Text Summarization of COVID-19 Medical articles using DistilBERT

[*]Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—The medical community is under increasing pressure to keep up with the accelerated proliferation of new coronavirus-related publications in light of the COVID-19 pandemic. In order to close the gap between researchers and the continuously expanding body of publications, the COVID-19 Open Research Dataset Challenge has made a corpus of academic articles available. By performing text summarization on this dataset, we take advantage of the most recent developments in BERT and DistilBERT to address this difficulty. Using Relevancy scores, we assess the outcomes. Based on keywords retrieved from the source articles, our model gives extractive and complete information. The medical community can benefit from our work by having access to concise summaries of papers for which there isn't already an abstract available.**

***Index Terms*—Text Summarization, BERT, DistilBERT**

I. INTRODUCTION

A severe acute respiratory syndrome coronavirus2 outbreak began in December 2019 in Wuhan, Hubei Province, China. It afterwards spread to the rest of China and the rest of the world. The World Health Organization (WHO) designated the condition as Coronavirus Disease 2019 (COVID-19) in February 2020 [41].The WHO proclaimed COVID19 to be a pandemic when there were 200,000 confirmed cases and more than 8,000 fatalities in more than 160 countries [32].We had very few resources that could tell us the sources, effects and handling this crisis. But our researchers were on it and started digging more about the treatment of this dangerous virus. As the researcher increases the covid related analysis published number also increases. But one cannot go through all these papers for his solution. Hence, we came up with the summarization model which summarizes the COVID-19 related articles which can answer the questions by summarizing the abstracts of the papers.

The goal of automatic text summarization is to reduce long texts to shorter ones while maintaining the essential information [39]. There are two standard methods. The first type of summary aims to extract and concatenate key passages from the source text.The second method focuses on creating fresh summaries that quote the original text. It has been demonstrated that the extractive strategy preserves a respectable level of accuracy and grammatically. Contrarily, the abstractive technique is far more difficult because the model must be able to describe the original text's semantic content and then use that representation to produce a paraphrase. The model might, however, learn how to use words creatively or how to draw conclusions from the source material.

In this study, a BERT-based and DistilBERT-based extractive summarization is proposed. The extractive method identifies the relevant parts of the text related to the question from user and summarizies the abstracts of the related papers to come up with a summarized answer. The proposed method has three phases : Pre-processing of the raw text, Taking in the User question and retrieving the relevant papers from the dataset, Summarizing the abstracts of the papers related to the user given question. This pipeline/procedure helps to find a summarizied answer to the question while looking up for relevant research papers from the dataset.

In the following, we will first review some of recent articles in the field of text summarization using abstractive,extractive methods and then we will discuss articles related to text summarization using deep learning methods and through health records.

Identify applicable funding agency here. If none, delete this.

II. Related Work

A. Text Summarization Methods

1) Extractive Summarization: Tan et al. [17] employed pre-trained BERT and GPT2 to synthesize Corona-related article summaries. Unsupervised extractive summarization and abstractive summarization make up the model's two components. A pre-trained BERT model is used for the first half, and the GPT-2 model for the second. The sentences are transformed into sentence embedding in the first stage using a pretrained BERT model. The set of sentence embeddings is then subjected to a k-medoid clustering to produce a set of cluster centers. An extractive summary is formed from the preceding sentences. Then, from the extractive summary, a number of keywords are extracted using POS-tagger. The GPT2 model is given keyword-reference summary pairs, and upon training, system summaries are produced.

A recurrent neural network-based extractive summarization is suggested in this study. The extractive technique locates the text's informative passages. For evaluating sequences like text, recurrent neural networks are particularly effective. Sentence encoding, rating of sentences, and compilation of summaries are the three stages of the suggested methodology. An approach called coreference resolution is utilized to enhance the performance of the summarization system. Identifying mentions in the text that relate to the same thing outside the text is known as coreference resolution. Finding the main theme of the text through this technique aids in the summarizing process. [2]

Lakshmi Krishna et al. [26] compared pre-processing and model building for extractive summarization tht were carried out for a small number of documents. BERT, Text rank, and GPT2 algorithms have been used to clean the text as much as possible. The performance of the GPT-2 algorithm produced the best results out of these models. This study is entirely reliant on pre-trained models; hence, training the model may result in a higher ROUGE score via better encoding representations of nodes.

Awane Widad et al. [37] presented a Question Answering tool based on BERT fine-tuned on the SQuAD benchmark. This tool takes CORD-19 dataset and uses 'Anserini', an open-source information search toolbox built around Lucene. This tool retrieves the relevant paragraph to the given question. Then this relevant set of paragraphs are sent to the BERT Text summarizer and then we use BERT pre-trained on SQuAD for answering of the questions.

