Abstractive Text Summary of COVID-19 Documents based on LSTM Method and Word Embedding

Saja Naeem Turky*
Informatics Institute for Postgraduate Studies (IIPS), Iraqi Commission for Computers and Informatics (ICCI), Baghdad, Iraq.
E-mail: ms201910522@iips.icci.edu.iq

Ahmed Sabah Ahmed AL-Jumaili
Department of Bio Informatics (BI), College of Bio Medical Informatics, University of Information Technology and Communications, Baghdad, Iraq.
E-mail: asabahj@uoitc.edu.iq

Rajaa K. Hasoun
University of information technology and communications, Baghdad, Iraq
E-mail: dr.rajaa@uoitc.edu.iq

Received May 12, 2021; Accepted September 15, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18I2/WEB18370

Abstract

An abstractive summary is a process of producing a brief and coherent summary that contains the original text's main concepts. In scientific texts, summarization has generally been restricted to extractive techniques. Abstractive methods that use deep learning have proven very effective in summarizing articles in public fields, like news documents. Because of the difficulty of the neural frameworks for learning specific domain- knowledge especially in NLP task, they haven't been more applied to documents that are related to a particular domain such as the medical domain. In this study, an abstractive summary is proposed. The proposed system is applied to the COVID-19 dataset which a collection of science documents linked to the coronavirus and associated illnesses, in this work 12000 samples from this dataset have been used. The suggested model is an abstractive summary model that can read abstracts of Covid-19 papers then create summaries in the style of a single-statement headline. A text summary model has been designed based on the LSTM method architecture. The proposed model includes using a glove model for word embedding which is converts input sequence to vector forms, then these vectors pass through LSTM layers to produce the summary. The results indicate that using an LSTM and glove model for word embedding together improves the summarization system's performance. This system was evaluated by rouge metrics and it achieved (43.6, 36.7, 43.6) for Rouge-1, Rouge-2, and Rouge-L respectively.

Keywords

Abstractive Summary, COVD-19, Glove Word Embedding Model, LSTM, Rouge Metric.

Introduction

The amounts of information posted on the Internet have recently grown to such levels that skimming through them will be a time-consuming process. And the propagation of coronavirus illness had caused havoc all over the world. Also, a large number of scientific papers have been posted and made freely accessible to the medical community since the coronavirus outbreak, the international health & medicine requires a manner to summarize the scientific text for developing prevention and treatment strategy from coronavirus. In addition, scientists are progressively in most need of summary systems that can collect data from scientific research papers and summarize them to reduce the time and effort they need to read them. With the accelerating development of information in textual form through the internet, particularly with the advancement of technology, a requirement for investigating those information has surfaced and summarized them (AL-Jumaili & Tayyeh, 2020). The automatic summary is a method for producing short, succinct summaries of a single or a group of a document. In the natural language processing (NLP) field the automated text summarization is one of its tasks. In natural language processing text, summarization is a well-explored area. When it comes to summarizing text, there are two primary approaches: extractive and abstractive. The extractive method takes the most important aspects and sentences of a document and incorporates or re-rank them to produce a summary (K & Mathew, 2020). While the abstractive method is to extract and paraphrase the input text to produce a summary with high accuracy as if a human-written summary (K & Mathew, 2020). There are two main challenges in text summary first, extracting from the source document's the most relevant information and principles, and second, putting all of this information into a logical summary. Abstractive methods have been recently demonstrated to be effective in general fields, for example, in news. There is a smaller effort that has been done to apply abstractive techniques for specific domains. In this work, an abstractive method has been applied upon COVID-19 documents for generating an abstractive summary model which can produce new brief summaries after it being totally aware of domain knowledge. This paper presents a simple summary process to create summaries in the style of a single-statement headline through training the model to create papers' headlines using the papers' abstracts as inputs of the model and the paper’s title as a target outputs.

The key contributions for this work are first using the recent methods in abstractive summary documents in medical fields such as covid-19 and second, producing an

abstractive summary model by using the LSTM method with word embedding. Also, the proposed method applicable through any other field.

