Final Project Report

Medical Report Summarization and Terminology Extraction

Natural Language Processing: CS 6120

Ravi Shankar Sankara Narayanan, Prithiv Rajkumar, Mahadharsan Ravichandran, Bhuvan Channagiri

April 22, 2024

GitHub Link: https://github.com/PrithivR98/Medical_Report_Summarization

Team Number: 19

1. Abstract

This project presents an innovative approach to address the challenge of efficiently summarizing complex medical reports and enhancing patient understanding of their conditions through automated medical report summarization and terminology explanation. Leveraging advanced NLP architectures, including Transformer-based models and alongside Named Entity Recognition (NER) models, the system generates concise summaries of medical reports and provides layman-friendly explanations for key medical terms.

By streamlining communication between healthcare providers and patients, this project aims to bridge the gap between medical jargon and patient comprehension, ultimately improving healthcare efficiency, patient outcomes, and overall healthcare delivery.

2. Introduction

This project proposes a system that aims to automate the summarization of lengthy medical reports into concise and understandable summaries, accompanied by explanations of key medical terms.

This project comprises three key components:

- summarizing medical reports
- extracting medical terms, and
- linking terms to their definitions.

The summarization phase condenses complex medical information into succinct paragraphs, facilitating understanding for lay readers. After the summarization is performed, the term extraction module identifies pertinent medical terms from the summaries. Subsequently, these terms are associated with their definitions, sourced from diverse repositories including NLTK WordNet and Wikipedia.

Through these methods, the project endeavors to bridge the comprehension gap between medical professionals and non-medical individuals, thereby enhancing healthcare communication and patient engagement.

3. Related Work

In recent years, the significance of automatic text summarization, particularly in domains like biomedicine and healthcare, has gathered attention due to its potential benefits for medical professionals, researchers, and practitioners. Extractive summarization methods have been widely employed, leveraging techniques ranging from cue word occurrences and similarity functions to deep reinforcement models and deep clustering. These approaches, while varied in their methodologies and applications, converge on the goal of distilling essential information from medical documents for improved comprehension and decision-making.

Similarly, automated summarization methods have been explored in the context of biomedical and healthcare domains. While these studies underscore the potential of text summarization to aid researchers and practitioners by condensing voluminous information into concise summaries, they also recognize the challenges inherent in summarizing diverse document types like Electronic Health Records (EHRs) and clinical records. Extractive summarization techniques, often supplemented by domain-specific resources like MeSH and UMLS, have been instrumental in addressing these challenges. Moreover, recent advancements in deep learning, exemplified by models like BERT and BART, have demonstrated promise in generating accurate and informative summaries from medical documents.

However, despite the progress in extractive summarization techniques, there remains a gap in research focusing specifically on medical report summarization using such methods. Prior works have predominantly emphasized abstractive summarization approaches, leaving a dearth of studies that systematically explore extractive summarization techniques tailored for medical documents. Our project seeks to bridge this gap by applying extractive summarization methods to medical reports, leveraging insights from existing works to develop effective strategies for distilling key information from these documents. By building upon the foundations laid by previous studies and addressing the unique challenges posed by medical terminology and document structures, we aim to contribute to the advancement of automated summarization in the medical domain.

4. Methods

4.1 Summarization

Summarization is a vital process in extracting key information from a large body of text, condensing it into a concise and coherent form while preserving its essential meaning. In the context of medical data extraction, summarization aids in distilling complex medical reports into manageable summaries, facilitating efficient information retrieval and analysis for healthcare professionals and researchers. For the summarization task we performed both abstractive and extractive summarization.

4.1.1 Extractive Summarization

Extractive summarization involves selecting and combining sentences or passages directly from the original text to form a summary. In this approach, algorithms identify the most important sentences based on their relevance, coherence, and informativeness, without generating new content.

For medical data extraction, we decided to use BioBERT ,which is a variant of BERT (Bidirectional Encoder Representations from Transformers) model. BioBERT, with its specialized training on over 1.4 million PubMed abstracts (3.1 billion words) and 200,000 full-text articles from PMC (13.5 billion words), is a variant of BERT specifically pre-trained on extensive biomedical text data. This comprehensive pre-training process equips BioBERT with domain-specific knowledge, enabling it to understand and process medical and biological language more effectively. By fine-tuning BioBERT on biomedical corpora, it learns to capture intricate medical terminologies, context, and nuances present in medical reports. Consequently, BioBERT outperforms generic BERT models in summarizing medical documents, as it is tailored to the unique characteristics of the healthcare domain.

