All Sciences Proceedings

5th International Conference on Applied Engineering and Natural Sciences

July 10-12, 2023 : Konya, Turkey

All Sciences Proceedings
http://as-proceeding.com/

https://www.icaens.com/

A Transformer based approach for Abstractive Text Summarization of Radiology Reports

Bilal Ahmed Khan Balouch [1], Fawad Hussain [2]

[1]*Department of Computer Engineering, University of Engineering and Technology Taxila, Pakistan*
[2]*Department of Computer Engineering, University of Engineering and Technology Taxila, Pakistan*

[1]*(bilal.ahmed2@students.uettaxila.edu.pk)* [2]*(fawad.hussain@uettaxila.edu.pk)*

Abstract – Abstractive text summarization has emerged as a promising approach for generating concise and informative summaries from radiology reports. The topic of research focuses on developing a mechanism for abstractive text summarization specifically tailored for radiology reports, to generate informative summaries from the voluminous and complex information contained within reports. Manual summarization is time-consuming and prone to errors, while automated techniques can save time, reduce human bias, and improve the overall quality of the generated summaries. The mechanism involves preprocessing the text with NLP techniques, utilizing deep learning-based architectures and transformer models, and generating summaries that capture the essence of the original reports in a more concise form. Challenges such as handling complex medical terminologies, incorporating contextual information, and evaluating the quality of generated summaries are important considerations in this mechanism. The potential applications of radiology report summarization include improving report readability, facilitating decision-making, and enabling large-scale data analytics. In this research, Biobart-V2 model is used for summarization and Rouge-L value of 69.42% is being achieved whereas the dataset used is MIMIC III. Further research is needed to address the remaining challenges in this domain and integrate summarization into clinical practice for more effective and efficient radiology report interpretation and patient care.

Keywords – Radiology reports, Abstractive summarization, Transformers, Natural language processing, Transfer learning

I. INTRODUCTION

Radiology reports contain important information about medical imaging findings, but they can be difficult for clinicians to quickly extract relevant details due to their complexity and verbosity. Abstractive text summarization, a specialized area of natural language processing (NLP), can potentially address this issue by generating concise and coherent summaries of radiology reports [1]. However, the details of how abstractive text summarization works for radiology reports are not well understood. Therefore, this re- search proposal aims to thoroughly investigate the intricacies of abstractive text summarization for radiology reports, with the goal of improving clinical decision-making and enhancing patient care.

Despite the potential benefits of abstractive text summarization for radiology reports, the underly- ing mechanism of how it works remains

largely un- explored [2]. Understanding the intricacies of this mechanism is essential for developing effective and accurate summarization models that can be integrated into clinical workflows. It requires investigating the complex interplay between various NLP techniques, such as natural language understanding, semantic representation, and language generation, in the context of radiology reports [3]. Additionally, it involves evaluating the performance of different abstractive summarization algorithms using appropriate metrics, such as ROUGE (Recall-Oriented Understudy for Evaluation) and obtaining user feedback through surveys or interviews to assess the usefulness and usability of the generated summaries in real-world clinical scenarios. The findings of this research will have significant implications for enhancing clinical decision-making and improving patient care. By providing clinicians with concise and relevant summaries of radiology re- ports, abstractive text summarization has the potential to save time, reduce cognitive load, and improve the accuracy and efficiency of clinical decision-making [4]. This can result in better patient outcomes, reduced medical errors, and improved healthcare re- source utilization. Additionally, the insights gained from this research can also contribute to the advancement of the field of NLP and inform the development of similar summarization techniques for other domains of medical literature, further extending the impact of this research beyond radiology reports. De- spite the critical significance of medical summarization, the application of NLP advancements to this task, particularly in the field of radiology [5], is limited. As a result, there is a lack of knowledge about language models specifically trained for summarizing radiology reports. The main aim is the development of a methodology for automatic impression from the finding section of radiology reports. State of art Transformer models are used for summarization and NER (Named Entity Recognition) Model for extracted data.

In this research, we propose novel solutions to address this gap by introducing a state-of-the-art Biobart-V2-based model fine-tuned on the MIMIC III dataset. Dataset preprocessing involves dropping the reports with no impressions. Extracting the meaningful information that is Findings and Impressions from all the reports. Data cleaning using techniques that is Tokenization, punctuation, stop words removal and stemming. Identified the subset of keywords in find- ings. Generating the sequence used in the model. Our model takes multiple fields from free-text radiology reports as input and utilizes a sequence-to-sequence architecture to generate abstract summaries.