Abdullah Javaid Chaudhry et al. [7] explored if "Termolator", a tool for extracting characteristic terms for a domain can be used to enhance the performance of extractive summarization approaches. They have used multiple approaches such as modified versions of TF-IDF vectors, BERT, modifications pf Word2Vec,K-Means clustering, Lex Rank, and template summarization. Evaluated the models with ROUGE family of metrics and concluded that the usage of characteristic terms for a domain as found by "Termolator" improves the performance of extractive summarization approaches with regards to the F-score.

A model known as CAiRE-COVID has been presented by Su et al. [33]. The three major modules of CAiRE-COVID are information retrieval, question-and-answer, and summarization. After receiving a user query, the information retrieval module retrieves the top n most pertinent paragraphs. The most pertinent sentences found in the preceding stage are listed as the answer by the question-answering module. To choose the pertinent sentences from each of the n paragraphs as the responses to the question, the question-answering module is applied to each of the n paragraphs. The top k paragraphs are then specified after these n paragraphs are once more reranked in accordance with the highlighted replies. These k paragraphs are provided to the summarizer module, which then uses them to produce an extractive summary and an abstractive summary. The abstractive summary is produced using the UniLM and BART models, and the extractive one is produced using the cosine similarity of the sentences to the query.

Attempt to complete the task of summarization [30], which generate texts to describe the disease articles coherently and concisely and aslso similar to the task of extracting titles from abstracts. Generally speaking, the important words can represent main concept of this article. Their approach combines a conventional Seq2Seq model with attention mechanism and a typical key phrase extraction method. Instead of generating summarization by Seq2Seq model, they consider the importance of words in source text. The method makes it easier to output important words in result. The experiment was conducted on COVID-19 dataset from Kaggle and performs both automatic and human evaluations.

A Opidi et al. [23]Text summary and how it helps to overcome the challenge of doing so manually have been briefly explored in this blog. Because Abstract-based summary must solve NLP issues for potential outputs, they employed extraction-based summarization in this instance. The text summarization process in extract-based summarization is broken down into a few parts.Here, the most basic NTLK library was used to help eliminate the majority of terms by importing stopwords and porterstemmer to help remove words from their root forms. used modules for wordtokenization and sentence tokenization to separate the set of sentences. They have utilized the BeautifulSoup library, which will parse the page, and the urllib library, which will obtain the HTML content, to fetch the article. In order to avoid collisions, they found the weighted frequencies of the most frequently occurring terms in the paper using dictation. They determine a threshold for the sentences to avoid those with scores below average.

S JUGRAN et al. [15] The goal of this project is to provide an extractive approach based model as a solution for text summarization, starting with natural language processing as the core model. The extractive method genuinely succeeds in presenting the summary using the same set of terms that are actually the most crucial ones found in the original text or archive as a result it provides the necessary information. Sometimes NLTK library is tedious and time consuming process, so they used SpaCy library which is more efficient than NLTK.The efficiency of various strategies for separating them based on size and summary accuracy is then exhibited from here. These techniques aim to comprehend the text first, mark the words according to their significance, then choose the sentences that contain the most significant words and use those words or the words substituted for them to create the actual summary, decreasing the length of the text. All approaches follow the same workflow, which is text as the input, text processing as the intermediary step, and summary as the output. Here they used extractive approach to produce a summary which is in the ratio to text to the summary 2:1 or even better.

V Dalal et al. [9] The purpose of this paper is to examine the approaches and techniques that scholars utilize to automatically summarize texts. Bio-inspired text summarizing techniques are given particular focus.The abstractive approaches carry out a thorough linguistic analysis of the text and provide summaries that resemble summaries produced by humans. They therefore perform better than extractive techniques but cost more to compute. The ability to optimize is a well-known trait of bio-inspired algorithms. Combining bio-inspired methods with an abstractive approach could reduce computation costs while producing summaries that are comparable to those produced by humans.

A Sinha et al. [29] For the purpose of summarizing a single article, we provide a completely data-driven method employing feedforward neural networks. We use the standard DUC 2002 dataset to train and test the model, and the results are comparable to those of the most recent models. The suggested architecture is scalable and can output the summary of texts of any size by dividing the original document into portions of defined sizes and then feeding those pieces recursively to the network. Without having access to any linguistic data, we suggested and tested the model on standard datasets, which produced results that are on par with those of state-of-the-art models. We showed that a straightforward and very simple method can get outcomes comparable to sophisticated deep networks/sequence-based models.