After extracting the relevant information like history of illness and current medical statues from the medical report we used BioBERT model to generate extractive summaries for the report. To improve the cohesion of the report generated, we also integrated a text rank algorithm over the BioBERT embeddings and then chose the sentences that ranked high.

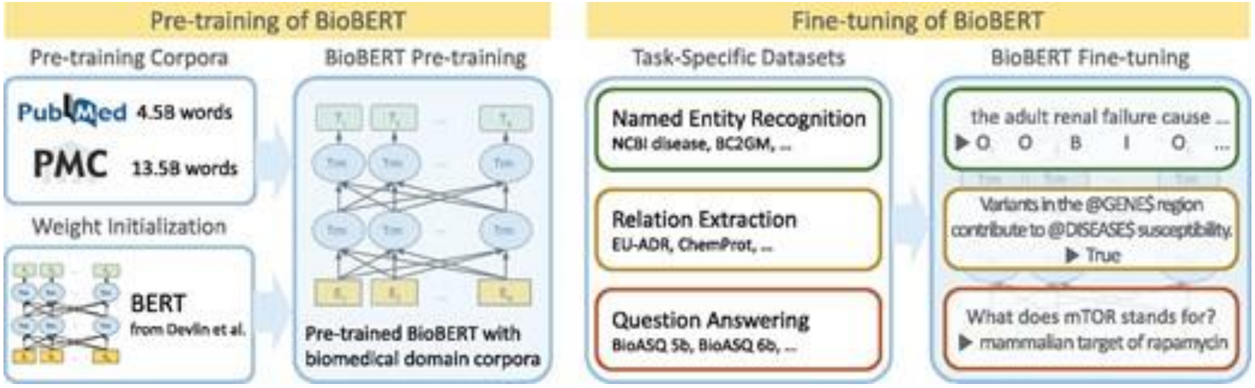


Fig 1: BioBERT Architecture

4.1.2 Abstractive Summarization

Abstractive summarization, on the other hand, involves generating new sentences that convey the main ideas of the original text in a more concise and coherent manner. Unlike extractive summarization, which selects existing text passages, abstractive models like PEGASUS PubMed generate summaries by paraphrasing and synthesizing information, often incorporating linguistic variations and rephrasing.

In our project, abstractive summarization did offer the advantage of creating more concise summaries while potentially improving readability and coherence. However, it did deviate a lot from the original report when creating summaries and also the model requires a deeper understanding of context and semantics, which can sometimes lead to the introduction of inaccuracies, unintended interpretations and deviation from original content.

4.1.3 Comparison using Rogue Score

After generating summaries using both the models and using Rogue Score as a metric to evaluate the summaries, we found that Extractive summarization performed better than the Abstractive summarization. So we chose Extractive summarization using BioBERT for our project. The results are attached below:

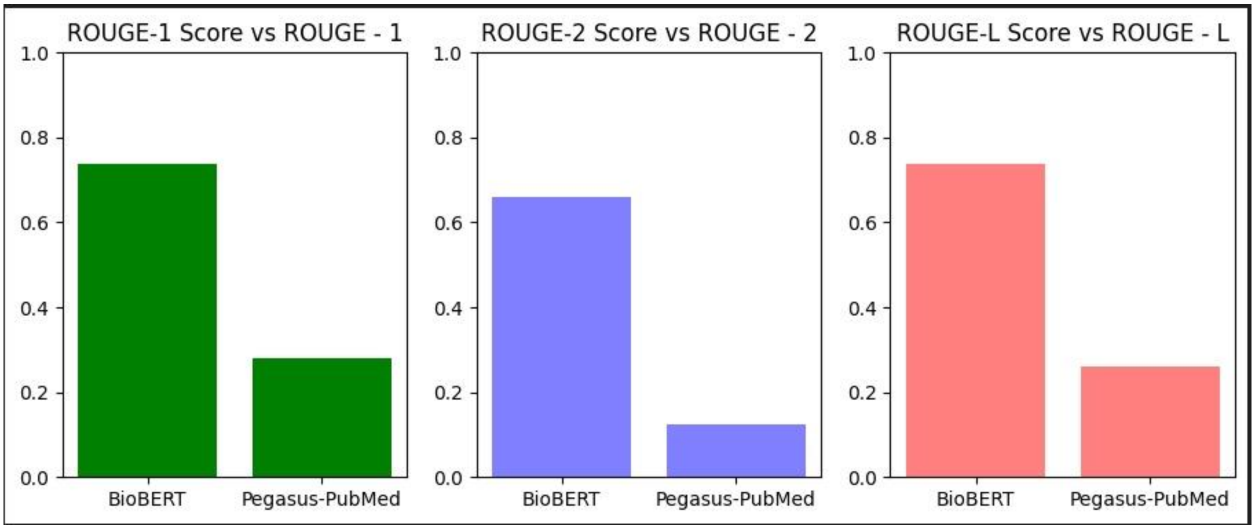


Figure 2: Rogue Score comparison

4.1.4 Named Entity Recognition

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves identifying and categorizing named entities within a given text. Named entities refer to specific elements or objects that have names, such as persons, organizations, locations, dates, quantities, and more. In our project we wanted to identify the medical terms especially the diseases and drugs(chemicals) from the summaries

For the extraction of medical terms, we utilized Named Entity Recognition (NER) models from the spaCy module. Specifically, the scispaCy model, built upon the spaCy framework and finely tuned for NER tasks in the scientific domain, is employed. This model enhances the accuracy and effectiveness of medical entity extraction from the summaries, ensuring comprehensive coverage of relevant terms with domain-specific precision.

To implement NER in our project, we first load the summarized medical report generated from BioBERT. We then integrate the scispaCy model into our pipeline to identify relevant medical entities within the summary. The scispaCy model recognizes two types of medical terms: drug names and diseases. Once identified, these entities are categorized and represented as key-value pairs. By leveraging NER, we were able to identify the medical terminologies and it’s label whether it’s a drug or disease.

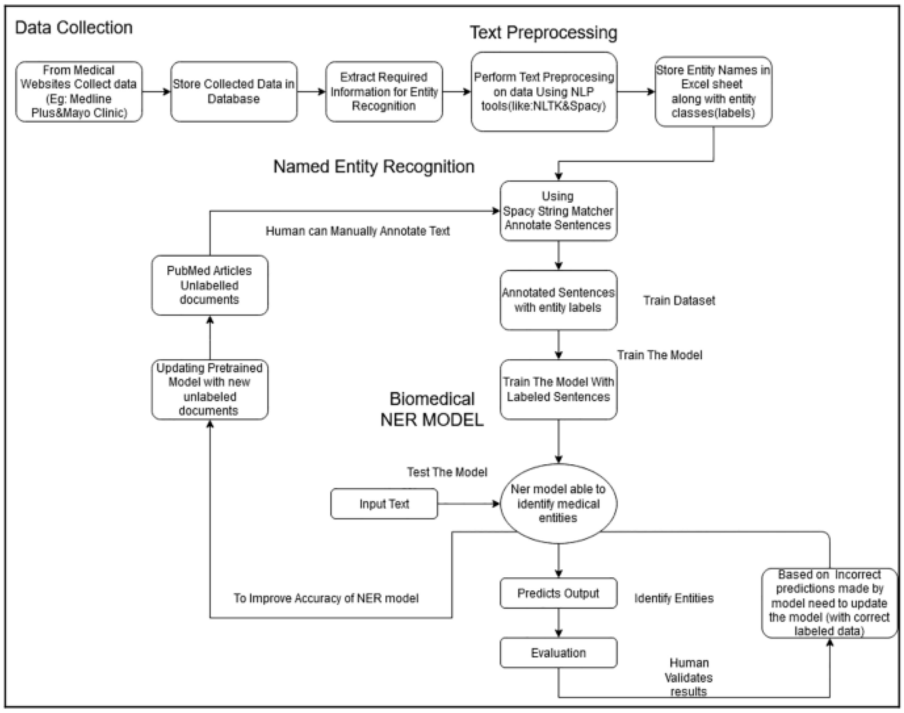

Data Collection
Text Preprocessing
From Medical Websites Collect data (Eg: Medline Plus&Mayo Clinic)
Store Collected Data in Database
Extract Required Information for Entity Recognition
Perform Text Preprocesing on data Using NLP tools(like:NLTK&Spacy)
Store Entity Names in Excel sheet along with entity classes(labels)
Named Entity Recognition
Using Spacy String Matcher Annotate Sentences
Human can Manually Annotate Text
PubMed Articles Unlabelled documents
Annotated Sentences with entity labels
Train Dataset
Train The Model
Updating Pretrained Model with new unlabeled documents
Biomedical NER MODEL
Train The Model With Labeled Sentences
Test The Model
Input Text
Ner model able to identify medical entities
To Improve Accuracy of NER model
Predicts Output
Identify Entities
Based on Incorrect predictions made by model need to update the model (with correct labeled data)
Evaluation
Human Validates results