The paper is structured as follows we first start by describing the MIMIC III dataset and techniques in- volved in cleaning of dataset. We then proceed to out- line the experiment carried out and present the results.

II. RELATED WORK

Research in healthcare has examined the success of deep learning applications in mainstream NLP tasks like name-entity recognition, semantic role labeling, and part-of-speech tagging. Recently, con- textual embedding structures have gained popularity as they offer customized solutions for NLP tasks [6]. For instance, in a study focused on healthcare domain, researchers generated contextual word embeddings from PubMed articles to improve classification of tweets during disease outbreaks, outperforming general pre-trained embedding models like Word2Vec and Glove [7]. Another relevant study proposed using LSTM-based attentive relation net- works to embed textual risk indicators based on mental disorders, with the goal of generating concise target text via generative or probabilistic search algorithms [8]. In our case, the input consists of FINDINGS and BACKGROUND sections of radiology reports, with the target output being the IMPRESSION paragraph/sentence. This problem can be referred to as a neural abstractive summarization or text sequence generation problem in machine learning literature [9]. Alongside leveraging natural language accuracy metrics such as ROUGE scores for model improvement endeavors, ensuring factual correctness in the predictions is given significant importance [10].

Previous work on the topic of summarizing radiology reports' findings into impressions have covered various approaches, including rule-

based, statis- tical, and machine learning methods . Deep learning- based architectures, such as convolutional neural net- works (CNNs), recurrent neural networks (RNNs), and transformer models, have been compared for their performance in automatic radiology report summa- rization [11]. Extractive and abstractive summarization techniques have been studied, with discussions on their trade-offs in terms of summary quality, read- ability, and coverage of findings [12]. Clinical natural language processing (NLP) techniques, including named entity recognition, relation extraction, and sentence generation, have been reviewed for their application in radiology report summarization [13].

Hybrid text summarization techniques that combine extractive and abstractive methods have also been proposed and evaluated [14]. Potential applications of automatic summarization in radiology practice include improving report readability, facilitating decision-making, and enabling large-scale data analytics. Challenges and limitations, such as the need for annotated datasets, biases in generated summaries, and diverse report formats, have been discussed in the literature. Overall, previous research has contributed to the understanding of various approaches for summarizing radiology reports' findings into impressions and their potential implications in clinical practice.

Abstractive text summarization is a challenging task in natural language processing (NLP) that in- volves generating concise and coherent summaries that capture the essential information from a given input document. Several studies have contributed to the advancements in abstractive text summarization. "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks" (2015) by Rush et al. This study introduces the use of attentive recur- rent neural networks (RNNs) for abstractive sentence summarization. The proposed model incorporates attention mechanisms to focus on important words and phrases while generating concise and coherent summaries.

The research demonstrates improved performance compared to existing methods [15]. "Get To The Point: Summarization with Pointer-Generator Net-works" (2017) by See et al. This study introduces pointer-generator networks, which integrate extractive and abstractive summarization approaches.

The model leverages a hybrid pointer network to copy words from the source document and generate novel words when necessary. The research demonstrates the effectiveness of this approach on various datasets [16]. "Deep Communicating Agents for Abstractive Summarization" (2018) by Celikyilmaz et al. The authors propose a deep communicating agent framework for abstractive summarization. The model utilizes a sequence-to-sequence architecture with an encoder-decoder structure. The encoder processes the input document, and the decoder generates the summary. The research investigates the impact of different communication methods between the encoder and decoder, resulting in improved summarization performance [17]. "Fine-tune BERT for Extractive Summarization" (2019) by Liu and Lapata. This work explores the fine-tuning of BERT, a pre-trained transformer-based model, for extractive summarization. The authors fine-tune BERT using a binary classification objective to identify the salient sentences for summarization. The experiments show promising results in producing informative summaries [18].

Evaluation of radiology report summarization

methods has been another key aspect of previous re- search. Different evaluation metrics have been used, including ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy), which are commonly used metrics for text summarization tasks [19]. However, evaluating the quality of summaries in the context of radiology reports can be challenging, as there may not always be a single correct summary due to the subjective nature of clinical impressions. Some studies have used ex- pert feedback or conducted user studies to assess the clinical relevance, accuracy, and readability of generated summaries.