L Gadasina et al. [11] In this study, they analyzed whether the coronavirus theme is unreliable information or whether it significantly affects the semantics of news stories. They extracted the most important details from news stories with a coronavirus subject. Their findings indicate that the presence of a popular unrelated broad topic has a detrimental effect on news headline extraction algorithms. It is required to remove the words most frequently connected with the dominant subject from the source texts in order to more accurately identify significant material during periods when there is a dominating theme. Otherwise, the summarization algorithms could generate results that are skewed and irrelevant to the demands of the users.Here An extractive approach-based TextRank algorithm was used. It is primarily employed for ranking web pages in search engine results. A link analysis technique called PageRank gives each component of a group of hyperlinked documents a numerical weighting to determine its relative value within the set. Any collection of objects with connected quotes and links can use the algorithm. The trust worthiness of web content was tested in experiments on extraction, and while the approach did not demonstrate top quality, the authors highlight that manual evaluation can be aided by it.

JW Park et al. [24] They proposed a brand-new BERT architecture that offers a concise but unique summary of large publications. The model adapts to the needs of the community by continuously learning on fresh data in an online setting while preventing catastrophic forgetting. The model provides a reliable synthesis of recent scientific literature, according to benchmark and manual evaluations of its performance.This feature can be useful for longer papers, which make up the majority of the material for COVID-19, since it allows readers to quickly grasp the main points of the papers while still saving a lot of time. The Continual BERT's scalable design also enables continuous learning over a longer period of time and with higher frequency to assimilate fresh research data and information more quickly. The evaluation shows that BERT outperforms generic scientific publications on papers with extensive medical terminology.

B. COBERT: COVID-19 Question Answering System Using BERT :

JA Alzubi et al. [4] They used COBERT, a retriever-reader dual algorithmic system that searches 59K documents of literature linked to the corona virus to provide responses to complicated queries. This paper is made available through the Coronavirus Open Research Dataset Challenge (CORD-19). The TF-IDF vectorizer used by the retriever collects the top 500 documents with the best scores. The reader, which is built on top of the HuggingFace BERT transformers and pre-trained on the SQuAD 1.1 development dataset, refines the sentences from the filtered documents before passing them to the ranker, which compares the logits scores to produce a short answer, the title of the paper, and the source article of extraction. Exact Match(EM)/F1 scores for the proposed DistilBERT version were 80.6 and 87.3, respectively, outperforming earlier pre-trained models.On the CORD-19 dataset, they refined the BERT uncased and DistilBERT versions of BERT in the Reader phase, and then compared documents based on cosine similarity to find the optimal solution. Users evaluated the system's performance on a wide range of questions and discovered that the results were reliable. They have demonstrated with the aid of empirical findings that DistilBERT, with its triple loss combining language modeling and an EM/F1 score of 81.6/87.3, is better able to grasp features interdependence of long documents.

1) Abstractive Summarization: C Limploypipat et al. [18] In this article, they described how an LSTM neural network was used to abstractly summarize Covid-19 news. Also incorporate an attention mechanism into the encoder-decoder neural network to help it focus on particular words and perform better. They produce training data sets with data augmentation and testing data sets from COVID-19 CBC News stories for our experiments. The early findings of the studies demonstrate that summarization can produce shorter paragraphs that are succinct and simple for readers to understand.

In order to help overworked medical professionals locate reliable scientific information, Andre Esteva et al. [10] introduced a tool called CO-Search, a semantic, multi-stage search engine. CO-Search is intended to handle sophisticated searches across the COVID-19 literature. The two sequential components that make up CO-Search are a hybrid semantic-keyword retriever, which uses an input query to provide a sorted list of the 1,000 documents that are the most relevant, and a re-ranker, which further ranks the documents by relevance. Each document receives a relevance score from the re-ranker, which is determined by comparing the results of an abstractive summarization module with a question-answering module that measures how well each item responds to the query.

Shengli Song et al. [31] proposed an LSTM-CNN based ATS framework (ATSDL) that can construct new sentences by exploring more fine-grained fragments than sentences. ATSDL is composed of two main stages the first one which extracts phrases from source sentences and the second generates text summaries using deep learning. LSTM-CNN based ATS framework, named ATSDL. We apply LSTM model that was originally developed for machine translation to summarization and combine CNN and LSTM together to improve the performance of text summarization. After training, the new model will generate a sequence of phrases. This sequence is the text summary that is composed of natural sentences. (ii) In order to solve the key problem of rare words, we use phrase location information, so we can generate more natural sentences. (iii) The experiment results show that ATSDL outperforms state-of-the-art abstractive and extractive summarization systems on both two different datasets.

