

QUESTION ANSWERING SYSTEM USING BERT ON CANCER DATA

Deva Kumar Salluri^{1,*}, Jyothi Swaroop Badiginchala², Veerendra Bellapukonda² and Kudaravalli Deepika²

Received: July 15, 2021; Revised: October 05, 2021; Accepted: November 10, 2021

Abstract

Question Answering systems can be used to generate an answer as a response to a question asked by the user, rather than searching from a large pool of files including answers. The main vision of the system in Natural Language Processing is to generate perfect answers for the queries asked by the users. The System needs to have information about each kind of Cancer disease. So, to achieve this there is a need for a system that accepts the questions and paragraphs as input and returns the answers as fast and accurately as possible. To achieve this, authors need a Language Model, which is a probability distribution over sequences of text sentences. The model gives a context understanding between the bag of words and sentences. Among the all-language models, authors chose the Bert-large-uncased-whole-word-masking-fine-tuned-squad model for the question answering system. By implementing this method will be helpful to identify diseases in the early stages which will be useful to the patients.

Keywords: Cancer diseases, BERT model, Natural Language Processing, question answering, medical domain, artificial Intelligence

Introduction

With the advancement of the medical industry, if diagnosed early, most life cautionary diseases are treated. It's uncommon to know all the causes of harmful diseases, which are misinterpreted as always harmful diseases. The the main goal is to resolve the knowledge gap of harmful diseases to the users and guide them using advanced AI. This information retrieval system will help the doctors to assist their patients in analyzing their initial symptoms and causes which might impact their health (Jackson *et al.*, 2017). Question Answering (QA) aims to supply queries with accurate answers. QA uses methodologies of Natural Language

Processing and Data Extraction (Athenikos *et al.*, 2009).

Remote diagnosing systems are getting more accurate, with many benefits like cost-effectiveness, medical call support, and treatment for illness, injury in general. With the increase of advancements in substitute intelligence (AI) techniques, the QA SYSTEM has become much more reliable for the communication between medical Representatives and patients (Omogegbe *et al.*, 2020). According to a study, 56% of American people seem to be searching for health information on the internet, among that seventy-seven of them used the overall

¹ Department of CSE, VFSTR deemed to be University, Andhra Pradesh 522213, India. E-mail: sdk_cse@vignan.ac.in.

² Department of CSE, Vignan's Lara Institute of Technology and Science, Andhra Pradesh 522213, India.

* Corresponding author

search engines (He *et al.*, 2019). Directly reading and understanding text remains a difficult drawback, particularly in complicated real-world situations. Real-world situations for reading comprehension square measure are typically far more complicated. One doesn't have paragraphs of text already tagged as containing the solution to the question. Rather, one must realize and extract relevant information from presumably large-scale text materials. The most self-made approaches are variants of disguised language models, that area unit victimization automobile encoders, that area unit trained to reconstruct text wherever a random set of the words has been disguised (Lewis *et al.*, 2020a).

During these past years, online websites have attracted several users and square measure thought of as valid resources by specialists from numerous knowledge bases. Those knowledge bases will give a platform for the end-users to share information. The user inquiring about an issue does not have information about the selected topic and specialists on that equivalent topic will provide the required information. Using this approach, queries and generated answers square measure the supply of data, substituting alternative sources like documents or databases (Annamoradnejad *et al.*, 2020). The medical field has invariably a vital developing direction of knowledge to the market the event of network and medical activities and touches the necessity of health consultants, it's important to increase the analysis of question answering system for the medical field. The medical Question Answering system (MediQA) is an intelligent service computer code that is based on most of the Medical domain knowledge and health question answering. It may extract medical-related data for resolving the Question of medical health for users (Zhou *et al.*, 2018).

Query and response method between the patient and medical knowledge base whose outputs are held on within the Electronic Medical Record within the type of explanation to help within the method of Clinical Decision Support. For the computing machines to method natural languages and neural networks are employed (Putra *et al.*, 2019). Knowledge Base Question Answering could be a Question answering over databases has additionally been well studied generally while not victimizing any other data collections. A recent study has proven that the increased databases with Wikipedia data, extend factual coverage once a given mental object is incomplete. They understood what will be loosely seen as a pattern of language. The generated answers the query supported by an outsized document set, and therefore the mental object is employed to raise the

similarities between completely different passages (Min *et al.*, 2019).

For resolving this issue, BERT comes into the picture. BERT is meant to pre-trained deep bidirectional representations from the unlabeled text by collectively learning on each left-right meaning altogether into layers. As a result, the BERT model of pre-trained may be fine-tuned with the output layers to form progressive models for good different types of tasks, like question respondent and language abstract thought, while not substantial task-specific design modifications (Devlin *et al.*, 2019). We propose the CANCER_QA, our reading comprehension task on cancer information to fill the information gap regarding completely different cancers. The main focus of our work was to create context classification models for a considerable variety of cancer symptoms, to permit automatic extraction regarding the kind of disease from the content. Later on, the Model will predict the spot answers as a response to the queries asked by users.

Related Work

Medical question answering (MQA) is extremely center of attention by researchers in recent days. To accomplish the knowledge gap, introduce a QA system that will be efficient for the job. Outcomes additionally showed their planned illustration proved that the accuracy of the systems by up to 5.4% exactness 2 and 5.9% Average exactness. Results additionally show the accuracy of the initiated natural illustration system. For the systems to analyze the word information, Natural Language Processing methods are used. The strategy will determine the causes of unwellness which is Named Entity Recognition (He *et al.*, 2019). Introduced DistilBERT, an all-purpose pre-trained version of BERT, four-hundredth smaller, 60% faster, that retains ninety-seven of the language understanding capabilities. showed that an all-purpose language model maybe with success trained with a distillation Associate in language context and flow than single direction language models analyzed the assorted parts with an ablation study (Sanh *et al.*, 2019). Explored RAG models that use the input sequence to retrieve text documents and use them as further context once generating the target sequence. Models leverage 2 parts are a gun dog with parameters that returns distributions over text passages given a question and a generator parametrized. For applied Statistical Machine Translation (SMT), the interpretation of unknown words has been the topic of intensive analysis. A large proportion of unknown word names, which may simply be

derived into the target text if each language shares an associated alphabet. If alphabets take issue, written text is needed. Character-based translation has conjointly been investigated with phrase-based models, which are well-tried particularly for closely connected languages (Burk, 1999).