Moreover, previous research has also explored

the integration of radiology report summarization into clinical workflow and decision support systems. Summarized impressions can be used to generate structured reports, populate electronic health records, and facilitate communication among health- care providers. They can also be used as input for downstream tasks, such as radiology report retrieval, information retrieval, and data analytics. The potential benefits of radiology report summarization include reducing the time and effort required for re- port interpretation, enhancing communication among healthcare providers, and improving decision-making in patient care.

BERT-based models are typically trained on corpora, such as wiki-data and literature datasets, using word-based tokenizers [20]. Tokenization is the process of breaking raw text into smaller chunks, typically words or sentences, called tokens [21]. These tokens are used to understand the context and develop NLP models. Tokenization aids in interpreting the meaning of the text by analyzing the sequence of words. Pre-trained models are large neural net- works that are widely used in various NLP tasks [22]. They follow a pretrain-finetune paradigm, where they are initially trained on a large text corpus and then fine-tuned using additional datasets for specific down- stream tasks. Despite the popularity of common architectures like BERT and T5 [23], they have not been pre-trained on specialized medical corpora. In our research, we have fine-tuned our model using the MIMIC III dataset, which is a publicly avail- able dataset containing free-text radiology reports and structured labels.

We assess the performance of our summarization generation using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric, specifically the F1 score. ROUGE has been widely used as a canonical metric for evaluating summarization tasks, as it has shown good correlation with human-evaluated summaries. In our evaluation, we specifically focus on the ROUGE-L variant, which measures the longest common subsequence overlap between the predicted and reference summaries, providing insights into the informativeness of the generated summaries.

Deep learning refers to a subset of machine learning techniques that involve training artificial neural networks with multiple layers to learn hierarchical representations of data. In the context of NLP, deep learning models, such as recurrent neural networks (RNNs) and transformer models, have been widely employed to capture the complex relationships and dependencies within language. These models excel at capturing contextual information, allowing them to understand the meaning of words and sentences based on the surrounding context. Unlike traditional NLP approaches that treat words in isolation, deep learning models can leverage the entire sentence or even the entire document to derive more accurate representations. Contextual embeddings, on the other hand, refer to word representations that are specific to the context in which they appear.

Traditional word embedding models, such as Word2Vec and GloVe, assign a fixed vector representation to each word, regardless of its context [24]. In contrast, contextual embedding models, such as ELMo (Embeddings from Language Models) and BERT (Bidirectional Encoder Representations from Transformers), generate dynamic embedding that capture the meaning of a word based on its context within a sentence [25]. These models are pre-trained on massive amounts of text data, learning to predict missing words or understand sentence coherence.

Overall, previous research on summarizing radiology reports' findings into impressions has covered a range of approaches, techniques, challenges, and potential applications. It has contributed to the development of automated summarization methods and their integration into clinical practice, with the ultimate goal of improving radiology report interpretation, communication, and patient care. Further research in this area can continue to advance the field and address the remaining challenges to enable more effective and efficient radiology report summarization.

III. PROPOSED METHODOLOGY

The problem addressed in this research is the need for effective abstractive text summarization techniques specifically designed for radiology reports. Radiology reports contain extensive and detailed information regarding medical imaging studies, making them lengthy and time-consuming to read and comprehend. Clinicians and researchers often require a concise summary of the findings and impressions from these reports, which can aid in decision-making, report triaging, and large-scale data analytics. In this research main aim to generate abstractive text summarization and for this purpose the methodology used is divided into following steps.

A. DATA ACQUISITION

The dataset used for model training is MIMIC-III. It was provided by Data Science Innovation Hub (DSIH) Lab, Computer Engineering Department. MIMIC-III includes information such as demographics, vital signs, laboratory results, medications, procedures, diagnoses, and outcomes. The data has been extensively cleaned and structured for research purposes. MIMIC-III has become a valuable resource for researchers in many fields, including clinical decision support, predictive modeling, natural language processing, and epidemiological investigation.

B. DATA CLEANING AND PRE-PROCESSING

MIMIC-III dataset contains reports with the categories of Radiology, Echo, Discharge Summary and Physician. Total count of reports is 20,00000. Table 1 shows the overview of dataset with respect to categories. As per research requirement reports needed for experimentation are Radiology Reports. After Radiology reports taken for experimentation data is further cleaned with the different techniques.