They apply the off-the-shelf attentional encoder-decoder RNN that was originally developed for machine translation to summarization, and show that it already outperforms state of-the-art systems on two different English corpora. Motivated by concrete problems in summarization that are not sufficiently addressed by the machine translation based model, [22] propose novel models and show that they provide additional improvement in performance and a new dataset for the task of abstractive summarization of a document into multiple sentences and establish benchmarks. They used 200 dimensional word2vec vectors (Mikolov et al., 2013) trained on the same corpus to initialize the model embeddings, but they allowed them to be updated during training. They did not use any dropout or regularization, but applied gradient clipping. They simply run the models trained on Gigaword corpus as they are, without tuning them on the DUC validation set and only change made to the decoder is to suppress the model from emitting the end-of-summary tag, and force it to emit exactly 30 words for every summary. They evaluated their models using the full-length Rouge F1 metric that employed for the Gigaword corpus, but with one notable difference: in both system and gold summaries, they considered each highlight to be a separate sentence. They used the Temporal Attention model of Sankaran et al.(2016) that keeps track of past attentional weights of the decoder and expliticly discourages it from attending to the same parts of the document in future time steps.

J Lin et .al [20] The traditional sequence-to-sequence model of neural abstractive summarization frequently experiences repetition and semantic irrelevance. We provide a global encoding architecture to address the issue, which regulates the information flow from the encoder to the decoder depending on the overall context information of the source. It is made up of a convolutional gated unit that performs global encoding to enhance the source-side information representations. Evaluations on the LCSTS show that their model performs better than the baseline models, and the analysis demonstrates that their model can produce higher-quality summaries while minimizing repetition.

X Cai et al. [6] In this research, they offer COVIDSum, a SciBERT-based summary technique that is linguistically enhanced for COVID-19 scientific papers. To be more precise, they create word co-occurrence graphs by first extracting important lines from source publications. Then, to encode sentences and word co-occurrence graphs, respectively, they use a SciBERT-based sequence encoder and a Graph Attention Networks-based graph encoder. In order to create an abstractive summary of each scientific work, they finally combine the two encodings mentioned before. When compared to other document summarizing methods, our proposed model performs significantly better when tested on the freely accessible COVID-19 open research dataset.

C Limploypipat et al. [19] In this article, they described how an LSTM neural network was used to abstractly summarize Covid-19 news. Also incorporate an attention mechanism into the encoder-decoder neural network to help it focus on particular words and perform better. They produce training data sets with data augmentation and testing data sets from COVID-19 CBC News stories for our experiments. The early findings of the studies demonstrate that summarization can produce shorter paragraphs that are succinct and simple for readers to understand.

C. Deep Learning for Text Summarization

Hayatin et al. [12] offered transformers as a core language model for producing abstractive summaries of COVID-19 news articles, utilizing architectural modification as the basis for developing the model, in research work related to the summarizing of COVID-19 news articles. They only used the MTDTG transformer model for abstractive text summarization in their research. The short summaries utilized for validation were insufficient to evaluate the summaries produced since they failed to capture the essence of the COVID-19 articles of the dataset.

Milad Moradi et al. [21] proposed an innovative method for summarizing that makes use of contextualized embeddings produced by the Bidirectional Encoder Representations from Transformers (BERT) model, a deep learning model that recently displayed cutting-edge outcomes in a number of natural language processing tasks. To find the most pertinent and instructive sentences within the input documents, they mix various BERT iterations with a clustering technique and compared the summarizer to a number of methods that have been previously reported in the literature using the ROUGE toolbox.

For extractive summarization, Rezaei et al. [25] used two deep learning architectures. Performing feature extraction and creating a feature-sentence matrix for the text sentences is the initial stage. Some of the most crucial sentence characteristics for text summarization are extracted at this stage, including sentence position, sentence length, TF-IDF, and title similarity. The Auto-Encoder neural network and the Deep Belief Network are the next two neural network types to receive this matrix as input. These networks augment the matrix. The sentence scores are calculated using this matrix, and the most significant and high-scoring sentences are chosen to be included in the summary.

A mechanism termed deepMINE has been proposed by Joshi et al [14]. The two primary components of this system are the Mine Article and the Article Summarization. The user enters the necessary keywords in the first section, and the system searches the article titles provided by CORD-19 to return related articles and links. The second component uses deep learning and natural language processing to summarize an input article.

