarization. *2013 International Conference on Soft Computing and Pattern Recognition, SoCPaR 2013*, pp.371–376.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L., (2019) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv*.

Li, H., Zhu, J., Zhang, J., Zong, C. and He, X., (2020) Keywords-Guided Abstractive Sentence Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3405, pp.8196–8203.

Lin, C. and Rey, M., (2001) ROUGE : A Package for Automatic Evaluation of Summaries.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z. and Tang, J., (2021) GPT Understands, Too. [online] Available at: http://arxiv.org/abs/2103.10385.

Liu, Y., (2019) *Fine-tune BERT for Extractive Summarization*. *arXiv*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., (2019) RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, 1.

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D., Weld, D.S., Etzioni, O. and Kohlmeier, S., (2020) CORD-19: The Covid-19 Open Research Dataset. *ArXiv*. [online] Available at: http://www.ncbi.nlm.nih.gov/pubmed/32510522%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7251955.

Miller, D., (2019) Leveraging BERT for Extractive Text Summarization on Lectures. [online] Available at: http://arxiv.org/abs/1906.04165.

Moawad, I.F. and Aref, M., (2012) Semantic graph reduction approach for abstractive Text Summarization. *Proceedings - ICCES 2012: 2012 International Conference on Computer Engineering and Systems*, pp.132–138.

Moratanch, N. and Chitrakala, S., (2016) A survey on abstractive text summarization. *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2016*, November 2017.

Moratanch, N. and Chitrakala, S., (2017) A survey on extractive text summarization. *International Conference on Computer, Communication, and Signal Processing: Special Focus on IoT, ICCCSP 2017*, January.

Narayan, S., Cohen, S.B. and Lapata, M., (2018) Ranking sentences for extractive summarization with reinforcement learning. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, pp.1747–1759.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J., (2001) BLEU: a method for

automatic evaluation of machine translation. *ACL*, pp.311–318.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., (2018) Improving Language Understanding by. *OpenAI*, [online] pp.1–10. Available at: https://gluebenchmark.com/leaderboard%0Ahttps://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Raffel, C., Roberts, A. and Liu, P.J., (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 21, pp.1–67.

Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P., (2015) SQuAD: 100,000+ Questions for Machine Comprehension of Text. ii.

Sahoo, D., Bhoi, A. and Balabantaray, R.C., (2018) Hybrid Approach to Abstractive Summarization. *Procedia Computer Science*, [online] 132Iccids, pp.1228–1237. Available at: https://doi.org/10.1016/j.procs.2018.05.038.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T., (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. [online] pp.2–6. Available at: http://arxiv.org/abs/1910.01108.

Scialom, T., Sylvain, P.D., Benjamin, L. and Jacopo, P., (2020) Discriminative Adversarial Search for Abstractive Summarization.

Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.Y., (2019) MASS: Masked sequence to sequence pre-training for language generation. *36th International Conference on Machine Learning, ICML 2019*, 2019-June, pp.10384–10394.

Subramanian, S., Li, R., Pilault, J. and Pal, C., (2019) On extractive and abstractive neural document summarization with transformer language models. *arXiv*.

Tan, B., Kieuvongngam, V. and Niu, Y., (2020) *Automatic text summarization of COVID-19 medical research articles using BERT and GPT-2*. *arXiv*.

Vaswani, A., Noam Shazeer, Niki Parmar and Jakob Uszkoreit∗, (2002) The Transformer-Attention Is All You Need. *IEEE Industry Applications Magazine*, 81, pp.8–15.

Vladislav, T. and Denis, S., (2020) *Combination of abstractive and extractive approaches for summarization of long scientific texts*. *arXiv*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R., (2018) *GLUE: A multi-task benchmark and analysis platform for natural language understanding*. *arXiv*.

Xiao, W. and Carenini, G., (2020) Extractive summarization of long documents by

combining global and local context. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp.3011–3021.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q. V., (2019) XLNet: Generalized autoregressive pretraining for language understanding. *arXiv*, NeurIPS, pp.1–18.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L. and Ahmed, A., (2020) Big Bird: Transformers for Longer Sequences. *arXiv*, NeurIPS.

Zhang, J., Zhao, Y., Saleh, M. and Liu, P.J., (2019) *PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization*. *arXiv*.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X. and Huang, X., (2020) Extractive summarization as text matching. *arXiv*.

Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M. and Zhao, T., (2018) Neural document summarization by jointly learning to score and select sentences. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1, pp.654–663.