Related Work

Paul Gigioli et al (Gigioli et al., 2019) proposed a deep-reinforced abstractive summary system for biomedical documents, they use two frameworks first, a sequence to sequence attention framework, second, a pointer-generator framework as their baseline neural network models. In the sequence-to-sequence attentional framework, they use a one-layer unidirectional recurrent neural network (RNN) and gated recurrent units (GRU) for each encoder and decoder. Their system is applied to the MEDLINE dataset and evaluates the result by using ROUGE, UMLS, MeSH, and TF-IDF word appearance, ROUGE metrics for this system achieved (42.43, 21.59, 36.89) for R-1, R-2, and R-L.

Joshi et al (Joshi et al., 2020) article and MINI article Summary are the two key sections of this framework. The first phase the client inserts the required `words and the system returns relevant articles and links based on a search of the titles of the articles given by CORD-19. The second section uses deep learning and Natural Language Processing (NLP) to execute a summary for an input article.

Mahsa Afsharizadeh et al (Afsharizadeh et al., 2021) proposed a system for extractive text summary by using a recurrent neural network. There are three stages to the proposed method sentence encoding, word ranking, and summary creation. They use coreference resolution to enhancing the performance of the system.

Tan et al (Tan et al., 2020) proposed a system for text summarization. Extractive summary and abstractive summary are two phases that have been involved in this system. The first phase is completed using a pre-trained BERT model, and the second phase is completed using the GPT-2 model. The use COVID-19 dataset. The BERT framework is used in the first phase to transform the sentences into sentence embedding. The set of sentence embedding is then clustered using k-medoid clustering to produce a set of k clusters. An extractive summary has been constructed by this group of sentences. The extractive summary is then analyzed with POS-tagger to extract a number of keywords. After training, system summaries are created using keyword-reference summary pairs given to the GPT2 model.

Dan Su1 et al (Su et al., 2020) have suggested a system for an extractive summary. This system has three major parts: retrieval information, the answer to questions, and the summary part. In the part of the retrieval information, it gets the user question and returns

the top n most important sections. In the second part, the answer to questions section determines a group of the high closely statements obtained as the answer in the previous phase. The answer to questions part is used to pick appropriate sentences from each of the n sections for responses to the question. The top k sections are then specified after these n sections are re-ranked based on theirs marked responses. The summary part takes these k sentences and creates an extractive and abstractive summary out of them. Abstract summary created based on using BART transformer and an extractive summary created based on cosine similarity.

Park et al (Park, 2020) suggested an extractive summary model based on the BERT framework for document summarization. This framework mostly continuous trains on new information. This adjective is more beneficial to COVID-19 documents which are published every day. This study employs two different BERT models with layer-wise connections. And uses a process of replacement training to reduce forgetting catastrophic. This model stacked a small Transformer encoder on top for extract sentences. And it achieved (35.2, 15.5, 33.8) for R-1, R-2 and R-L.

Ju Zhang and Zhengzhi Lou (Zhengzhi Lou, 2020) proposed an abstractive summary model for summarizing scientific literature topics. The model has been implemented based on the coronavirus dataset. They investigated how to produce summaries on the COVID-19 academic literature corpus using three deep-learning-based abstractive summarizing models: PropheNet, T5, and BART. For each paper, they employ deep learning-based models to produce a headline from the paper's abstract. Then they evaluate the suggested system with ROUGE metrics. Then they found that the fine-tuned ProphetNet performs well in the score of ROUGE recall and f-1 measures, while the fine-tuned T5 model, on the other hand, has a good ROUGE precision score.

Methodology

The suggested system has been trained on an appropriate dataset consisting of a huge value of statements. The one-of-a-kind words are used to build a vocabulary. Hence, features are retrieved and the LSTM method is used to create a summary. The suggested system phases are explained in Fig 1.