[34] They are developing semantic visualization techniques in order to enhance exploration and enable discovery over large datasets of complex networks of biomedical relations. They apply visualization techniques to analyses the recently released Harvard INDRA Covid-19 Knowledge Network (INDRA CKN) dataset, the Blender lab Covid knowledge graph dataset (Blender KG), and the COVID-19 Open Research Dataset (CORD-19). They include several kinds of visualizations: data tables for tracing evidence associated with relations, metrics panes to display the count of evidences and the count of unique articles, tag clouds and heat maps for some metadata, type-level and phrase-level visualizations that enable users to drill down into the elements in the relations, dense visualizations for functional types, and visualizations of upstream regulators. They presented the indexing the output of NLP readers in several pipelines, in order to give semantically typed elements for tag clouds and heat maps. A unique technique developed here is the application of parameter reduction operations to the extracted relations, creating relation containers, or functional entities that can also be visualized using the same methods, allowing the visualization of multiple relations, and both partial and complete protein-protein pathways

[5]They summarised all the articles in the WHO Database through an extractive summarizer followed by an exploration of the feature space using word embeddings which were then used to visualize the summarized associations of COVID-19 as found in the text. EDA was also carried out to understand the patterns of publication and venues.(CLOUD VISUALIZATION). A low-dimensional representation of the corpus trained through word2Vec algorithm[34] which was then visualized in 3 dimensions in order to aid the exploration of feature space.Latent Semantic Analysis (LSA), is used for text summarization of abstract. limitation of the study is the availability of only the abstracts on the WHO resource and the relatively small size of the resource. Future work in this direction will include full-texts of the available peer reviewed articles, primarily for the purpose of better model tuning.

S Adhikari et al. [1] Text summarization has become crucial for extracting the proper amount of information from lengthy texts as a result of the abundance of data available nowadays. In news websites, blogs, customer review websites, and other places, we encounter lengthy pieces. This paper offers several methods for producing summaries of lengthy books. Many approaches to text summarization that have been employed up to this point have been examined in several articles. The techniques presented in this study typically result in abstractive (ABS) or extractive (EXT) summaries of texts. Techniques for query-based summarization are also covered.The majority of the discussion in the paper is focused on structured-based and semantic-based approaches to text document summarization. The summaries generated by these models were tested using a variety of datasets, including the CNN corpus, DUC2000, single and multiple text documents, etc. They have researched these techniques as well as their trends, successes, past work, and potential applications in both the field of text summarization and other areas.In this paper to determine which summary is more accurate and concise they used ROGUE scores to compare their accuracy. In some instances TDIDF scores has also been applied. Summaries we get using these models not always been the best ones, so they suggested GAN's and transfer learning for better results.

E Zolotareva et al. [40] This paper utilizes both extractive and abstractive methods to automate text summarizing. The former strategy makes use of submodular functions and the BERT language representation model, while the latter strategy makes use of the GPT-2 language model. We use two different kinds of datasets: Podcast, which consists of podcast episode transcripts, and CNN/DailyMail, a dataset of benchmarked news articles. The GPT-2's output on the CNN/DailyMail dataset produced results that are competitive with the state-of-the-art. Along with the quantitative assessment, they also conduct a qualitative study in the form of a human evaluation and look at the trained model to see if it can acquire plausible abstractions.The research described in this thesis shows that a Transformer language model that was pretrained on a vast amount of data may be simply adjusted to produce accurate summaries. They used a method which makes use of the Transformer architecture's transfer learning capabilities, which have been demonstrated on numerous other tasks. Also, When summarize news items, they observe that the summaries generated by the suggested system are high in fluency and consistent with the input documents, as one could anticipate from a system built on the Transformer architecture. The quality of the summaries produced after using this approach on podcast data varies greatly.

T Xu et al. [38] In this paper, they suggest a new architecture to improve the sequence-to-sequence attention model by fusing adversarial generative networks and reinforcement learning. First, they employ a hybrid pointer-generator network that accurately reproduces information without compromising the ability of generators to produce new words. This network copies words directly from the source text. Second, they penalize summarized content and deter repetition by using both intra-temporal and intra-decoder attention. We apply our model to the COVID-19 article title summarizing assignment that we have created and get very close approximations to the ROGUE model while improving readability.

N Wang et al. [36] In this study, we first go over the available datasets and multi-document summarizing techniques' limitations. Our studies show that our proposed flexible three-stage framework outperforms currently available unsupervised algorithms in the more general task of summarizing papers with heterogeneous content. We use the TextRank algorithm for the initial stage of information extraction. The word graph and kshortest path-based sentence fusion method are used for the second stage of information condensation to combine comparable information, while hierarchical clustering is used to group similar information together. Finally, we provide a pointer generator network-based reverse coreference model for the third repetition elimination stage.

D. Information Extraction from CORD-19 Using Hierarchical Clustering and Word Bank :

R Jain et al. [13] By creating a summary of the most important sections of a medical research paper, this paper intends to relieve the strain on doctors by preventing them from having to read the lengthy study materials. A text summarizing algorithm is always based on scoring the phrases and quantifying them in some way. They quantify sentences using the TF-IDF quantification, which is a common technique. In comparison to a threshold, they choose sentences with a high score and reject those with a low score. Several phrases in a medical study report might receive a low grade, but if they contain biological things, they might still be significant. They created a significantly better summary than some existing tools by using a dataset built on COVID-19 and biomedical terminology.