Due to their ability to capture syntactical and natural data of words from massive scale untagged text, pre-trained word vectors are a typical element of most progressive information science architectures, together with question answering, Question deduction, and grammatical category labeling. However, these approaches for learning word vectors solely permit one context free illustration for every word (Peters *et al.*, 2018). Build a logical VQA dataset that contains thirty-two, seven hundred and eighty-nine question and answer pairs of seven classes, collected from four, nine hundred and eighty-nine images. The overall queries in our collection of data are openly ended, moving a nice opportunity for the medical field VQA analysis. Their data collections are publicly out there towards their main aim of developing ML systems that pass the certificated Board exams of the Yankee Board of Pathology and fostering analysis in the MediVQA. They developed a system to retrieve pathology pictures and titles from books and create QA pairs from the titles victimization text extraction process (Lewis *et al.*, 2020b). Proposed Medical Visual Question Answering supported the event of VQA generally pictures, the medical VQA task inherits similar techniques and achieves sure achievements. Specifically, the eye mechanisms like MCB, SAN, BAN, or CTI are applied to tell joint illustration between medical visual data and queries. Recently, approaches that directly solve different aspects of medical VQA are introduced, together with reasoning diagnosis model behavior, multi-modal fusion, dedicated framework style, and generative model for managing abnormality queries (Do *et al.*, 2021). Advanced Question Answering Systems are more similar to chatbots, which are the foremost well-liked software for this purpose. Chatbots assist the patients to spot their causes and their service front desks that will direct the user to accept care services (Bao *et al.*, 2020). One type of web-based psychological invention is a synchronous word-based, matched speech communication ("chat") in its completely different forms: as Short Message Service text electronic messaging services using SMS, portable applications, or webpages (Hoermann *et al.*, 2017). Many applications of chatbots embody patient assistants for medical areas that facilitate patients to spot their symptoms, service front desks that direct the patient to accept care service, and specialists (Chaix *et al.*, 2019). Their methodology is actuated by the central role complete by alignment in Machine Translation

and former approaches to sentence similarity modeling natural communication abstract thought, and natural parsing. The neural counterpart to alignment, which could be a key part of our approach, was originally projected and has been preponderantly employed in conjunction with LSTM and to a lesser extent with CNN. In distinction, our use of attention is only supported by word embeddings and our methodology consists of feed-forward networks that operate for the most part severally of ordination (Parikh *et al.*, 2016). Willer QA is victimization constant techniques to answer cancer connected queries asked by users and therefore the data through that answers are extracted cancer be concurrent as a result of the data extraction is completed dynamically from the updated trustworthy websites that make it a lot more reliable for patient care (Lee *et al.*, 2006).

Proposed the paragraph vector (PV) model to plant variable-length text. Model area units trained to find a word given in its same meaning or words showing during a tiny window supported a vector illustration of the supply document. Unlike most different strategies, during this work sentences are a unit thought of as atomic units rather than as an integrative operation of its words. Socher planned algorithmic autoencoders that cipher Associate in input sentences employing an algorithmic encoder and a decoder reconstructing the hidden states of the encoder (Logeswaran and Lee, 2018). Proposed A framework for understanding and acquiring the word vectors. The context of 3 words is employed to find out the fourth word. During this framework, each word is associated with a singular vector, painted by a column during a matrix. The column is indexed by the position of the word within the vocabulary. The concatenation or addition of the vectors is then used as options for the prediction of the following word during a sentence (Djuric *et al.*, 2015). Proposed a model that deals with the high accuracy on different organ systems and is like old pediatricians in the identification of common diseases. The effect is also mostly seen in area units wherever tending suppliers are in a short period. AI-based strategies have much more advanced tools to rework treatment. Though machine learning techniques have already incontestably sturdy results in diagnoses, and large electronic health record information will be difficult (Liang *et al.*, 2019). An electrical device-based language model consists of stacked Transformer blocks. Every block contains a multi-head self-attention layer followed by a completely connected point feed-forward network. Thus, existing approaches add a point bias input word embedding so that each input word is painted by a vector whose worth depends on its content and position (He *et al.*, 2020b). Here the

framework utilizes the unsupervised Denoising Auto-Encoder and hence we use supervised Meta-Learning. The main advantage of this approach is the big quantity of pictures which will be a helpful advantage of Meta-Learning is to find out weights that will adapt quickly with restricted knowledge. By investing in the benefits of those approaches, it permits the projected methodology to be expeditiously trained to employ a tiny training set (Karri and Kumar, 2020). The experiments have proven that the projected technique considerably outperforms the progressive medical VQA Most approaches for medical VQA square measure to directly apply the models to the medical knowledge base (Nguyen *et al.*, 2019).

Simple semi-supervised technique victimization pre-trained neural network models are used to compose token representation in a sequence of the connected models. Their technique considerably outperforms the present state-of-the-art models in 2 common data sets used as Named Entity Recognition and configuration. The results show that adding a backward luminous flux unit additionally to ancient LMs systematically improves accuracy (Peters *et al.*, 2017). Reported the analysis development, implementation, and a pilot analysis of the medical question answering system (MedQA). Proposed a convolutional neural network tensor spec to cipher the words in the naturals area and model their communication with a tensor layer. The model integrates sentence modeling and naturals associating it into one model, which may not solely capture the helpful data with convolutional and pooling layers, however, additionally understands the same accuracy and precision between the queries and its results (Qiu and Huang, 2015).