APPENDIX A: RESEARCH PLAN

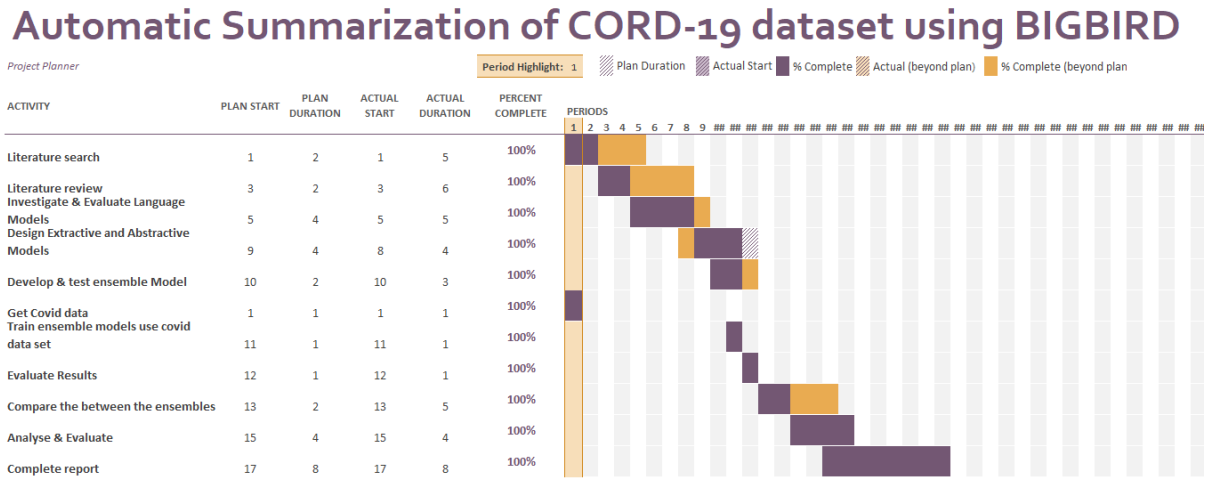

Figure 1. 1 Research Timeline Milestone

APPENDIX B: RESEARCH PROPOSAL

Abstract

The aim of this research work is to explore and build a method to generate abstract summaries of Covid data set by leveraging both Extractive and Abstractive strategies, the benefits of which can eventually be extended to Journals and articles datasets of healthcare domain. The intent here is to leverage Keywords generation, followed by extraction summary step which can be used to condition the abstractor module. Hence the intent here is to leverage and combine the existing Language Models for an ensemble and then design an approach to innovate. The intent here is to improve summarization and achieve good performance metrics (ROUGE Scores)

1. Introduction

Automatic summarization of lengthy documents using Machine Learning e.g. Legal data sets, Science / Medical Journals etc., has gained popularity since few years as the concept of chunking and filtering only the important information without losing the context from exhaustive factual documents not only saves time but also contributes to an improved understandability for target audience. Extractive Summarization and Abstractive Summarization are two existing approaches that has been implemented using Machine Learning and has found to be doing as extraction and abstraction concepts are subjective to domain at times, this always creates a scope of improvisation. Extractive summary retains the most important i.e., those sentences or phrases from the original text that are conveying the context of the original texts. In Abstract summarization new sentences are generated from the original text and at the same time preserving the context. Text Summarization is an important feature in Medical domain where there is a need for summarization long scientific journals, articles, and related texts for research purpose. Leading Research Groups in collaboration with White House have prepared CORD-19 i.e., COVID-19 Open Research Dataset and it has been made public for the research communities across the world. This data set contains over 4000,000 researched scholarly pandemic articles including over 150000 full length articles on coronaviruses including COVID-19 and SARS-CoV-2. This content can be leveraged by the Research communities across the world to synthesize new insights to battle the ongoing pandemic and help in insulating from the pandemics that may come up in the future. By leveraging Natural Language Processing and other AI techniques on this data, a lot of useful work can be done one of the ways to achieve this is by performing efficient Text

Summarization. An efficient Text summarization here will not only save time for the researchers in going through the exhaustive content but also leverage them to get hold of the context and help them draw new insights of each article which otherwise can be lost in exhaustive reading of the original content.

2. Background and related research

Text Summarization attempts to condense long texts without losing the context and at the same time preserving the important information. Initial works in areas of text summarization focused on extractive techniques aiming to retain the most important sentences from the documents.