E. Health Records

Abeed Sarker et al. [28] suggested a quick and straightforward extractive summarizing method that is simple to deploy and use and may help medical experts and researchers quickly access the most recent study information. Instead of computationally expensive pre-processing and resource-intensive knowledge bases, the system uses similarity measurements produced from pre-trained medical domain-specific word embeddings along with simple features at runtime. Although their approach is implemented simply, automatic evaluation using ROUGE—a summary evaluation tool—on a public dataset for evidence-based medicine demonstrates that its performance is statistically comparable with the state-of-the-art.

Emily Alsentzer et al. [3] proposed an automatic text summarization of discharge notes. Developed a LSTM model to sequentially label topics of history of present illness notes. This model achieved an F1 score of 0.876, which shows that this model can be employed to create a dataset for evaluation of extractive summarization methods

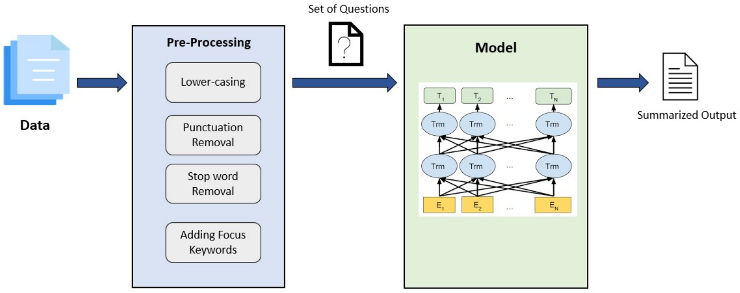

Fig. 1. Model Architecture

PHA Chen et al. [8] In this paper, a brand-new automatic text summarization system called OATS is proposed. OATS first makes use of an ontology-based method using concept graphs to represent the knowledge from unstructured text and a matching algorithm to identify relevant documents. It then makes use of the answers from a question-answering model based on ”questions” specified by users to generate text summaries that are in line with the user's focus. Considering the difficulties of the task and the absence of domain-specific fine-tuning of the QA model, the results are generally positive. The QA model technique has the drawback of requiring a set of uniform questions to access data that may be presented differently in different documents. Also it has some limitations other than that the model would provide good summarization.

III. Methodology

Unsupervised system for comprehending scientific literature that accepts questions in natural language with a focus keyword and retrieves precise responses from the CORD19 corpus of scientific papers.

A. Data Pre-processing/Tokenization Methods

Data preprocessing is a vital step in building a Machine Learning/Deep Learning model. The quality of the preprocessing determines the model's performance [16]. There are various techniques that can be employed so as to clean the raw text we have. We choose to build a Text pre-processing pipeline in which we firstly lower-case the corpus we have at hand for uniformity and eliminating the punctuation marks and then we get rid of the most frequent stop words that do not add great significance to the context of the text. Stemming and Lemmatization are few very well-known text pre-processing methods, but instead of doing it manually. We employed the BERT models to the Stemming and Lemmatization. After these basic text pre-processing techniques. We have added Focus Keywords which can focuses specifically on few things related to the COVID and its symptoms. We included these specific keywords because the dataset is huge and has many research papers. So as to keep our focus on few topics would make our search and summaries precise.

As part of cleaning the data, we dropped Null value columns, duplicate titles and consider research papers from the year 2020. We create a data frame in which we try to hold the abstracts of the papers which contains terms related to the COVID and its symptoms.

Created a Data Frame such that, the data frame contains the abstracts of the paper which focuses on the given focus words. Later we used sentence similarity from the SpaCy library to calculate the similarity of the given sentence to that of the summarized answer.

B. Text summarization with BERT and Distilled BERT

BERT - Bidirectional Encoder Representations from Transformers. BERT is a free and open-source machine learning framework for natural language processing. BERT uses the surrounding text to provide context in order to help machines understand the meaning of ambiguous words in text. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets.

Transformer, an attention mechanism that recognizes contextual relationships between words in a text, is used by BERT. Transformer's basic design consists of two independent mechanisms: an encoder that reads the text input and a decoder that generates a job prediction. Only the encoder mechanism is required because BERT's aim is to produce a language model.

DistilBERT: a distilled version of BERT DistilBERT is a more compact general-purpose language representation model that, like its larger competitors, can be tweaked and performs well on a variety of applications. While the majority of earlier research focused on using distillation to create task-specific models, DistilBERT demonstrates that a BERT model's size can be reduced by 40 while still preserving 97 percent() of its language understanding skills and being 60 percent() faster. [27]

DistilBERT, the student, shares BERT's basic architecture. The pooler and token-type embeddings are eliminated, and the number of layers is decreased by a factor of 2. The investigations revealed that variations on the last dimension of the tensor (hidden size dimension) have a smaller impact on computation efficiency than variations on other factors like the number of layers. The majority of the operations used in the Transformer architecture (linear layer and layer normalisation) are highly optimized in modern linear algebra frameworks. DistilBERT thus concentrates on minimizing the amount of layers.