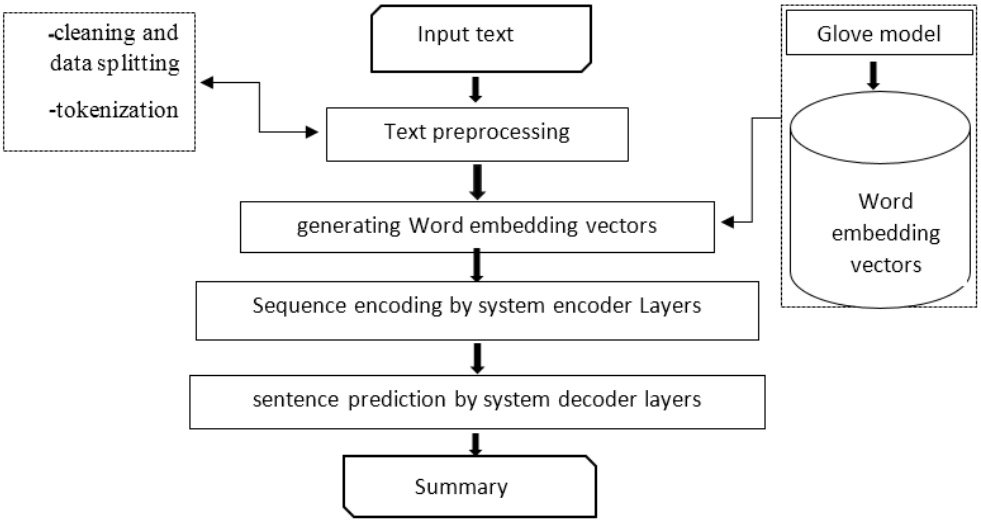

Figure 1 phases of the suggested system

The total phases of the system are subdivided into main units which are explained in the following:

A. Dataset Preprocess

At this phase, the processing process has been used for the gathered data to prepare them for training the system. This unit is subdivided into several units which are explained in the following:

1. Data cleaning: in this step, all unnecessary data will be removed.
2. Tokenization: the process of divide the text to appropriate modules (words). The modules are called tokens, the token then can be represented by a specific number.
3. Padding is a process executing to get an equal length for the outcomes that have been generated in the previous stage. A process of pre-pad the sequences have been done in this work based on the dataset's largest sequence length (Das et al., 2020).

B. Word Embedding

Every word should be converted into a numerical form that the machines can understand. Creating the appropriate form for words is an important part of the NLP task's final performance. Word embedding is a way to convert words into dense real vectors. The Global Vectors (GloVe) method is a common word embedding model used in the process of word embedding. The Glove model is an unsupervised technique that generates word vector representations by using word co-occurrence ratio possibilities. This research has used pre-trained model word embedding which was derived from Wikipedia 2014 and Gigaword 5 data based on the Glove (Pennington et al., 2014). The dimensional of the

GloVe model in this system is set to 100 to create vectors, after the vectors are generated by the glove model they will transfer to recurrent layers.

C. Encoder-Decoder Model

The sequence-to-sequence model is implemented to the encoder's hidden states and used to provide context data for the decoder. For the encoder, the suggested model uses three layers of LSTM, and for the decoder, a single layer of LSTM has been used. The encoder receives the words which have been entered and preprocessed, the words have been mapped to their word embedding vectors by the Glove word embedding model and then passing to the embedding layer. The sequence then will be encoded in the form of a static vector representation. After that, the decoder will receive them. On other hand, the decoder work to creates the output sequence by expecting a word at each time step based on the word expected previously, prior hidden state, and a generated context vector by the module.

The suggested system has multiple layers inside it in a sequential way, these various layers will describe below.

1. Embedding Layer: Word embedding is a technique for mapping a set of words into vector format to improve neural network ability (Rong, 2016). Working with data in form of numbers rather than data in form of text is more effective for neural networks.
2. Long Short-Term Memory (LSTM) Layer: In this model LSTM layer comes after an embedded layer. The output space dimension to the LSTM layer in this system equals 300. Three LSTM layers in the encoder and one LSTM layer in the decoder have been used. LSTM network (Farahani et al., 2020) contains three gates as shown in fig 2.