C. TOKENIZATION

Tokenization is a fundamental step in data cleaning and preprocessing. In this technique, text is broken down into smaller units called tokens. Tokens can be individual words, sub words or even characters. After tokenization, data is further cleaned with the help of punctuation and by handling the special character. Figure 1 shows the example of tokenization performed on dataset. In this text is tokenized into individual words. Each word is treated as separate token, resulting in a list of tokens. This tokenization process breaks down the text into smaller units, allowing for further analysis or processing.

D. STOP WORD REMOVAL

In Tokenization removal of stop words have been carried out such as articles and prepositions. Stop words are words that are considered insignificant and do not carry much meaning in given language. After identification of the stop words, tokens have been assigned to individual words and phrases and tokens have been compared against each stop word list. If token matches a stop words then it is removed from the text.

Table 1. Overview of Reports in MIMIC III dataset.

Categories	**Entries**
Nursing/ Other	822497
Radiology	522279
Nursing	223556
ECG	209051
Physician	141624
Discharge Summary	59652
Echo	45794
Respiratory	31739
Nutrition	9418
General	8301
Rehab Services	5431
Social Work	2670
Case Management	967
Pharmacy	103
Consult	98

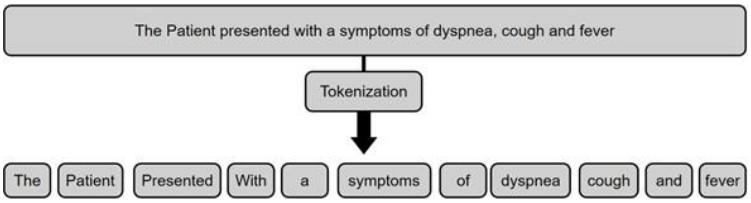

Fig. 1 Tokenization Process On Dataset

E. FEATURE ENGINEERING AND VECTORIZATION

In feature Engineering technique raw data is being transformed into set of features that can be effectively used by machine learning algorithms. Relevant information was extracted and meaningful information was created from the input data. After feature engineering vectorization is applied to text that transform the text data into format that is suitable for machine learning algorithms. Feature Engineering involves transforming the text data into meaningful features that capture the relevant information while in vectorization features converted into numerical representation.

F. IMPLEMENTATION DETAILS

The research revolves around the analysis of radiology reports extracted from the MIMIC-III dataset. This dataset comprises a total of 2,000,000 reports, out of which 586,000 are specifically radiology reports. These radiology reports underwent a thorough cleaning process, resulting in a final set of 1,690,000 reports that were utilized for experimentation purposes. The dataset underwent rigorous cleaning, which involved extensive removal of irrelevant reports lacking impressions. Subsequently, the dataset was processed using general NLP preprocessing techniques, including tokenization, punctuation removal, stop word elimination, and stemming. These steps aimed to extract meaningful information such as findings and impressions from all the reports while ensuring the data's quality and coherence.

Biobart-V2 model was used for summarization on large corpus of data. Biobart-V2 is an architecture designed for report summarization in the biomedical domain. It is an extension of BART (Bidirectional and Autoregressive Transformer), a popular sequence-to- sequence model based on the transformer architecture. Biobart-V2 is pretrained on a large corpus of biomedical text, including scientific articles, clinical notes and their biomedical literature.

Table 2. Hyperparameters For Biobart-V2

Batch size	2
Learning rate	2e - 5
Maximum sequence length	256
Number of epochs	7

This pre- training enables the model to learn domain-specific knowledge and language patterns. Similar to the original BART model, Biobart-V2 follows an encoder- decoder architecture. The encoder takes the input report and encodes it into a fixed-size representation, capturing the contextual information of the text. The decoder then generates a summary by autoregressively predicting the next token based on the encoder's output and previously generated tokens. Biobart-V2 employs masked language modeling during pretraining. This technique involves randomly masking certain tokens in the input text and training the model to predict those masked tokens based on the surrounding context. Masked language modeling helps the model learn to understand and generate coherent and contextually appropriate summaries. After pretraining, Biobart-V2 is further fine-tuned using specific summarization objectives and datasets. This finetuning process adapts the model to the task of summarizing radiology reports. It involves training the model with supervised learning, where pairs of in- put reports and corresponding summaries are used to

optimize the model's performance on generating ac- curate and concise summaries. Biobart-V2 utilizes attention mechanisms between the encoder and de- coder to capture the dependencies and relationships between different parts of the report. This attention mechanism allows the model to focus on relevant in- formation during the summary generation process, improving the quality and coherence of the generated summaries. In this research, Biobart-V2 is trained on 169,0000 reports with 7 epochs. The model's performance was evaluated on the test set using rouge metric. The results of the experiments were compared and analyzed to determine the effectiveness of the Biobart-V2 model for analyzing radiology reports in the MIMIC III dataset. Figure 2 shows the methodology and steps that have been performed while experimentation.

