

NLP Solutions on Articles

Murat Çelik

b21827263@cs.hacettepe.edu.tr

Department of Computer Engineering,

Hacettepe University

Ankara, Turkey

ABSTRACT

Along with the technologies in recent years, artificial intelligence technologies have affected our lives in many areas. Research on artificial intelligence studies in the field of health continues rapidly. One of these areas is Natural language processing (NLP). Mining studies in clinical documents; diagnosis of the disease by searching many sources; identification of disease through automatic labeling of patient complaints; answering patients' questions via chatbot; They are divided into different tasks such as summarizing, tagging and titling texts. In this project, we made use of these algorithms by creating a website. The development stages of the project, the facilities it provides and the details are documented. The focus of the study is to make it easier for people far from the field of health to be informed and to save time.

KEYWORDS

NLP, Summarization, Keyword, Question-Answer

1 INTRODUCTION

With the technologies in recent years, artificial intelligence technologies have affected our lives in many areas. The research of artificial intelligence studies in the field of health continues rapidly and even as a result of recent developments, it has even started to be used in the decision-making mechanism. The part of these technologies that focuses on text is called Natural language processing (NLP). Mining studies in clinical documents; diagnosis of disease by searching many sources; identification of disease by automatic labeling of patient complaints; answering patients' questions via chatbot; They are divided into different tasks such as summarizing, tagging and titling texts. With these developments, it is important to increase success in the field of health, to put less burden on health workers and to disseminate correct information.[6][4]

2 PROBLEM

The focus in this project is to apply Natural Language Processing solutions on health articles. The main purpose is to create a more simplified and practical website for people who cannot find time, miss important parts, and cannot improve themselves in research. On the website developed within the scope of the project, additional studies were carried out and a demo was prepared in order to appeal to the end user with plain texts.

2.1 Dataset

1198 plain health articles published by MedicalNewsToday were used. The articles are focused on different topics and have a length of between 4000-10000 characters. Article data can be accessed at the following address. 2k clean medical articles (MedicalNewsToday)

3 METHOD

The project focused on an accessible and easy-to-use website. Long articles with an average of 7000 characters are not read and understood by most people. This was an incentive to use the Natural Language Processing field to solve the problem. These articles were handled as titling, article tagging and summarizing. This showed that user accessible areas can be extracted from a long plain text. In addition, a scenario was designed where the user can ask a question about an article he/she chooses. It was expected that the model would answer the question by researching it through the article. This provided an interactive scenario with the user.

3.1 Title Generation

When a user wants to read an article, the first thing they notice is the title. The model has been prepared in order to identify the most important parts of the article itself and to bring it to the title format. Fabiochui's pretrained model named "t5-small-medium-title-generation" was used. The model was trained with 16000 articles and 1000 articles that were not used in the train stage were used in the evaluation step. A score of 26.9% was obtained in the Recall-Oriented Understudy for Gisting Evaluation-L metric. The ROUGE-L score is the metric type that targets the longest clusters by calculations of different metrics.

3.2 Summarization

Summarizing allows us to present the shorter version of the texts without losing their informational substance. There are two types of paths in these algorithms. One is the algorithm prepared by extracting sentences from the article called Extractive. The other model is the algorithms called Abstractive/Generative and aiming to create completely new text. It was studied with the pretrained model of Sshleifer named "distillbart-cnn-12-6". This model is in the architecture of the BART model in the article [2]. It is an extractive model.

3.3 Keyword Extraction

Determining the keywords of the texts is important for categorization. In this way, the user can search for an article in a desired area and examine the options. KeyBERT is an algorithm defined on the BERT model and its purpose is to extract and score keywords. The model has some shortcomings. Since the same words were scored with different suffixes, a Lemmatization application was made on the text. For each article, 5 popular tags were determined and categorized.[1][3]

3.4 Question Answer

The main scenario of the study is a question-answer study. This study is to answer the question received from the user on the article. In our project, this scenario is realized in two ways. The first scenario is that the user can ask a question from an article page and get an answer. In the second scenario, the user selects the question and a keyword, the answer is searched in all articles related to that keyword, and the answer with the highest answer score is shown to the user.

The model is based on the "distillbert-base-cased-distilled-squad" model. It has an extractive structure. The answer to the question asked is searched in the text and after the answer is found, it is scored. The algorithm tells us the answer, the score and the position of the answer in the text. [5] [7]

4 PROJECT DETAILS

The project is built on Django architecture. The libraries and versions of the project are given in the "requirements.txt" within the project. Basically, 3 pages were determined in the project.