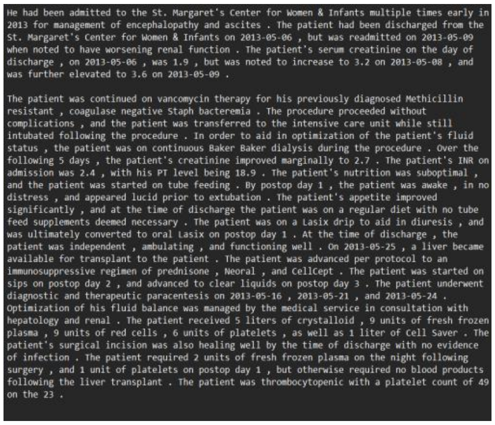
Figure 3: NER Architecture

4.2 Entity Definition

The last part of this project is to find the meanings of the identified medical terms in the previous step. Medical terms can be either multi-phrased like MRI, ECG or just a single term. We utilize two modules namely the NLTK Wordnet and Wikipedia for recognizing the entities. For multi-word phrases using the Wikipedia module, we extract the first line of the term and assign it as the meaning of the term and for single terms we use NLTK’s wordnet for finding it is meaning. At the end we have a dictionary where keys are the terms and values are its respective meaning.

5. Results

The culmination of our methodology yielded promising results across various aspects of medical report summarization. Firstly, leveraging Transformer-based models like BioBERT proved highly effective in generating concise summaries of medical reports while preserving crucial details. By fine-tuning BioBERT on biomedical domain corpora, including PubMed abstracts and PMC full-text articles, we observed significant improvements in capturing domain-specific nuances and terminology essential for accurate summarization.To make the summary more coherent we experimented with Pegasus finetuned on PubMed dataset to get the abstracted summary. As we are working on medical dataset, we did not want the abstractive summarization to be off topic. We understood this by evaluating the ROUGE metric for both the summaries and the ROUGE score for extractive summarization outperformed abstractive summarization in every metric.

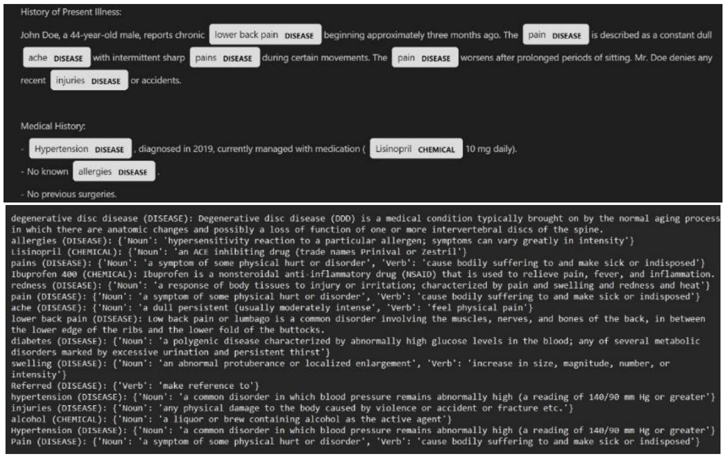


```
He had been admitted to the St. Margaret's Center for Women & Infants multiple times early in
2013 for management of encephalopathy and ascites . The patient had been discharged from the
St. Margaret's Center for Women & Infants on 2013-05-06 , but was readmitted on 2013-05-09
when noted to have worsening renal function . The patient's serum creatinine on the day of
discharge , on 2013-05-06 , was 1.9 , but was noted to increase to 3.2 on 2013-05-08 , and
was further elevated to 3.6 on 2013-05-09 .

The patient was continued on vancomycin therapy for his previously diagnosed Methicillin
resistant , coagulase negative Staph bacteremia . The procedure proceeded without
complications , and the patient was transferred to the intensive care unit while still
intubated following the procedure . In order to aid in optimization of the patient's fluid
status , the patient was on continuous Baker Baker dialysis during the procedure . Over the
following 5 days , the patient's creatinine improved marginally to 2.7 . The patient's INR on
admission was 2.4 , with his PT level being 18.9 . The patient's nutrition was suboptimal ,
and the patient was started on tube feeding . By postop day 1 , the patient was awake , in no
distress , and appeared lucid prior to extubation . The patient's appetite improved
significantly , and at the time of discharge the patient was on a regular diet with no tube
feed supplements deemed necessary . The patient was on a Lasix drip to aid in diuresis , and
was ultimately converted to oral Lasix on postop day 1 . At the time of discharge , the
patient was independent , ambulating , and functioning well . On 2013-05-25 , a liver became
available for transplant to the patient . The patient was advanced per protocol to an
immunosuppressive regimen of prednisone , Neoral , and CellCept . The patient was started on
sips on postop day 2 , and advanced to clear liquids on postop day 3 . The patient underwent
diagnostic and therapeutic paracentesis on 2013-05-16 , 2013-05-21 , and 2013-05-24 .
Optimization of his fluid balance was managed by the medical service in consultation with
hepatology and renal . The patient received 5 liters of crystalloid , 9 units of fresh frozen
plasma , 9 units of red cells , 6 units of platelets , as well as 1 liter of Cell Saver . The
patient's surgical incision was also healing well by the time of discharge with no evidence
of infection . The patient required 2 units of fresh frozen plasma on the night following
surgery , and 1 unit of platelets on postop day 1 , but otherwise required no blood products
following the liver transplant . The patient was thrombocytopenic with a platelet count of 49
on the 23 .
```