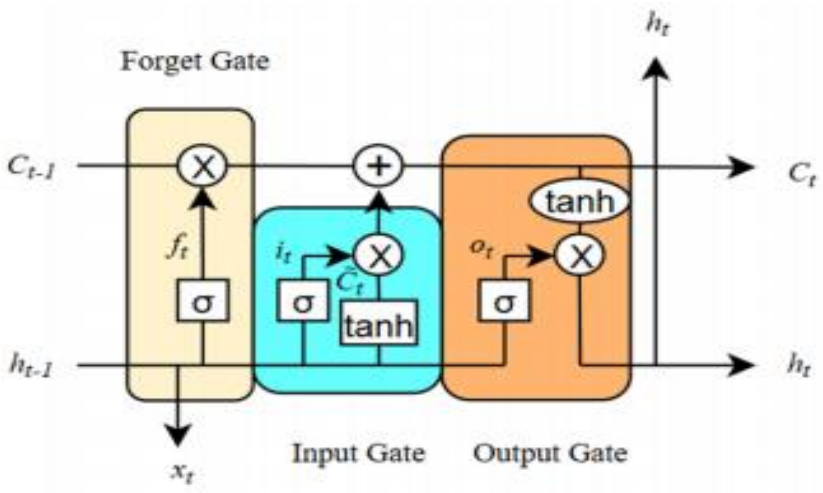

Figure 2 LSTM Cell Component

These gates will explain in the following:

a) The Input Gate: Here, the sigmoid function specifies which input values should be used to modify the memory. The input gate is represented by i_t.

b) The Forget Gate: In this gate, the sigmoid function defines the values that must to eliminated or didn't take into consideration. The forget gate is represented by f_t.

c) The Output Gate: Here the output will be determined by using the memory block and inputs. The output gate is represented by o_t.

3. Dense Layer: This model has applied a dense Layer with the "softmax" activation function which a non-linear activation function. The length of the vocabulary is equaled to the dimensionality of outer space. The model is compiled with the "adam" optimizer (Kingma & Ba, 2015).

The suggested model has been built with many orderly layers as the input layer, word embedding layer, LSTM, and dense layer. The embedding layer in this system gets the words vectors from the glove model after transformed the sequence of words into words vectors by the glove model. Then the Embedding layer feeds the next LSTM layer the production of static length size vector encoded statements. Long Short Term Memory (LSTM) layer begins the computation procedure to all input in the specific individual cell by the memory cell and the three gates. Then it transforms the outcomes to the following layer in the system which is the Dense layer. By train and validating the system products its parameters, and the prediction summary of the system will be produced.

System Evaluation

For model evaluation, the Rouge metric has been used, which calculates the overlap between produced summaries and original summaries was written by a human, rouge was created to evaluate summary systems (C. Lin,2004). Although widely recognized as a useful metric for summary evaluation.

$$\text{Rouge -N}= \frac{(\sum_{\text{Reference Summary}} \sum_{\text{N gram}} \text{count}_{\text{match}}(\text{N gram}))}{(\sum_{\text{Reference Summary}} \sum_{\text{N gram}} \text{count}\,(\text{N gram}))}$$

Dataset Description

CORD-19 (COVID-19 Open Research Dataset) is recourses for papers on COVID-19 or associated illnesses. It is a collection of science documents linked to the coronavirus (Wang et al., 2020). This database is intended to aid in creating an information retrieval framework

and text mining frameworks for COVID-19-related scientific papers. CORD-19 aims to bring together the machine learning community with biomedical experts in order to speed up the discovery of knowledge from scientific journals. To derive valuable information from this dataset, users use a range of artificial intelligence-based techniques. More than 59,000 research papers are included in this dataset, containing over 47,000 full papers on COVID-19 or associated illnesses. In this work, 12000 samples from this dataset have been used. Table 1 displays different summary methods that have been executed on the covid-19 dataset.