The learning technique in the natural language process was Universal Language Model Fine-tuning for Text Classification technique that tunes the language model on a replacement knowledge base once training a language model on a knowledge base, like Wiki. Here, at last, the finely tuned model may be utilized in a task for a replacement knowledge base (Annamoradnejad *et al.*, 2020). Increasing the model size once pre-training natural language representations usually leads to improved performance on downstream tasks. However, for some purpose more models will increasingly become more durable thanks to GPU/TPU memory limitations and longer coaching times. As a result, our greatest model establishes new progressive results on the GLUE, RACE, and SQUAD benchmarks whereas having fewer parameters compared to BERT-large (He *et al.*, 2020a). Increasing the performance of Albert, they introduced a self-talented loss for sentence order prediction. Sentence order prediction primarily aims at internal sentence coherence and is meant to

handle the ineffectualness of future sentence prediction(NSP) loss planned within the original BERT (Lan *et al.*, 2019).

Worked on question answering and reading comprehension within the field of human language technology. The Text Retrieval Conference (TREC) analysis tasks and therefore the MC Test are eminent early efforts at making datasets for answering queries supported text documents. Varied models are planned for this type of task. Earlier tries to directly use LSTM to method text and use attention to fetch information from the passage illustration LSTM layers to pre-process text input and predict answers. A span is chosen from a passage by predicting the beginning and finish positions of the solution. They studied and tested the supervised metric capacity unit classifiers, that use pre-existent categorized knowledge to derive learning (Zhang *et al.*, 2018). A large range of tools and frameworks exist for general-purpose information extraction from clinical dictionaries, like cTAKES, NOBLE and MedLee. They introduced the CRIS-CODE project, which contains the semi-permanent primary goal of providing comprehensive information science models for psychological state constructs. The main objective of the initial program of labor represented here was to generate the classification models for sentences and also for a considerable variety of Symptomatology, to permit automatic extraction for several of the foremost informative symptoms from the patient details and symptoms (Evans *et al.*, 2020).

Methodology

Our planned work is to create a Question-answering system for cancer information. Our system aims to get responses/answers to the various cancer-connected queries that were asked by the overall public or Medical Practitioners. Cancer is one in every one of the life-threatening diseases that medical help which helps to fill their information gap is extremely abundant. Many people can have different queries relating to these cancers. To clear those doubts, they could browse the Web, try and browse different articles. However, the question here is that the actual doubt might not be cleared once directly finding out their doubt Question, Even the solution found that will vary with another article in another website. Even a trustworthy and approved website article found the user required to browse and search the full content for locating the solution. So, in our model, we have chosen a trustworthy medical website with the URL "<https://www.mayoclinic.org/>". Mayo Clinic is a Non-Profitable American academic medical center of integrated

health care, education, and research. Which, maintains elaborated information relating to the cancer sort and its summary, causes, symptoms, risk factors, prevention measures, identification, and treatment. Any other website that contains required Data can be chosen for the information extraction. But Government-authorized trustworthy websites are recommended to use for accurate information.

Cancer Data Collection

The authors have hand-picked five sorts of cancers are Skin Cancer, Breast Cancer, Prostate Cancer, Colorectal Cancer, and Lung Cancer. As a result, among all Registered cases, these five types are the highest most registered cancer sorts in keeping with the SEER (Surveillance, Epidemiology, and End Results) Cancer Statistics Review until April nineteen, 2021 expressed by NCI (National Cancer Institute). Our approach aims to create use of the important information that's offered online, we tend to use Web scraping techniques to gather information into numerous varieties of cancer. However, such information would be inefficient for usage promptly thanks to potential noise. We tend to analyze the data while not eliminating the stop-words, avoiding different meanings inside the passage.

Web Scraping

Instead of the normal method of extracting the information, Web scraping is an automatic information extraction technique that retrieves tons of, millions, or perhaps billions of information records on the Web. Web scraping is an information scraping technique that is used for information extraction from the Mayo clinic website for the different cancer types and extracts the information of cure, prevention, symptoms, and causes of each sort of cancer. The online software system for scraping could directly access the Web using the HTTP protocol. Unlike screen scraping, which solely copies all the pixel data that will be displayed on the screen, Web scraping will extract underlying Hypertext Mark-up Language code and, with it, information. The hand tool will then replicate the entire website content. By using Web scraping to gather information associated with different cancer diseases from the website. It contains 3 stages:

Step1: GET the Web-request is distributed from the server so it gets the response from an online website.

Step2: Hypertext Mark-up Language code parsing of the website in a tree web structure.

Step3: Python library is then wont to seek for trees that are parsing.

Pre-processing

After collecting data, we tend to performed pre-processing the information while not removing stop words, wherever special symbols are going to be removed, and also the knowledge remaining constant. We've not removed any stop words since there are tons of impacts on the model that is recommended by BERT why because the BERT system wants to stop words to spot the context from each Left to Right and Right to Left.

Topic Modelling

As a part of data collection, we tend to organize knowledge into 5 classes of cancers like skin cancer, lung cancer, Breast cancer, Prostate cancer, Colorectal cancer. We knew vital keywords exploitation Topic modeling. Topic Models can be used for file clustering, Structuring of Question knowledge base, information retrieval will change the text which is unstructured and then have feature selection. They're being employed to prepare giant data of emails, feedback reviews. We've used Latent Dirichlet Allocation for Topic Modelling. LDA assumes documents are area units created from a group of topics. Given a large number of documents, LDA will backtrack and try to work out what topics would produce those documents within the initial place. LDA tells that every word in each document comes from a subject, also the topic is chosen from a file word distribution over topics. thus, we've 2 matrices:

1. $\Theta_{td} = P(t \setminus d)$ that is that the chance probabilistic of topics in documents
2. $\Theta_{wt} = P(w \setminus t)$ that is that the chance probabilistic of words in topics

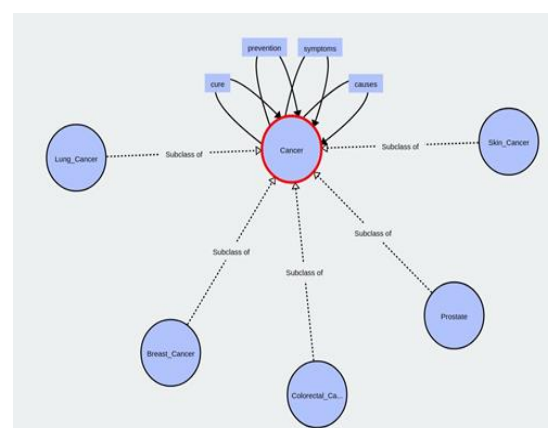


Figure 1. OWL Diagram of cancer data

And, we will take that the chance of a text is given as a file i.e. $P(w|d)$ is adequate to $P(w|t,d)$ $P(t|d)$. Here T is the total variety of topics. Also, let's assume W is the variety of words in our words for all over the files.