MORE TO BE ADDED AND BERT MODEL FIGURE

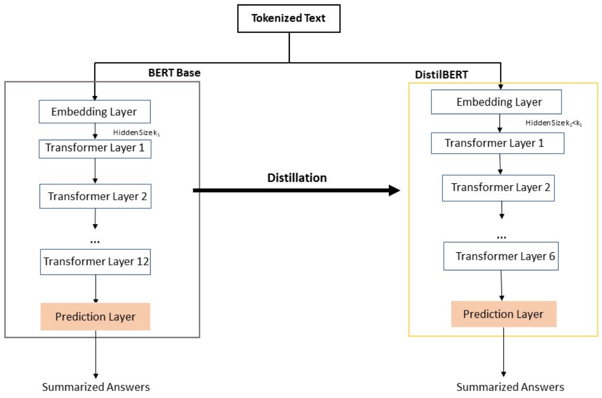

Fig. 2. BERT and DistilBERT

	No. Transformer Layers	No. of Hidden Units in each Layer	No. of
BERT-Base	12	768	12
BERT-Large	24	1024	16
DistilBERT	6	768	12

TABLE I
SPECIFICATIONS OF BERT AND DISTILBERT

IV. EXPERIMENTATION AND RESULTS

A. Dataset

In order to develop a treatment and preventative measures against the COVID-19 [35], the scientific literature needs to be surveyed by the global health and research community. The COVID-19 Open Research Dataset (CORD-19) was created by the White House and top research organizations in response to this challenge in order to bring in the NLP expertise to help uncover the solution within the literature or provide insights to the general public. Over 59,000 research articles, including over 47,000 full-text articles about the COVID-19 or associated disorders, are included in this dataset.

The dataset contains research papers from way before 2020. Hence, we segmented the dataset and dropped the research papers that are before 2020.

B. Evaluation Metrics

REFERENCES

[1] Surabhi Adhikari et al. Nlp based machine learning approaches for text summarization. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 535–538. IEEE, 2020.

[2] Mahsa Afsharizadeh, KOMLEH HOSSEIN EBRAHIMPOUR, and Ayoub Bagheri. Automatic text summarization of covid-19 research articles using recurrent neural networks and coreference resolution. 2020.

[3] Emily Alsentzer and Anne Kim. Extractive summarization of ehr discharge notes. *arXiv preprint arXiv:1810.12085*, 2018.

[4] Jafar A Alzubi, Rachna Jain, Anubhav Singh, Pritee Parwekar, and Meenu Gupta. Cobert: Covid-19 question answering system using bert. *Arabian journal for science and engineering*, pages 1–11, 2021.

[5] Raghav Awasthi, Ridam Pal, Pradeep Singh, Aditya Nagori, Suryatej Reddy, Amogh Gulati, Ponnurangam Kumaraguru, and Tavpritesh Sethi. Covidnlp: A web application for distilling systemic implications of covid-19 pandemic with natural language processing. *MedRxiv*, 2020.

[6] Xiaoyan Cai, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. *Journal of Biomedical Informatics*, 127:103999, 2022.

[7] Abdullah Javaid Chaudhry, Shehryar Hanif, and Muhammad Ali. An exploration in extractive text summarization and sentence vectors with specific reference to covid-19 medicinal articles.

[8] Po-Hsu Allen Chen, Amy Leibrand, Jordan Vasko, and Mitch Gauthier. Ontology-based and user-focused automatic text summarization (oats): Using covid-19 risk factors as an example. *arXiv preprint arXiv:2012.02028*, 2020.

[9] Vipul Dalal and Latesh Malik. A survey of extractive and abstractive text summarization techniques. In *2013 6th International Conference on Emerging Trends in Engineering and Technology*, pages 109–110. IEEE, 2013.

[10] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine*, 4(1):1–9, 2021.

[11] Lyudmila Gadasina, Vladislav Veklenko, and Pasi Luukka. Summarization algorithms for news: A study of the coronavirus theme and its impact on the news extracting algorithm. In *International Conference on Computational Data and Social Networks*, pages 351–360. Springer, 2021.

[12] Nur Hayatin, Kharisma Muzaki Ghufron, and Galih Wasis Wicaksono. Summarization of covid-19 news documents deep learning-based using transformer architecture. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 19(3):754–761, 2021.