Table 1 Text summary methods that apply to the COVD-19 dataset

Model	**methods**	**summary type**	**Rouge metrics degree**
Joshi (Joshi et al., 2020)	Natural Language Processing (NLP) and deep learning method.	Extractive summary	Not available
Dan Su1 (Su et al., 2020)	ALBERT transformers for choose statement they use cosine similarity degree.	Both extractive and abstractive summary	Not available
Tan (Tan et al., 2020)	BERT and GPT2 transformers for choose statement they use k-medoid clustering.	Both extractive and abstractive summary	Not available
Jong Won Park (Park, 2020)	BERT transformers	Extractive summary	(35.2 ,15.5 ,33.8) for R-1, R-2 and R- L
Mahsa Afsharizadeh (Afsharizadeh et al., 2021)	a recurrent neural network	Extractive summary	R-1 achieved (0.53383, 0.34383, 0.25357) for precision, f-measure, and recall. R-2 achieved (0.17829, 0.11616, 0.08614) for precision, f-measure and recall. R-L achieved (0.25287, 0.18803 ,0.14966) for or precision, f-measure and recall

Experimentations and Parameter Settings

In this work, the glove model has been used for word embedding, glove model has been trained on Wikipedia 2014 and the Gigaword5 dataset. The dimension vector of the glove model 100. The size of the vocabulary generated from the model is 29519. This system contained three LSTM layers at the encoder and one LSTM layer at the decoder. The dimension of the LSTM layer is set to 300. The total number of parameters generated by

this system is 9,904,842 parameters. The batch size is set to 512. The dropout rate is set to 0.5. Adam optimizer is used to adapt parameters of system with learning rate = 0.001, This work constructed the suggested neural network by using the TensorFlow library. Google colab has been used which provides a Tesla K80 GPU of about 12GB. Each epoch in this system takes 30 seconds.

Results and Discussion

The system's training set contains 12000 samples of the dataset. The loss rate is determined after the training data is fitted into the model. It can be seen from Fig 3. which shows the rate of the loss function that is getting lower and being close to zero after each epoch. To illustrate the effect, the rate is mapped into the graph.

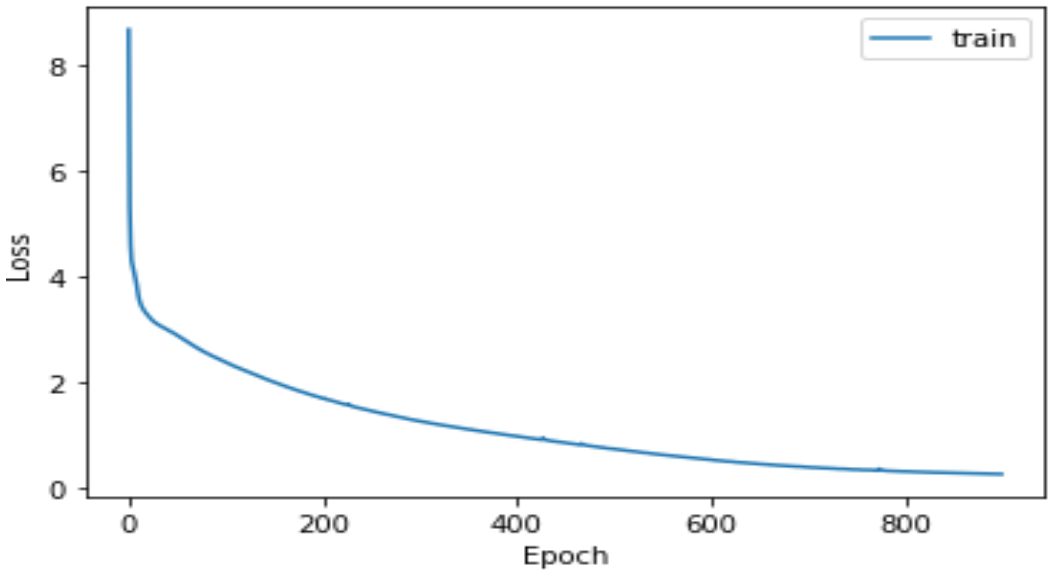

Figure 3 Loss function for 900 epochs for the suggested system

According to Fig1, which illustrates the total phases of the system. Firstly, the table has been show the applied preprocessing process to the input text. Preprocessing is used to decrease the features that are not important (Hayatin et al., 2021), also to reduce the total volume of memory required to complete the training phase. Secondly, after the step of preprocessing the table clear that the glove model has been used for word embedding which maintaining global concepts of the text by generates an array of the global co-occurrence through calculating the likelihood of a given term co-occurring with other terms. In this system, the dimensional of the GloVe model is set to 100. The total number of vocabulary generated from the model is 29519. Finally, the outcomes of the glove model will be fed to the encoder-decoder model which then uses them to generate the summaries.