Here is in Figure 1 the OWL (Web Ontology Language) Diagram illustration of the Cancer knowledge set. Cancer is the main category and Subclasses of Lung cancer, Breast cancer, Colorectal cancer, Prostate cancer, and Skin Cancer the subclasses and each having properties like cure, prevention, symptoms, and causes. These are sorted in individual order to urge Associate in Organized cancer knowledge set.

Fine-Tuned BERT Model

BERT itself could be an immense model with Three forty million parameters and it is a pre-trained model that might need the next level of computation power for training the model. Here the goal is to be approachable for the low-end computer systems presently accessible at user's interests for the preprocessing and have a tendency to propose the use of BERT.

Here a fine-tuned BERT model is used for the information retrieval system and applied on a cancer knowledge set that uses a transformer model which implements the BERT QA System. It's fine-tuned by exploiting the NER (Named entity recognition), SQUAD (Stanford Question answering Dataset), and MNLI (Multi-language Inference) to form BERT helpful for Question answering represented in Figure 2. First, we tend to initiate the training model of the transformers. That additionally supports Sequence Classification, Token Classification, Question answering. This model takes the input of the question and connected Cancer Text knowledge wherever the solution ought to be fetched. Whenever we tend to decide the model, The Question and Cancer knowledge are going to be tokenized and the distillation of BERT is finished by terribly giant exploitation masking.

Applying the tokenizer to the input text, treating them as a text-pair, and looking at the input_ids for the primary instance of the '[SEP]' token. Then coverage however long the input sequence is and run through the model then the generated answer is going to be displayed. Then those tokens are going to be useful to make your mind up the context understanding in step with the question that was asked and retrieve the solution consequently. During this section, we tend to aim to enhance the model accuracy by sterilizing the output BERT. The projected methodology could be a changed model of the BERT fine-tuned QA system. During

this section, we tend to aim to enhance the model accuracy by sterilizing the output on BERT. Our Best performed design of the BERT QA System for cancer data contains 3 important layers which area units delineated below are intimately shown in Figure 3.

Embedding Layer

When using the pretraining method of BERT, the transformers will generate the results that complete the order of hidden stages like the multi-head attention block. Hidden states are the vector-connected illustration of the context and query $T_1, T_2 \dots T_n \in R^h$. This BERT will be shown because the embedding layer isn't updated throughout pre-training within the model.

Highway Network

We tend to use a 2 layered Network that transforms each hidden layer T_i . The network transformations will be used two times and every time with unique parameters. This tells us we tend to use them on top of transformations two times using unique parameters. By using this highway Network, we will have to refine the BERT model illustration with the transformation mechanism and modify the structure of LSTM with larger step transformation depths to optimize the training methods.

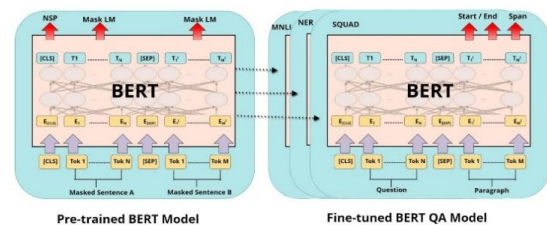


Figure 2. Fine tuning of BERT model

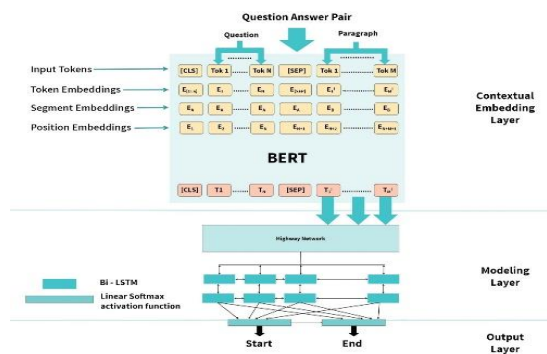


Figure 3. BERT question answering system architecture

Modeling layer

The modeling layer carries with its 2 layers of bi-LSTM with a hidden layer length of h which looks at the vector house transformation before the highway network layer. It's maintained to be the finding relation between paragraph and question. The output of the modeling layer is $M \in R^{2h \times L}$. When modeling a layer, we tend to conjointly operate to not use the over-fit.

Output Layer

In the output layer, we tend to find the solution to begin, finish chance probabilistic severally using 2 head-attentions, every performance part a linear representation followed by activation function of softmax. The label changes will be represented as $P = \text{softmax}(CWT)$ If the predicted label is zero as a not possible answer. Otherwise, the output answer given from the QA system would be eliminated.

Metrics and Analysis

Here the crucial issue is the accuracy of the generated answer. It's better to not offer any answer instead of sharing a false one. So, that's the rationale that the generated answer should be legitimate.