Abstractive approach on the other hand involves summarizing the whole context in a more condensed structure drastically reducing the length of the content. Obviously abstractive summarization is a challenging task as generating creating abstractive demands command on the domain and as well as the natural language which is a tedious task for the machine.

Scientific document summarization is a special case of Summarization as characteristics of scientific papers – length, writing styles, scientific terms and discourse structure demands an exclusive model consideration to maintain the context and at the same time retaining the accuracy of the topic. Researchers have engineered different approaches to address the challenges in Scientific document Summarization. The metamorphosis of Long Short-Term Memory networks to attention mechanism combined with sequence-to-sequence framework was a pivotal in improvising language modelling tasks. Introduction of transformer architecture coupled with novel self-attention mechanism was a significant leap in language modeling task.

Extraction:

In a recent work Ming Zhong et al. (2020) proposes "Extractive Summarization as Text Matching", a novel summary framework which scores and extracts sentences one by one to form a summary, a strategy to formulate extractive summarization in form of semantic text matching problem. In this Siamese network structure and basic BERT have been combined to form Siamese-BERT architecture to compute the similarity between the source document and the candidate summary. Wen Xiao et al. (2019) approaches extractive summarization for lengthy structured content by leveraging both local and global context from the entire document. This approach is inspired by natural topic-oriented structure of long documents which are created using human

intelligence, where the binary conclusion of whether the sentence should be part of the summary is dependent on the sentence itself, the entire document and the current topic. The representation of document is cascading of the last ‘n’ hidden states of the forward and backward RNNs, while the representation of topic segment is done by leveraging LSTM-Minus method.

Abstraction:

There have been a number of research work that illustrates improvised Abstractive Summarization. In a recent work (Zhang et al., 2019) proposes " Pre-training with Gap-Sentences for Abstractive Summarization i.e., PEGASUS" model, in which the key sentences are removed or masked from an input source and are generated collectively as single output sequence from the remaining sentences, this approach is similar to an extractive summary. In another work "Discriminative Adversarial Search for Abstractive Summarization" (Scialom et al., 2020) usage of Beam search is proposed which is de-facto algorithm used to decode generated sequences of text. Beam search has led to performance improvements of State of Art models Q&A Generation, Text Summarization and Neural Machine Translation. In one of the recent successful models "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension” by (Mike Lewis et al., 2019), it performs pretraining of sequence-to -sequence models by denoising autoencoder. Training of BART is done by corrupting text with an arbitrary noising function and Learning a model is made to reconstruct the original text. BART is one of the best performing transformers as it generalizes BERT, GPT and many other most pre-training schemes.

Enhancing Abstraction using Extraction:

In a recent work by Sandeep Subramanian and et al. (2020) titled "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models", author proposes combining extraction and abstraction strategies to come up with a more effective abstract summarization. The author has used encoder decoder architecture for Extractive summary where a sentence encoder is implemented using bi-directional LSTM and the decoder is implemented as autoregressive LSTM taking the sentence-level LSTM hidden state of the previously extracted sentence as input and predicting the next extracted sentence. For abstraction they have trained a Transfer Language Models with -220M parameters with 20 layers, 768 dimensional embeddings, 3072-dimensional position-wise MLPs and 12 attention heads. This model is trained on 4

components: Introduction, extracted sentences, abstract and rest of the paper. For creating abstracts of a long document, the trained model uses "Introduction" of the document as proxy to contain enough information for abstract along with the extracted content from extraction model.

In another similar work by Tretyak Vladislav and et al. (2020) the Combination of abstractive and extractive approaches have been explicitly explored for text summarization of long scientific texts", the author proposes combining extractive and abstractive summary strategies with the usage of pre-trained Transformer Models. Extractive Model is trained as a classification to generating abstractive summary. They have experimented on 3 different architectures BERT, RoBERTa, and ELECTRA for extractive Summary. Word piece tokenizer is used where BERT and ELECTRA are experimented. In the process they have added special tokens like [math], [graph], [table], [equation] extracted using regex. Rouge Score is calculated using extracted summary and Ground Truth Abstract. As an outcome of the experiment BERT exhibited the highest performance Metrics as compared to other models. Extractive Summarization output was further fed into abstractive summarization model. The Abstract Summarization various pre-trained autoregressive language models were experimented out of which GPT-2 and BART were found to perform better as compared to other models. Combining BERT with BART and conditioning was done on input combination of introduction and conclusion (from the original document) along with the extractive summary (derived from the extraction model). The experiment resulted in the best ROUGE score.