One of them is the main page (Figure:2) where the articles are published in catalog. On this page, articles are shown with titles, abstracts and tags published by NLP models. The articles are also cataloged through the selected tags.

Another page is a specially created content page (Figure:1) for each article. Here, the title of the article, the article itself, the article summary and tags are published. In addition, there is a question bar on the page. Here, the user asks a question over the article and the answer is automatically observed on the screen.

Another page is the question and answer panel(Figure:3). Here the user writes his question and selects a tag. The answer to the question is searched on the articles with that tag and the best answer is returned to the user. The main purpose of this page is to create a platform where the user can get information without going to the doctor.

5 PROJECT RESULTS

The overall results of the project are positive. The hashing, keyword extraction, and titling algorithms work well. The question and answer area also provides a nice experience.

In the summarization algorithm, the performance of the article decreases with the length of the article. Due to the length of some articles, the model cannot work and as a solution, the article is given to the model piece by piece.

Due to the length of the articles in the question and answer area, there may be difficulties in answering. It is seen that the model has difficulty in general questions. More specific questions are more successful. For example, "What is the name of the abortion pill?" answers the question as "Misoprostol". But the general question "What is abortion?" He answers the question "It occurs when the pregnancy ends". In general, the answers are in line with the clarity of the question.

6 CONCLUSIONS

The use of Natural Language Processing in the field of health has been determined as the main theme. The main purpose is to establish and present an accessible, easy and fast health platform to the

user. It has been determined that the results are generally positive and its use will also be beneficial. Since it is aimed to add new articles to the system dynamically, it has been an efficient work in terms of usage and maintenance. Some changes and additions are planned in the later stages of the project. Interface correction, a faster and more stable site experience is aimed. In the next stage of the study, it is aimed to add another scenario. Within the scope of this scenario, it is aimed to add user comments to the platform and tag them. These labels; The relationship between the article and the comment will be on determining the accuracy and usefulness of the comment, and filtering false information. These considerations are important for future work.

As a result of the advancement of technology and new ideas, it is certain that artificial intelligence technology will achieve greater success in the field of health in the future. It is also seen that the work in this area will continue without slowing down.

REFERENCES

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018). <https://doi.org/10.48550/ARXIV.1810.04805>
- [2] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. (2019). <https://doi.org/10.48550/ARXIV.1910.13461>
- [3] Piotr Pęzik, Agnieszka Mikołajczyk-Bareła, Adam Wawrzyński, Bartłomiej Nitoń, and Maciej Ogrodniczuk. 2022. Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer. <https://doi.org/10.48550/ARXIV.2209.14008>
- [4] Ana Sabina Uban and Cornelia Caragea. 2021. Generating Summaries for Scientific Paper Review. (2021). <https://doi.org/10.48550/ARXIV.2109.14059>
- [5] Zhen Wang. 2022. Modern Question Answering Datasets and Benchmarks: A Survey. (2022). <https://doi.org/10.48550/ARXIV.2206.15030>
- [6] Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2021. Natural Language Processing for Smart Healthcare. (2021). <https://doi.org/10.48550/ARXIV.2110.15803>
- [7] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. <https://doi.org/10.48550/ARXIV.2101.00774>

Articles
Question Answer Page

Medicare Covers the Costs of Screening Coloscopies at Specific Time intervals, based on a person's risk for colon cancer

Medicare covers the costs of screening colonoscopies at specific time intervals, based on a person's risk for colon cancer. For those with Medicare, the test is usually free. However, a person may have to pay out-of-pocket costs if they need polyp removal or use anesthesia services for their colonoscopy. According to the Centers for Disease Control and Prevention (CDC), 15 million colonoscopies took place in 2012 across the United States. A screening colonoscopy plays a vital role in identifying colon cancer and providing a person with treatment opportunities early in the progression of the disease. The CDC also note that health authorities in the US. are currently aiming to perform screening for 80% of people between 50 and 75 years of age by the year 2024. This target may not be achievable, however, as many people cannot afford a colonoscopy. To remedy this, Medicare may make colon cancer screening more accessible for those with an increased colon cancer risk, such as those aged 50–75 years. In this article, we explain how Medicare covers colonoscopy costs and what a person can expect to pay out-of-pocket for the procedure. Medicare covers screening colonoscopy costs as long as the doctor "accepts assignment." Accepting assignment means the doctor who will perform the colonoscopy agrees to Medicare reimbursing them at a standard rate for screening colonoscopies. Medicare will cover screening colonoscopies at the following intervals: Once every 24 months: This interval is for people who have a high risk of colorectal cancer due to a family or personal history of the disease. , Once every 120 months, or 10 years: Medicare will fund this after a previous colonoscopy screening or every 48 months after a person has had a flexible sigmoidoscopy. This is an examination in which the doctor inserts the colonoscope into the sigmoid colon but no deeper. . If a doctor accepts assignment and does not view or remove polyps during a colonoscopy, a person with Medicare does not pay anything for the test. Polyps are growths in the lining of the rectum and colon. While many polyps are not cancerous in the beginning, they may develop into colon cancer over time. It is challenging for a doctor to predict the presence of polyps before a colonoscopy, and they are usually so tiny that a person will not be aware of them. For this reason, colon cancer screenings are vital for identifying polyps. An estimated one-third of people receiving a colonoscopy do so under anesthesia, according to the