To validate the solution generated by our model, we tend to use the metrics of Recall-Oriented Understudy for Gisting Evaluation and Bilingual Evaluation Understudy scores. Bilingual Evaluation Understudy, maybe a metric for scrutiny of user tokenization of the text to a lot of tokens. Being developed for tokenization, it will be accustomed to analyzing the text generated by a set of Natural language processing tasks. During this initial stage, we tend to enforce the BERT model wherever we are going to predict the accuracy score by evaluating and marking text victimization in the nltk module in Python language. Here, it solely depends on the actual similarities between words. Precision and Recall are calculated as a quantitative relation of the words within the words that are similar and also within the translation, and therefore the range of text within the translation severally. Recall-Oriented Understudy for Gisting analysis is a metric for evaluating texts. furthermore, in artificial intelligence. It is working by the comparison of an automatically produced answer against a reference answer (human predicted). Precision and Recall are calculated as the ratio of non-overlapped text and all the text in the Machine.

Results and Discussion

In the Execution part, we need a Cancer knowledge data set. That data was extracted using Web scraping additionally referred to as Web extraction of data, which is a method of retrieving knowledge from the web resources like Internet sites. We tend to extract exploitation from some website links mentioned below within Figure 4.

We performed pre-processing on the information while not removing stop words, wherever special symbols are going to be removed, and also the knowledge remains similar. We did not eliminate any stop words because they play a great deal in the model BERT. Thus, we tend to apply pre-processing to the text files shown in Figure 5.

We have known vital keywords exploitation topic modeling by exploitation LDA algorithmic program. Topic modeling could be a variety of applied math procedures for locating the different "topics" that occur during an assortment of files. Latent Dirichlet allocation is a type of topic model and is employed to classify the text into topics. During this model, we want more stop words for language naturals understanding by BERT. There's no probability that the subject names are to stop words. Thus, we tend to simply unnoticed the stop words as a subject name by the LDA model. We have chosen the num_of_words as a pair of and

```
C:\Users\Suranaree\AppData\Local\Programs\Python\Python37\python.exe 0:/4-2/Project/information_extraction/extract_web.py
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/breast-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/breast-cancer/]
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/lung-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/lung-cancer/]
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/prostate-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/prostate-cancer/]
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/skin-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/skin-cancer/]
writing file breast_cancer.txt
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/lung-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/lung-cancer/]
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/lung-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/lung-cancer/]
writing file lung_cancer.txt
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/prostate-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/prostate-cancer/]
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/prostate-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/prostate-cancer/]
writing file prostate_cancer.txt
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/skin-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/skin-cancer/]
extract from https://www.ncbi.nlm.nih.gov/diseases/conditions/skin-cancer/ [https://www.ncbi.nlm.nih.gov/diseases/conditions/skin-cancer/]
writing file skin_cancer.txt
Process finished with exit code 0
```

Figure 4. Web extraction

```
C:\Users\Suranaree\AppData\Local\Programs\Python\Python37\python.exe 0:/4-2/Project/information_extraction/extract_web.py
pre-processing the file breast_cancer.txt
pre-processing the file colorectal_cancer.txt
pre-processing the file lung_cancer.txt
pre-processing the file prostate_cancer.txt
pre-processing the file skin_cancer.txt
Process finished with exit code 0
```

Figure 5. Preprocessing of cancer data

```
C:\Users\Suranaree\AppData\Local\Programs\Python\Python37\python.exe 0:/4-2/Project/information_extraction/extract_web.py
Topic modelling for the file breast_cancer.txt
topic 1 = (0.000*breast + 0.000*cancer)
topic 2 = (0.000*breast + 0.000*cancer)
Topic modelling for the file colorectal_cancer.txt
topic 1 = (0.000*colorectal + 0.000*cancer)
topic 2 = (0.000*colorectal + 0.000*cancer)
Topic modelling for the file lung_cancer.txt
topic 1 = (0.000*lung + 0.000*cancer)
topic 2 = (0.000*lung + 0.000*cancer)
Topic modelling for the file prostate_cancer.txt
topic 1 = (0.000*prostate + 0.000*cancer)
topic 2 = (0.000*prostate + 0.000*cancer)
Topic modelling for the file skin_cancer.txt
topic 1 = (0.000*skin + 0.000*cancer)
topic 2 = (0.000*skin + 0.000*cancer)
Process finished with exit code 0
```

Figure 6. Topic modelling of cancer data

also the topics have to be compelled to be fetched as a pair to match the subject names. We have got the cancer names because the topic names for every cancer type are perfectly shown in Figure 6.

Here a fine-tuned BERT model is used for the information retrieval system and applied on a cancer knowledge set that uses a transformer model which implements the BERT QA System. It's fine-tuned by exploiting the NER (Named entity recognition), SQUAD (Stanford Question answering Dataset), and MNLI (Multi-language Inference) to form BERT helpful for Question answering. Bilingual Evaluation Understudy, maybe a metric for scrutiny of user tokenization of the text to a lot of tokens. Being developed for tokenization, it will be accustomed to analyzing the text generated by a set of Natural language processing tasks Bilingual Evaluation Understudy, maybe a metric for scrutiny of user tokenization of the text to a lot of tokens. Being developed for tokenization, it will be accustomed to analyzing the text generated by a set of Natural language processing tasks. Fine-Tuned the BERT model with SQUAD (Stanford Question Answering Dataset), NER (Named entity recognition), MNLI (Multi Natural Language Inference) to remodel it into BERT Question Answering System. Model and Tokenizer initialization, Query/Question tokenization, Answer Prediction were the most activities we've got tired of this BERT Question Answering System.

We prepared different queries manually for the testing of the Question respondent system. Questions varied for every cancer type like Lung cancer, Breast cancer, Colorectal cancer, Skin cancer, and Prostate cancer. For each type of Cancer, the queries are ready for every property like Causes, Prevention, Symptoms, and treatment.

Finally, ready queries were held on during a JSON (JavaScript Object Notation) Format File specified with which kind of Cancer and on that attribute the question is to be connected. Installed Hugging face transformers library and from those transformers loaded pre-trained BERT-large model. We have taken some queries as a JSON format, so we tend to enforce our BERT model on those queries, and once implemented on those queries the generated Answers for every several questions are derived onto a computer file. The results are shown in Figure 7.