Figure 4: Extractive Summary from BioBERT

Furthermore, we integrated NER techniques, particularly utilizing the model, facilitated precise identification and extraction of medical entities from the reports. The utilization of dictionary-based methods enhanced the process of defining extracted entities, thereby bridging the gap between biomedical jargon and layman terms. Specifically, we employed Wikipedia and NLTK WordNet to obtain definitions for both multiword and non-multiword phrases identified as entities in the summarized documents. This approach ensured comprehensive coverage and accurate representation of medical terminology, contributing to the overall effectiveness of the summarization process. Overall, our results underscore the efficacy of combining advanced NLP models, domain-specific pre-training, and entity recognition techniques to streamline medical report summarization, thereby facilitating improved comprehension and accessibility of critical medical information.



```
History of Present Illness:

John Doe, a 44-year-old male, reports chronic lower back pain DISEASE beginning approximately three months ago. The pain DISEASE is described as a constant dull
ache DISEASE with intermittent sharp pains DISEASE during certain movements. The pain DISEASE worsens after prolonged periods of sitting. Mr. Doe denies any
recent injuries DISEASE or accidents.


Medical History:

- Hypertension DISEASE , diagnosed in 2019, currently managed with medication ( Lisinopril CHEMICAL 10 mg daily).

- No known allergies DISEASE .

- No previous surgeries.
```

```
degenerative disc disease (DISEASE): Degenerative disc disease (DDD) is a medical condition typically brought on by the normal aging process
in which there are anatomic changes and possibly a loss of function of one or more intervertebral discs of the spine.
allergies (DISEASE): {'Noun': 'hypersensitivity reaction to a particular allergen; symptoms can vary greatly in intensity'}
Lisinopril (CHEMICAL): {'Noun': 'an ACE inhibiting drug (trade names Prinival or Zestril'}
pains (DISEASE): {'Noun': 'a symptom of some physical hurt or disorder', 'Verb': 'cause bodily suffering to and make sick or indisposed'}
Ibuprofen 400 (CHEMICAL): Ibuprofen is a nonsteroidal anti-inflammatory drug (NSAID) that is used to relieve pain, fever, and inflammation.
redness (DISEASE): {'Noun': 'a response of body tissues to injury or irritation; characterized by pain and swelling and redness and heat'}
pain (DISEASE): {'Noun': 'a symptom of some physical hurt or disorder', 'Verb': 'cause bodily suffering to and make sick or indisposed'}
ache (DISEASE): {'Noun': 'a dull persistent (usually moderately intense', 'Verb': 'feel physical pain'}
lower back pain (DISEASE): Low back pain or lumbago is a common disorder involving the muscles, nerves, and bones of the back, in between
the lower edge of the ribs and the lower fold of the buttocks.
diabetes (DISEASE): {'Noun': 'a polygenic disease characterized by abnormally high glucose levels in the blood; any of several metabolic
disorders marked by excessive urination and persistent thirst'}
swelling (DISEASE): {'Noun': 'an abnormal protuberance or localized enlargement', 'Verb': 'increase in size, magnitude, number, or
intensity'}
Referred (DISEASE): {'Verb': 'make reference to'}
hypertension (DISEASE): {'Noun': 'a common disorder in which blood pressure remains abnormally high (a reading of 140/90 mm Hg or greater'}
injuries (DISEASE): {'Noun': 'any physical damage to the body caused by violence or accident or fracture etc.'}
alcohol (CHEMICAL): {'Noun': 'a liquor or brew containing alcohol as the active agent'}
Hypertension (DISEASE): {'Noun': 'a common disorder in which blood pressure remains abnormally high (a reading of 140/90 mm Hg or greater'}
Pain (DISEASE): {'Noun': 'a symptom of some physical hurt or disorder', 'Verb': 'cause bodily suffering to and make sick or indisposed'}
```