Table 2 comparative analysis with previous techniques

Methods	Rouge-1			Rouge-2			Rouge-L		
ProphetNet (Zhengzhi Lou, 2020)	P	R	F-1	P	R	F-1	P	R	F-1
	29.18	65.29	39.01	18.89	42.98	25.31	23.70	54.06	31.82
T5 (Zhengzhi Lou, 2020)	57.11	14.11	21.75	22.29	5.33	8.22	41.58	10.12	15.61
BART (Zhengzhi Lou, 2020)	31.15	11.04	15.61	8.54	2.79	4.02	26.06	9	12.8
Proposed method LSTM+Glove model	**56.3**	**36.4**	**43.07**	**46.9**	**29.8**	**35.4**	**55..6**	**35.9**	**42.4**

Table 2 performs a comparative analysis with previous techniques T5, ProphetNet, and BART. The table contains the highest scores achieved by the previously used methods and scores achieved by the suggested model and illustrated that our technique gets a higher degree in precision and f-1 measures than the previous methods. Till now, only a few researchers have proposed methods for systematically summarizing COVID-19 documents as illustrated in table 1 and most of them were on extractive summary and little of them on an abstractive summary. This paper provides an abstractive summarization which evaluated by using Rouge metrics measurement. And according to outcomes of the entire system the rouge metrics achieved (43.3,36.9,43.3) for Rouge-1, Rouge-2and Rouge-L respectively.

Tables 3and 4 show examples of a source abstract with an original headline and the generated headline produced by the suggested model. The result indicates the proposed system-created summary looks very similar to the original summary.

Table 3 EX. 1 of the Model

Source abstract
Background yields of virus induced interferon if n by leukocyte cultures were previously suggested to be associated with recurrent respiratory infections in children et al objectives to investigate if the observed ifn producing capacity was secondary to the underlying disease and consequently would be after recovery of the child from the chain of infections study design forty eight year old children suffering from recurrent upper respiratory tract infections acute otitis media included were followed up for years their clinical condition and virus induced interferon production in cultures of peripheral blood leukocytes were examined at the beginning and end of this period results in 24 children the health improved during the follow up in 12 children mild improvement place while 12 children remained ill ifn yields in cultures stimulated with and respiratory viruses improved along with the clinical situation of the children parallel cultures induced with influenza or rhinoviruses did not show similar correlation conclusion these results suggest that the relationship between interferon production by leukocyte cultures and recurrent infections is complex and may be virus specific.
Original summary
Start virus induced interferon production in leukocyte cultures from children with recurrent respiratory infections follow up study end
Generated summary
Start virus induced interferon production in leukocyte cultures from children with recurrent respiratory infections follow up study with end

Table 4 EX.1 of the Model

Source abstract
1 secretory inhibitor of canonical signaling plays critical role in certain bone loss diseases studies have shown that serum levels of are significantly higher in rheumatoid arthritis ra patients and are correlated with the severity of the disease which indicates the possibility that bone in ra may be inhibited by neutralizing the biological activity of in this study panel of twelve peptides have been selected using the software 7 1 and screened high affinity and immunogenicity epitopes in vitro and in vivo assays furthermore four cell epitopes have been optimized to design novel dna vaccine and evaluated its bone protective effects in collagen induced arthritis mouse model of ra high level expression of the designed vaccine was measured in supernatant of cells in addition immunization of balb c mice with this vaccine was also highly expressed and sufficient to induce the production of long term igg which neutralized natural in vivo importantly this vaccine significantly attenuated bone in mice compared with positive control mice these results provide evidence for the development of dna vaccine targeted against to attenuate bone.
Original summary
Start designation of novel dkk1 multi epitope dna vaccine and inhibition of bone loss in collagen induced arthritic mice end
Generated summary
Start designation of novel dkk1 multi epitope dna vaccine and inhibition of bone loss in collagen induced arthritic mice mice end.