A generated Answer computer file is ready by taking the question from the JSON File that we've provided, and also the answers are going to be generated by taking each question and several attribute files which may be seen below in Figure 8.

```

question-what is breast cancer?content file-breast_cancer/overview.txt
Query has 223 tokens.

question-what is the average limit for breast?content file-breast_cancer/prevention.txt
Query has 47 tokens.

question-are we more likely to get breast cancer?content file-breast_cancer/risk_factors.txt
Query has 616 tokens.

question-what is breast lobular carcinoma?content file-breast_cancer/types.txt
Query has 314 tokens.

question-what is lymphatic?content file-breast_cancer/treatment.txt
Query has 472 tokens.

question-what is colorectal cancer?content file-colorectal_cancer/overview.txt
Query has 206 tokens.

question-what are colon cancer risk reducing methods?content file-colorectal_cancer/prevention.txt
Query has 268 tokens.

question-what type of pain occurs by colorectal cancer?content file-colorectal_cancer/symptoms.txt
Query has 332 tokens.

question-what is cause of colorectal cancer?content file-colorectal_cancer/causes.txt
Query has 134 tokens.

question-which people can have greater risk of getting colorectal cancer?content file-colorectal_cancer/risk_factors.txt
Query has 316 tokens.

question-what is lung cancer causing agent?content file-lung_cancer/causes.txt
Query has 296 tokens.

question-what do do to reduce your risk of lung cancer?content file-lung_cancer/prevention.txt
Query has 344 tokens.

question-what is the leading cause of lung cancer?content file-lung_cancer/risk_factors.txt
Query has 274 tokens.

question-what are symptoms of lung cancer?content file-lung_cancer/symptoms.txt
Query has 84 tokens.

question-what are the tests for lung cancer?content file-lung_cancer/diagnosis.txt
Query has 330 tokens.

question-what is prostate cancer?content file-prostate_cancer/overview.txt
Query has 126 tokens.

question-how about doctors know that prostate cancer occur?content file-prostate_cancer/causes.txt
Query has 125 tokens.

question-how can you reduce risk of prostate cancer?content file-prostate_cancer/prevention.txt
Query has 164 tokens.

question-how can you have a higher risk of prostate cancer?content file-prostate_cancer/risk_factors.txt
Query has 225 tokens.

question-what is the advanced symptoms of prostate cancer?content file-prostate_cancer/symptoms.txt
Query has 79 tokens.

question-what is test cancer?content file-test_cancer/overview.txt
Query has 125 tokens.

question-when does cancer occur?content file-test_cancer/causes.txt
Query has 208 tokens.

question-what are the Prevention measures for avoiding this cancer occur?content file-test_cancer/prevention.txt
Query has 447 tokens.

question-what are the signs of your parents or a sibling has had cancer, is there any risk of acquiring this cancer?content file-test_cancer/risk_factors.txt
Query has 360 tokens.

question-what are priority areas getting this cancer?content file-test_cancer/symptoms.txt
Query has 464 tokens.

```

Figure 7. Generated output after executing the Cancer Question Answering System

what is breast cancer?
cancer that forms in the cells of the breasts

what is the acceptable limit for alcohol?
no more than one drink a day

who are more likely to get breast cancer?
women

what is invasive ductal carcinoma?
granular tissue called lobules

what is lymphoma?
removing the breast cancer

what is colorectal cancer?
a tumor that sometimes starts cancer and rectal cancer

what are colon cancer risk reducing methods?
eat a variety of fruits, vegetables and whole grains

what type of pain cause by colorectal cancer?
persistent abdominal discomfort

what is cause of colorectal cancer?
healthy cells in the colon develop changes (mutations) in their dna

which people can have greater risk of getting colorectal cancer?
african - americans

what is lung cancer causing agent?
smoking

what to do to reduce your risk of lung cancer?
what you do to reduce your risk of lung cancer?
quit - i smoke , i quit - no more smoking , don't start - talk to your children about not smoking is that they can understand how to avoid this major risk factor for cancer

what are the leading cause of lung cancer?
exposure to radon gas

what are symptoms of lung cancer?
a new cough that doesn't go away, coughing up blood, even a small amount, shortness of breath, chest pain, hoarseness, losing weight without trying, joint pain

what are the tests for lung cancer?
ct scan, x-ray image, sputum cytology biopsy

what is prostate cancer?
cancer that occurs in the prostate

How about doctors know that prostate cancer occurs?
when cells in the prostate develop changes in their dna

How can you reduce risk of prostate cancer?
if you choose a healthy diet full of fruits and vegetables

who may have a higher risk of prostate cancer?
prostate are older

what are the advance symptoms of prostate cancer?
trouble urinating, increased times in the stream of urine, blood in the urine, blood in the semen, bone pain, losing weight without trying and erectile dysfunction

what is skin cancer?
the abnormal growth of skin cells

when skin cancer occurs?
when errors (mutations) occur in the dna of skin cells

what are the immediate measures for avoiding skin cancer occurs?
avoid the sun during the middle of the day, you should use radiation protection - tanning, and clouds offer little protection from damaging rays

if one of your parents or a sibling has had skin cancer, is there any risk of acquiring skin cancer?
you may have an increased risk of the disease

what are probably areas getting skin cancer?
sun - exposed skin

Figure 8. Generated answer results for different questions on cancer-related questions