In an attempt to combine extractive and abstractive summary another improvisation that has been experimented is Keywords-Guided Abstractive Sentence Summarization by Haoran Li et al. (2020) where he proposes extracting overlapping words between the input and the reference as the ground-truth keywords followed by Multi-task learning i.e., generating summary using the input sentence and the ground-truth keywords. Keywords are generated using the trained keywords extractor for the input sentence in the training set and then fine-tuning the sentence summarizer using the original sentence and the predicted keywords.

During testing, first keywords are generated using the trained keywords extractor for the input sentence and then the summary is produced using the input sentence and the predicted keywords. Similar Keyword based extraction followed by abstraction is proposed by Bowen Tan et al. in his work related to COVID-19 Medical Research dataset's abstract text summarization. He proposes a model where initially source text is scanned to extract keywords using token classification tools such as part of speech tagging packages of NLTK, or part of speech tagging of fine-tuned BERT token classifier. The extracted keywords are categorized into nouns, verbs and noun and verbs. Subsequently the keywords are paired with the gold summary abstract and model is processed using GPT-2. extraction and abstraction strategies to come up with a more effective abstract summarization. The author has used encoder decoder architecture for Extractive summary where a sentence encoder is implemented using bi-directional LSTM and the decoder is implemented as autoregressive LSTM taking the sentence-level LSTM hidden state of the previously extracted sentence as input and predicting the next extracted sentence. For abstraction they have trained a Transfer Language Models with -220M parameters with 20 layers, 768 dimensional embeddings, 3072-dimensional position-wise MLPs and 12 attention heads. This model is trained on 4 components: Introduction, extracted sentences, abstract and rest of the paper. For creating abstracts of a long document, the trained model uses "Introduction" of the document as proxy to contain enough information for abstract along with the extracted content from extraction model. For a smaller document the introduction is the entire document. The combination has given a better rouge score.

In another similar work by Tretyak Vladislav and et al. (2020) the Combination of abstractive and extractive approaches have been explicitly explored for text summarization of long scientific texts", the author proposes combining extractive and abstractive summary strategies with the usage of pre-trained Transformer Models. Extractive Model is trained as a classification to generating abstractive summary. They have experimented on 3 different architectures BERT, RoBERTa, and ELECTRA for extractive Summary. Word piece tokenizer is used where BERT and ELECTRA are experimented. In the process they have added special tokens like [math], [graph], [table], [equation] extracted using regex. Rouge Score is calculated using extracted summary and Ground Truth Abstract. As an outcome of the experiment BERT exhibited the highest performance Metrics as compared to other models. Extractive Summarization output was further fed into abstractive summarization model. The Abstract Summarization various pre-trained autoregressive language models were experimented out of which GPT-2 and BART were found to perform better as compared to other models. Combining BERT with BART and conditioning was done on input combination of introduction and conclusion (from the original document) along with the extractive summary (derived from the extraction model). The experiment resulted in the best ROUGE score.

3. Research Questions (If any)

The following research questions are suggested for each of the research objective as highlighted as follows.

1. Can leveraging Keyword's extractions to filter the important keywords relevant to the domain improve the extractive summary, which can then be conditioned for a better abstraction module (which will also include the rest of the document)?

2. Does the proposed ensemble approach work well for data sets of other domains too? (Cross Data set experiments)

3. Which transformer models and word embeddings techniques contributes to the best performance metrics?

4. Aim and Objectives

The main aim of this research is to propose a generate improved abstract summaries of COVID-19 data set (Scientific document) by leveraging both Extractive and Abstractive strategies. In the wake of recent pandemic, the need for robust research from large volume of data within a short time span has become inevitable. Extractive and Abstractive summarization approaches through recent advances in Open AI and NLP, can leverage processing and retrieval of comprehensive information from healthcare domain within a short time frame. This research attempts to explore the abstract summary competency by leveraging keywords extraction followed by combining of both extractive and abstractive approaches to perform automated text summarizing comprehensive using the existing language models

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

- To explore the ensemble of keywords extractions, extractive and abstractive approaches
- To identify a suitable word embeddings approach to optimize the result
- To identify the best Language model combination/ensemble for extraction and abstraction (e.g., BERT for extraction and BART for abstraction)
- To evaluate the performance of based on available metrics i.e., rouge score to assess the performance of the abstracted result with that of the conventional abstraction technique.