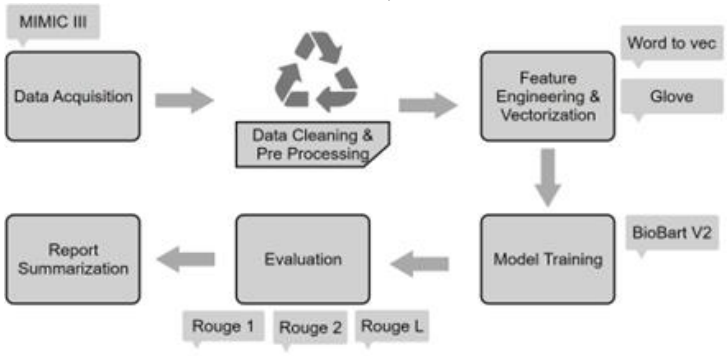

Fig. 2 Process overview for impression finding

G. EVALUATION METHOD:

The result evaluated by the metric called Rouge. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric is commonly used for evaluating and comparing the quality of text summarization systems. It measures the similarity between the system generated summary and one or more reference summaries, which are typically human-generated. Summarization aims to capture the key information from a source text within a concise summary. The ROUGE metric focuses on recall, measuring the overlap between the system-generated summary and the reference summaries. It evaluates the ability of the system to include important information from the source text in the summary. ROUGE calculates similarity using noneless of summarization systems. The ROUGE-L is a version of the ROUGE metric that assesses the quality of a summary generated by a model. It measures the amount of overlap in the longest common subsequence between the machine-generated summary and the reference summary. This overlap indicates how informative the generated summary is compared to the reference summary. ROUGE-L has two components: precision and recall. Precision measures the proportion of the longest common subsequence between the generated summary and the reference summary to the total length of the generated summary. Recall measures the proportion of the longest common subsequence be- tween the generated summary and the reference summary to the total length of the reference summary. The equations for precision and recall consider the maximum length of the longest common subsequence between the reference and candidate summaries, as well as the length of the reference and candidate summaries.

ROUGE has several variants, such as ROUGE-N, ROUGE-L, and ROUGE-S, which focus on different aspects of summarization. ROUGE-N measures n-gram overlap, ROUGE-L considers longest common subsequences, and ROUGE-S evaluates skip-bigram-based similar- ity. These variants provide more nuanced insights into the quality and effectively.

IV. RESULTS AND DISCUSSION:

The final set of results are shown below. It has been shown that Biobart-V2 outperforms all models that have been used. Table 3 shows the results of ROUGE performed against different models. We fine-tuned the following models on larger corpus (169,000 Re- ports with 7 epochs).

TABLE 3. Rouge scores against different models experimented

Models	Rouge - L
Pubmed-PEGASUS	27.23
Textrank	33.86
Biobart	36.19
Bigbird	37.08
Bertsum	56.13
Bert to Bert	52.08

V2 gives the best results of experimentation. Table 4 shows the rouge value against models BART LARGE XSUM and Biobart-V2 whereas Biobart-V2 model performs best.

Figure 3 shows the comparison of all models against the Rouge values. In this research Biobart- V2 outperforms with MIMIC III dataset trained with 7 Epochs.

V. MANUAL EVALUATION:

It is sometimes not possible to access the results generated by the AI models as compared to the reference, like Rouge is good to access the summarization but human for example the radiologists can confirm the efficiency and accuracy of the summarized version. For this we shared the results to a team of radiologists and asked them to access the accuracy of the model. We have 210 evaluations in total: 3 radiologists and 70 reports. We compared the scores provided by the radiologists to determine if they were the equal, better or worse for our model vs. ground truth and our model vs. BART LARGE XSUM.