[13] Rushit Jain, Bhavesh Bellaney, and Parth Jangid. Information extraction from cord-19 using hierarchical clustering and word bank. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2021.

[14] Bhrugesh Joshi, Vishvajit Bakarola, Parth Shah, and Ramar Krishnamurthy. deepmine-natural language processing based automatic literature mining and research summarization for early-stage comprehension in pandemic situations specifically for covid-19. *bioRxiv*, 2020.

[15] SWARANJALI JUGRAN, ASHISH KUMAR, BHUPENDRA SINGH TYAGI, and VIVEK ANAND. Extractive automatic text summarization using spacy in python & nlp. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 582–585. IEEE, 2021.

[16] Ammar Kadhim. An evaluation of preprocessing techniques for text classification. *International Journal of Computer Science and Information Security,*, 16:22–32, 06 2018.

[17] Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *arXiv preprint arXiv:2006.01997*, 2020.

[18] Chatchawarn Limploypipat and Nuttanart Facundes. Abstractive text summarization for covid-19 news with data augmentation. In *2022 International Conference on Digital Government Technology and Innovation (DGTi-CON)*, pages 56–59, 2022.

[19] Chatchawarn Limploypipat and Nuttanart Facundes. Abstractive text summarization for covid-19 news with data augmentation. In *2022 International Conference on Digital Government Technology and Innovation (DGTi-CON)*, pages 56–59. IEEE, 2022.

[20] Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. Global encoding for abstractive summarization. *arXiv preprint arXiv:1805.03989*, 2018.

[21] Milad Moradi, Georg Dorffner, and Matthias Samwald. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. *Computer methods and programs in biomedicine*, 184:105117, 2020.

[22] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[23] Alfrick Opidi. A gentle introduction to text summarization in machine learning. *Blog, FloydHub, April*, 15, 2019.

[24] Jong Won Park. Continual bert: Continual learning for adaptive extractive summarization of covid-19 literature. *arXiv preprint arXiv:2007.03405*, 2020.

[25] Afsaneh Rezaei, Sina Dami, and Parisa Daneshjoo. Multi-document extractive text summarization via deep learning approach. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pages 680–685, 2019.

[26] Deepika S, Lakshmi Krishna N, and Shridevi S. Extractive text summarization for covid-19 medical records. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–5, 2021.

[27] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[28] Abeed Sarker, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Aamir Abbas. A light-weight text summarization system for fast access to medical evidence. *Frontiers in digital health*, 2:585559, 2020.

[29] Aakash Sinha, Abhishek Yadav, and Akshay Gahlot. Extractive text summarization using neural networks. *arXiv preprint arXiv:1802.10137*, 2018.

[30] Guohui Song and Yongbin Wang. A hybrid model for medical paper summarization based on covid-19 open research dataset. In *2020 4th International conference on computer science and artificial intelligence*, pages 52–56, 2020.

[31] Shengli Song, Haitao Huang, and Tongxiao Ruan. Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools and Applications*, 78(1):857–875, 2019.

[32] A Spinelli and G Pellino. Covid-19 pandemic: perspectives on an unfolding crisis. *Journal of British Surgery*, 107(7):785–787, 2020.

[33] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management. *arXiv preprint arXiv:2005.03975*, 2020.

[34] Jingxuan Tu, Marc Verhagen, Brent Cochran, and James Pustejovsky. Exploration and discovery of the covid-19 literature through semantic visualization. *arXiv preprint arXiv:2007.01800*, 2020.

[35] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William Cooper Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas A. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.

[36] Ning Wang, Han Liu, and Diego Klabjan. Large-scale multi-document summarization with information extraction and compression. *arXiv preprint arXiv:2205.00548*, 2022.

[37] Awane Widad, Ben Lahmar El Habib, and El Falaki Ayoub. Bert for question answering applied on covid-19. *Procedia Computer Science*, 198:379–384, 2022. 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare.

[38] Tianyang Xu and Chunyun Zhang. Reinforced generative adversarial network for abstractive text summarization. *arXiv preprint arXiv:2105.15176*, 2021.

[39] Mahmood Yousefi-Azar and Len Hamey. Text summarization using unsupervised deep learning. *Expert Systems with Applications*, 68:93–105, 2017.

[40] Ekaterina Zolotareva, Tsegaye Misikir Tashu, and Tomás Horváth. Abstractive text summarization using transfer learning. In *ITAT*, pages 75–80, 2020.

[41] Zi Yue Zu, Meng Di Jiang, Peng Peng Xu, Wen Chen, Qian Qian Ni, Guang Ming Lu, and Long Jiang Zhang. Coronavirus disease 2019 (covid-19): a perspective from china. *Radiology*, 296(2):E15–E25, 2020.