Summary of Article

Medicare covers colonoscopies at specific intervals, based on a person's risk for colon cancer . For those with Medicare, the test is usually free for a person with Medicare . However, a person may have to pay out-of-pocket costs if they need polyp removal or use anesthesia services for their colonoscopy . A colonoscopy can help a doctor screen for colon cancer and remove polyps . The ability to identify and remove precancerous growths before they grow and become malignant is vital to colorectal cancer prevention . Not all colonoscopies are for screening but can be a diagnostic procedure .

KeyWords : [#colonoscopy](#) [#screening](#) [#medicare](#) [#colonoscopy](#) [#medicaid](#)

Ask a question to AI

The answer of AI

Once every 24 months

Figure 1: This is the page where the article details appear. The title is written in blue font. Below is the original text. The summary and keywords of the text are given on the right. Below them, there is an area to ask questions about the article. As seen in the example, the answer to the question is given below.

Articles
Question Answer Page

Menu
cancer
coronavirus
dietary
inflammation
arthritis
covid
diet
disease
bowel
cardiovascular

Chronic obstructive pulmonary disease, or COPD, is a group of chronic diseases that make breathing difficult.

Chronic obstructive pulmonary disease, or COPD, is a group of chronic lung diseases . It is a progressive condition that gets worse over time . COPD has a range of effects on the lungs that reduce their ability to take in oxygen . The diseases that make up COPD include emphysema, chronic bronchitis, and refractory asthma . Physiology describes the changes a disease or condition causes in a person's physical function . Inhaling any pollutant can cause COPD, whether it is cigarette smoke, industrial chemicals, cooking fumes, or heavy air pollution . Genetics may also play a role in the development of COPD . People who smoke tend to have more exacerbations than those who do not . People with refractory asthma cannot return the airways to their natural state .


How doctors work out the life expectancy for people with COPD

COPD is a term for several chronic health conditions that reduce lung function . The outlook for a person with COPD depends on the stage of the disease and their overall health . COPD causes airflow obstruction, impacting a person's ability to get enough oxygen into their lungs and move it through their body . There is no cure for COPD, but medicines can help to reduce severe symptoms . A person with COPD is likely to experience episodes in which their symptoms suddenly become worse . These attacks are known as COPD exacerbations . Exacerbations may require different drugs, hospitalization, or ventilator support until a doctor can control the flare . Pulmonary rehabilitation involves sessions with a respiratory therapist or lung specialist .

KeyWords : [#copd](#) [#pulmonary](#) [#asthma](#) [#respiratory](#) [#lung](#)

KeyWords : [#copd](#) [#pulmonary](#) [#respiratory](#) [#lung](#) [#expiratory](#)

Figure 2: This is the area where the articles are displayed. You can access the articles of that tag by selecting a tag on the left. When the option bar on the left is removed, the title, summary and tags of the text are displayed on the page. This information is completely extracted by algorithm. There are 3 articles on each page.

Articles Question Answer Page Search 

Ask AI

With a selected tag, your question will be reviewed and answered in all articles in that field.

Your Question is here:

You can choose your tag in here:

Submit

Answer of AI

Exactly 54 articles were reviewed. And found the best answer.

The question is : What is the cause of cancer?

The answer is : high cholesterol

Figure 3: This is the Question Answer Page. The user writes a question in the left input field. The user selects a tag in the right input field. The user sees the answer in the field below. In the sample picture, the question is scanned from 54 different articles and given.