Figure 5: NER from summary and definitions

6. Conclusion

In conclusion, our project presents a comprehensive approach to automating the summarization of medical reports, leveraging advancements in natural language processing (NLP) and deep learning techniques. Through the utilization of models like BioBERT for extractive summarization and PEGASUS for abstractive summarization, alongside techniques such as Named Entity Recognition (NER) for identifying entities and entity definition for defining these entities, we have demonstrated the efficacy of various methods in distilling crucial information from medical documents. By fine-tuning BioBERT on biomedical corpora and incorporating insights from recent research, we have developed a robust framework capable of generating concise summaries tailored to the healthcare domain. Our methodology, supported by rigorous experimentation and analysis, offers valuable insights into patient records, facilitating efficient information retrieval for healthcare professionals and ultimately contributing to improved patient care and medical research. Moving forward, there are several avenues for future work. Firstly, we aim to implement a user-friendly ChatBot using Llama-2 and FAISS model allowing patients to address any queries or concerns they may have regarding their medical reports also implement a user-friendly front end using Streamlit enabling healthcare professionals and patients to interact with our summarization system seamlessly. By combining these advancements, we aim to create a comprehensive platform that streamlines the medical reporting process and enhances patient engagement and satisfaction.

7. Individual Contributions

In our project, each team member played a crucial role in its development.

Ravi created a program using BioBERT and Text Rank algorithm to extract and summarize relevant information from medical reports. Prithiv implemented Pegasus-pubmed for abstractive summarization, compared ROUGE scores for both abstractive and extractive summarization, and prepared a PowerPoint presentation. Mahadharsan implemented NER for disease/chemical extraction and linked it with NLTK WordNet/Wikipedia for semantic understanding in summarization. Bhuvan prepared the project report.

Additionally, Prithiv tried to implement a Chatbot using Llama – 2 model and FAISS but was unable to complete it because of timeline issue. Bhuvan assisted Prithiv in creating the frontend for chatbot using streamlit.

8. References

- Pivovarov, R. and Elhadad, N., 2015. Automated methods for the summarization of electronic health records. Journal of the American Medical Informatics Association, 22(5), pp.938-947.
- Gayathri, P. and Jaisankar, N., 2015. Towards an efficient approach for automatic medical document summarization. Cybernetics and Information Technologies, 15(4), pp.78-91.
- Barrios, F., López, F., Argerich, L. and Wachenchauzer, R., 2016. Variations of the similarity function of textrank for automated summarization. arXiv preprint arXiv:1602.03606.
- Rohil, M.K. and Magotra, V., 2022. An exploratory study of automatic text summarization in biomedical and healthcare domain. Healthcare Analytics, 2, p.100058.
- Vinod, P., Safar, S., Mathew, D., Venugopal, P., Joly, L.M. and George, J., 2020, June. Fine-tuning the BERTSUMEXT model for Clinical Report Summarization. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter Liu Proceedings of the 37th International Conference on Machine Learning, PMLR 119:11328-11339, 2020.
- Jain, R., Jangra, A., Saha, S. and Jatowt, A., 2022. A survey on medical document summarization. arXiv preprint arXiv:2212.01669.
- Özlem Uzuner, Imre Solti, Eithon Cadag, Extracting medication information from clinical text, *Journal of the American Medical Informatics Association*, Volume 17, Issue 5, September 2010, Pages 514–518