Conclusion

In this work, a deep learning system has been suggested of an abstractive summary model which can generate domain-aware summaries of COVD-19 documents. And the Rouge metric has been used to evaluate the results and it achieved (43.6, 36.7, 43.6) for rouge-1, rouge-2, and Rouge-L. The final results indicated that enriching the representation of word with glove word embedding vectors and using them in LSTM-based summary systems led to enhanced results of summary. While the study scope of this research was restricted to the creation of one-sentence titles or headlines, the techniques used here can be applied in the future to the creation of multi-sentence summaries and also applicable through any other fields.

References

AL-Jumaili, A.S.A., & Tayyeh, H.K. (2020). A hybrid method of linguistic and statistical features for Arabic sentiment analysis. *Baghdad Science Journal*, *17*(1), 385–390. https://doi.org/10.21123/BSJ.2020.17.1(SUPPL.).0385

Kurian, S.K., & Mathew, S. (2020). Survey of scientific document summarization methods. *Computer Science*, *21*(2), 141-177.

Gigioli, P., Sagar, N., Rao, A., & Voyles, J. (2019). Domain-Aware Abstractive Text Summarization for Medical Documents. *Proceedings - IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, 2338–2343. https://doi.org/10.1109/BIBM.2018.8621539

Joshi, B., Bakarola, V., Shah, P., & Krishnamurthy, R. (2020). deepMINE - Natural language processing based automatic literature mining and research summarization for early-stage comprehension in pandemic situations specifically for COVID-19. *BioRxiv*, 1–9. https://doi.org/10.1101/2020.03.30.014555

Afsharizadeh, M., Ebrahimpour-Komleh, H., & Bagheri, A. (2021). Automatic Text Summarization of COVID-19 Research Articles Using Recurrent Neural Networks and Coreference Resolution. *Frontiers in Biomedical Technologies*, *7*(4), 236–248. https://doi.org/10.18502/fbt.v7i4.5321

Tan, B., Kieuvongngam, V., & Niu, Y. (2020). Automatic text summarization of COVID-19 medical research articles using BERT and GPT-2. *ArXiv 2006.01997 v1 [Cs.CL]*.

Su, D., Xu, Y., Yu, T., Siddique, F. Bin, Barezi, E., & Fung, P. (2020). *CAiRE-COVID: A Question Answering and Query-focused Multi-Document Summarization System for COVID-19 Scholarly Information Management*. https://doi.org/10.18653/v1/2020.nlpcovid19-2.14

Park, J.W. (2020). Continual BERT: Continual Learning for Adaptive Extractive Summarization of COVID-19 Literature. *ArXiv*.

Zhengzhi Lou, J.Z. (2020). Abstractive Summarization on COVID-19 Publications. *CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX Template Borrowed from NIPS 2017)*, 1–6.

Das, S., Partha, S.B., & Imtiaz Hasan, K.N. (2020). Sentence Generation using LSTM Based Deep Learning. *IEEE Region 10 Symposium, TENSYMP 2020*, 1070–1073. https://doi.org/10.1109/TENSYMP50017.2020.9230979

Pennington, J., Socher, R., & Manning, C.D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/v1/d14-1162

Rong, X. (2016). word2vec Parameter Learning Explained. *ArXiv:1411.2738v3 [Cs.CL]*., 1–21.

Farahani, M., Farahani, M., Manthouri, M., & Kaynak, O. (2020). Short-term traffic flow prediction using variational LSTM networks. *ArXiv*.

Kingma, D.P., & Ba, J.L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Lin, C.Y. (2004). Rouge: A package for automatic evaluation of summaries. *In Text summarization branches out*, 74-81.

Wang, L.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., & Kohlmeier, S. (2020). CORD-19: The COVID-19 open research dataset. *ArXiv*.

Hayatin, N., Ghufron, K.M., & Wicaksono, G.W. (2021). Summarization of COVID-19 news documents deep learning-based using transformer architecture. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(3), 754–761.

Narayanaswamy, R., & Weaver, K.D. (2015). The impact of information and communication technologies on book challenge trends in the United States: An analysis. *Webology*, *12*(2), 1-13.