Table 1. Comparison of performances of the proposed model with some other models

	Accuracy	AUC Score	Precision (Micro Avg / Macro Avg)	Recall (Micro Avg / Macro Avg)	F1 Score (Micro Avg / Macro Avg)
BERT-base (cased)	0.8884	0.9624	0.66 / 0.34	0.64 / 0.34	0.65 / 0.34
BERT-base (uncased)	0.8728	0.9604	0.61 / 0.31	0.71 / 0.38	0.65 / 0.34
bare BERT-base (cased)	0.8898	0.9602	0.68 / 0.60	0.68 / 0.51	0.65 / 0.51
bare BERT-base (uncased)	0.8721	0.9749	0.60 / 0.60	0.78 / 0.58	0.66 / 0.50
bare BERT-large (cased)	0.8895	0.9601	0.62 / 0.53	0.73 / 0.51	0.67 / 0.51
bare BERT-large (uncased)	0.8931	0.9807	0.69 / 0.60	0.73 / 0.52	0.71 / 0.52
DistilBERT (cased)	0.8766	0.9649	0.63 / 0.34	0.68 / 0.37	0.66 / 0.34
DistilBERT (uncased)	0.8549	0.9741	0.59 / 0.48	0.73 / 0.39	0.65 / 0.34

```

C:\Users\Suranaree\Documents\Programs\Python\bert\python.exe 0.0.0.0\bert\bert_model_question_answering.py
bert has 228 tokens

.....
Abstract: Pain in the Q1 System
.....
What lung cancers don't cause symptoms until the disease has advanced, in part because the lungs have few nerve endings. When lung cancer does cause signs in its early stages, they may vary from person to person but commonly include: A new cough that is persistent or worsens, or a change in an existing chronic cough that produces blood. Pain in the chest, back or shoulders that worsens during coughing, laughing or deep breathing. Shortness of breath that comes on suddenly and doesn't always subside. Unexplained weight loss. Feeling that you are tired or weak. Loss of appetite. Lung infections such as pneumonia or symptoms that won't go away. Hoarseness or wheezing. Late cancer symptoms of lung cancer may include: Swelling in the face or neck. Difficulty swallowing or pain while swallowing. Changes in the appearance of fingers, called finger clubbing. Although most of these symptoms are more likely to be caused by something other than lung cancer, it's important to see a doctor. Discovering lung cancer early may mean more treatment options are available.

.....
Question Asked to System
.....
What are late cancer symptoms of lung cancer?
.....
Manual answer given by user
.....
Swelling in the face or neck. Difficulty swallowing or pain while swallowing. Changes in the appearance of fingers, called finger clubbing.
.....
System answer given by user
.....
Swelling in the face or neck. Difficulty swallowing or pain while swallowing. Changes in the appearance of fingers, called finger clubbing.

.....
Bleu Score for the generated answer
.....
Precision is 0.970605970605
Recall is 0.970605970605
F Score is 0.970605970605
(0.970605970605, 0.970605970605, 0.970605970605)

.....
Bleu Score for the generated answer
.....
0.970605970605

***** Finished with exit code 0

```

Figure 9. Performance using BLUE and ROUGE metrics

To evaluate the performance of the QA System enforced by the BERT Model with our custom knowledge, Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) are useful to gauge the accuracy score and evaluating automatic summarizing of texts. Question Answer system is verified with the exploitation of the BERT model and manual data similarly. It returns for the queries on cancer knowledge and performance levels that are cross-checked with ROUGE and BLEU metrics. Information retrieval systems have nice potential in health care scenarios. BERT has been tested to be able to come through a better score of preciseness allowing achieving information metrics tasks during a better and a lot of economical approaches. The precision and Recall are calculated using ROUGE models on sample cancer knowledge. Exploiting this model, we have observed that precision as 96.80% and recall scores as 98% shown in Figure 9.

Question answer system for cancer data, we are implementing a model called pre-trained BERT that has good accuracy when compared to the remaining models. It shows the results for the questions on cancer data and performance levels

that are checked using Recall-Oriented Understudy for Gisting Evaluation and Bilingual Evaluation Understudy metrics. Information Retrieval systems have great potential in hospital scenarios. BERT proved that it can get the highest score of accuracy and precision by Natural processing tasks in the best and most accurate way.

Table 1 summarizes the performance of the models based on the cased type and uncased type versions. Here we used a BERT large, uncased version. Supporting this comparison, the BERT large, uncased model has given better results over the remaining models. We tend to use this model with the fine-tuned BERT-large model.

Conclusions

In this work authors have done during this paper, authors aimed to resolve the knowledge gap of the people regarding totally different cancer types.

The authors used the Bert – large - uncased – whole – word – masking – fine – tuned - squad model for the question answering system. In this Question Answer System Model, authors need to cross-check the results given by both the BERT model and the manually taken answer. The results given by the model are nearly similar for many of the questions we have given and performance levels that we need to verify using the metrics. Information retrieval systems have great potential in healthcare scenarios. BERT has achieved a higher performance score in a better and more efficient way.

References

- Annamoradnejad, I., Fazli, M., and Habibi, J. (2020). Predicting Subjective features from questions on qa websites using BERT. In: 2020 6th International Conference on Web Research, (ICWR), 2020, p. 240-244. doi: 10.1109/ICWR 49608.2020.9122318.
- Athenikos, S.J., Han, H., and Brooks, A.D. (2009). A framework of a logic-based question-answering system for the medical domain (LOQAS-Med). In: Proceedings of the ACM Symposium on Applied Computing, p. 847-851. doi: 10.1145/1529282.1529462.