5. Research Methodology

Leading Research Groups in collaboration with White House have prepared CORD-19 i.e., COVID-19 Open Research Dataset and it has been made public for the research communities across the world. This data set contains over 4000,000 researched scholarly pandemic articles including over 150000 full length articles on coronaviruses including COVID-19 and SARS-CoV-2. This content can be leveraged by the Research communities across the world to synthesize new insights to battle the ongoing pandemic and help in insulating from the pandemics that may come up in the future. By leveraging Natural Language Processing and other AI techniques on this data, a lot of useful work can be done one of the ways to achieve this is by performing efficient Text Summarization.

An efficient Text summarization here will not only save time for the researchers in going through the exhaustive content but also leverage them to get hold of the context and help them draw new insights of each article which otherwise can be lost in exhaustive reading of the original content.

The proposition of the research here is that a model, leveraging Keyword extraction followed by extractive approach and abstractive approach is built. Experiments using pre-trained transformer models will be performed in the research. Initially an efficient keyword extractor model will be trained that can identify the important keywords of COVID data or scientific keywords. Based on these keywords extractive summary model will be trained to perform effective extraction summary. For the abstractive summarization task, the best extraction model will be taken and inferred on the original data set. ROUGE metrics will be used as metrics for performance of summary evaluation.

6. Expected Outcomes

The outcome expected are as below:

1. An optimized ensemble combining extraction and abstraction model with a good rouge score and score should be better than most of the conventional summarization approach

2. Satisfactory Cross data set evaluation performance- The trained model should perform well for other scientific journal datasets and for data belonging to another domain as well.

7. Requirements / resources:

- • Cloud lab with GPUs

8. Research Plan

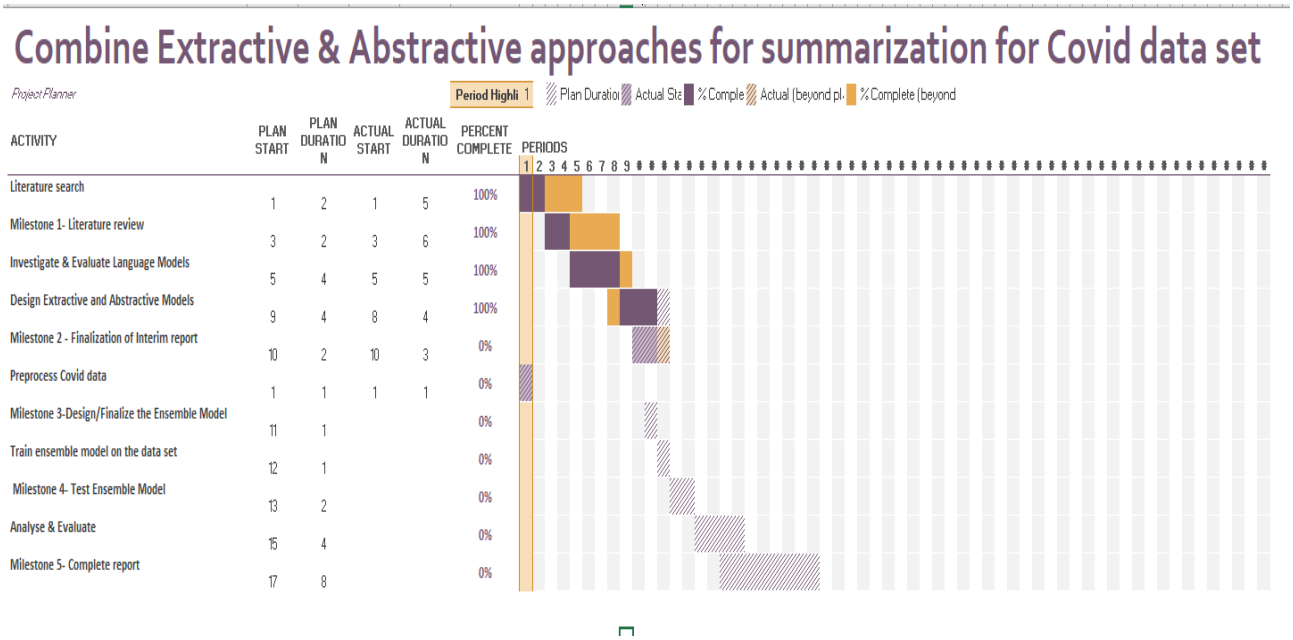

Figure 1. 2 Research timeline and milestone- RP