- Bao, Q., Ni, L., and Liu, J. (2020). HHH: An Online Medical Chatbot System Based on Knowledge Graph and Hierarchical Bi-Directional Attention. In: Proceedings of the Australasian Computer Science Week Multiconference (ACSW '20). Association for Computing Machinery, New York, NY, USA, Article no. 32:1-10. <https://doi.org/10.1145/3373017.3373049>.
- Burk, H. (1999). Das Hunderttage-Stadion: Entstehungsgeschichte des Bad Nauheimer Kunsteisstadions unter Colonel Paul R. Knight. Stadt Bad Nauheim, 63p.
- Chaix, B., Bibault, J., Pienkowski, A., Delamon, G., Guillemassé, A., Nectoux, P., and Brouard, B. (2019). When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR Cancer*, 5(1):e12856. doi: 10.2196/12856.
- 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, 1(Mlm):4,171-4,186.
- Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Doc2Vec. In: Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion, 32:29-30.
- Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., and Nguyen, A. (2021). Multiple meta-model quantifying for medical visual question answering. *Comp. Vis. Patt. Recogn.*, (i):1-5. <https://doi.org/10.48550/arXiv.2105.08913>
- Evans, H.P., Anastasiou, A., Edwards, A., Hibbert, P., Makeham, M., Luz, S., Sheikh, A., Donaldson, L., and Carson-Stevens, A. (2020). Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. *Health Inform. J.*, 26(4):3,123-3,139. doi: 10.1177/1460458219833102.
- He, J., Fu, M., and Tu, M. (2019). Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC Med. Inform. Decis. Mak.*, 19(Suppl 2):52. doi: 10.1186/s12911-019-0761-8.
- He, P., Liu, X., Gao, J., and Chen, W. (2020a). DeBERTa: decoding-enhanced BERT with disentangled attention. *Comp. Langu.*, p. 1-21. <https://doi.org/10.48550/arXiv.2006.03654>.
- He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. (2020b). PathVQA: 30000+ questions for medical visual question answering. *Comp. Langu.*, p. 1-12. <https://doi.org/10.48550/arXiv.2003.10286>.
- Hoermann, S., McCabe, K.L., Milne, D.N., and Calvo, R.A. (2017). Application of synchronous text-based dialogue systems in mental health interventions: systematic review. *J. Med. Int. Res.*, 19(8):e267. doi: 10.2196/jmir.7023.
- Jackson, R.G., Patel, R., Jayatileke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R.J., and Stewart, R. (2017). Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (CRIS-CODE) project. *BMJ Open* 7(1):e012012. doi: 10.1136/bmjopen-2016-012012.
- Karri, S.P.R. and Kumar, B.S. (2020). Deep learning techniques for implementation of chatbots. In: 2020 International Conference on Computer Communication and Informatics, (ICCCI) p. 1-5. doi: 10.1109/ICCCI48352.2020.9104143.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *Comp. Langu.*, p. 1-17. <https://doi.org/10.48550/arXiv.1909.11942>.
- Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., and Yu, H. (2006). Beyond information retrieval--medical question answering. *AMIA ... Annual Symposium Proceedings / AMIA Symposium*. AMIA Symposium, p. 469-473.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020a). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 7,871-7,880. doi: 10.18653/v1/2020.acl-main.703.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020b). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Comp. Langu.*, p. 1-19. <https://doi.org/10.48550/arXiv.2005.11401>.
- Liang, H., Tsui, B.Y., Ni, H., Valentim, C.C.S., Baxter, S.L., Liu, G., Cai, W., Kermany, D.S., Sun, X., Chen, J., He, L., Zhu, G., Tian, P., Shao, H., Zheng, L., Hou, R., Hewett, S., Li, G., Liang, P., Zang, X., Zhang, Z., Pan, L., Cai, H., Ling, R., Li, S., Cui, Y., Tang, S., Ye, H., Huang, X., He, W., Liang, W., Zhang, Q., Jiang, J., Yu, W., Gao, J., Ou, W., Deng, Y., Hou, Q., Wang, B., Yao, C., Liang, Y., Zhang, S., Duan, Y., Zhang, R., Gibson, S., Zhang, C.L., Li, O., Zhang, E.D., Karin, G., Nguyen, N., Wu, X., Wen, C., Xu, J., Xu, W., Wang, B., Wang, W., Li, J., Pizzato, B., Bao, C., Xiang, D., He, W., He, S., Zhou, Y., Haw, W., Goldbaum, M., Tremoulet, A., Hsu, C.N., Carter, H., Zhu, L., Zhang, K., and Xia, H. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Med.*, 25(3):433-438. DOI: 10.1038/s41591-018-0335-9.
- Logeswaran, L. and Lee, H. (2018). An efficient framework for learning sentence representations. In: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, p. 1-16.
- Min, S., Chen, D., Zettlemoyer, L., and Hajishirzi, H. (2019). Knowledge guided text retrieval and reading for open domain question answering. *Comp. Langu.*, p. 1-12. <https://doi.org/10.48550/arXiv.1911.03868>.
- Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., and Tran, Q.D. (2019). Overcoming data limitation in medical visual question answering. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11767 LNCS:522-530. doi: 10.1007/978-3-030-32251-9_57.
- Omorgbe, N.A.I., Ndaman, I.O., Misra, S., Abayomi-Alli, O.O., and Damaševičius, R. (2020). Text messaging-based medical diagnosis using natural language processing and fuzzy logic. *J. Health. Eng.*, doi: 10.1155/2020/8839524.
- Parikh, A.P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. In: EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, p. 2,249-2,255. doi: 10.18653/v1/d16-1244.
- Peters, M.E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1:1,756-1,765. doi: 10.18653/v1/P17-1161.
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1:2,227-2,237. doi: 10.18653/v1/n18-1202.
- Putra, F.B., Yusuf, A.A., Yulianus, H., Pratama, Y. P., Humairra, D.S., Erifani, U., Basuki, D.K., Sukaridhoto, S., and Budiarti, R.P.N. (2019). Identification of symptoms based on natural language processing (NLP) for disease

- diagnosis based on international classification of diseases and related health problems (ICD-11). In: 2019-International Electronics Symposium: The Role of Techno-Intelligence in Creating an Open Energy System Towards Energy Democracy, Proceedings, p. 1-5. Doi: 10.1109/ELECSYM.2019.8901644.
- Qiu, X. and Huang, X. (2015). Convolutional neural tensor network architecture for community-based question answering. In: IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua(Ijcai):1,305-1,311.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *Comp. Langu.*, p. 2-6. <https://doi.org/10.48550/arXiv.1910.01108>
- Zhang, X., Wu, J., He, Z., Liu, X., and Su, Y. (2018). Medical exam question answering with large-scale reading comprehension. In: 32nd AAAI Conference on Artificial Intelligence, AAAI, p. 5,706-5,713. doi: <https://doi.org/10.1609/aaai.v32i1.11970>.
- Zhou, X., Wu, B., and Zhou, Q. (2018). A depth evidence score fusion algorithm for chinese medical intelligence question answering system. *J. Health. Eng.*, 2018:1205354. doi: 10.1155/2018/1205354. PMID: 30123438; PMCID: PMC6079581.