It has been find out that Biobart-V2 has clearly better than the BART LARGE XSUM: 12.5% of cases are better, 3.13% are worse. Biobart-V2 exceeds the BART LARGE XSUM in 25% (vs. 15.6% “lose”) of evaluations. Biobart-V2 is only slightly worse than ground truth in overall quality (better: 25%, worse: 28.13%). There has been a lot of research on medical document summarization, with various models being developed to tackle the task. These documents can be divided into different categories, each presenting its own unique challenges. While some types of medical documents, such as research articles, radiology reports, and medical dialogue, have been extensively studied, others like electronic health records and consumer health questions are less explored due to difficulties in obtaining datasets for these sub-tasks. The techniques used for medical document summarization can also be categorized based on in- put, output, and method. Most of the current work focus on using a single document as input, generating abstractive summaries as output, and utilizing deep learning or transformer-based models as their foundation. Some approaches also incorporate external knowledge bases, such as medical databases or knowledge graphs, to enhance performance.

Moreover, some works are domain-specific, focusing on medical fields. Recently, there has been a shift to- wards hybrid approaches, combining both extractive and abstractive summarization methods to improve summary faithfulness. Regarding evaluation metrics, it has been observed that standard metrics are inadequate to capture the unique aspects of medical summaries. There is also inconsistency in human evaluation, with different researchers assessing different aspects of the summary. Therefore, it is important to consider all these aspects in human evaluation, along with the use of medical domain-specific metrics. In medical document summarization, it’s important to ensure that the generated summary accurately reflects the true impression of the original document.

Our research has shown that there are cases where the generated impression matches the true impression, while in other cases, there may be different findings but the same true impression. Interestingly, we also found examples where the generated summary has the same findings and true impression as the original document. This highlights the importance of evaluating summaries not only based on their factual accuracy, but also on how well they capture the

overall impression of the original document. By considering both aspects, we can better evaluate the effectiveness of medical document summarization models.

VI. CONCLUSION

In this study, it was investigated utilizing the Biobart- V2 model for text summarization on a vast dataset. The objective was to assess the capability of the model to comprehend medical documents and generate accurate summaries of their findings. To measure the quality of the generated summaries, we employed the ROUGE score, a widely used metric that evaluates the overlap of words between the generated summaries and the reference summaries. The results of this study provide compelling evidence that our model has achieved state-of-the-art performance in medical text summarization. We obtained an impressive ROUGE-L score of 69.42, which signifies a high degree of word overlap between our generated summaries and the reference summaries. This suggests that our model can effectively condense the essential information from medical documents, providing concise and accurate summaries of the findings.

TABLE 4. Rouge scores against BART LARGE

XSUM and Bio Bart-V2

SR. No	Model	Rouge-1	Rouge-2	Rouge-L
1	BART LARGE XSUM	56.06	46.33	57.08
2	**Biobart-V2**	**66.34**	**61.20**	**69.42**

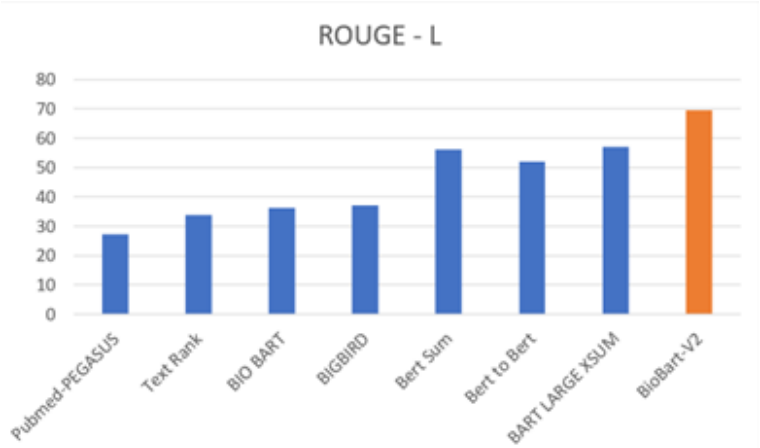

Fig. 3 Comparison of Models against Rouge L

H. IMPLICATIONS OF RESEARCH

The implications of our findings are significant for radiologists and other medical professionals in the healthcare industry. The ability to quickly and accurately summarize medical documents can greatly enhance productivity and decision-making processes. With our model's exceptional performance, it has the potential to serve as a valuable tool in clinical settings, allowing healthcare professionals to efficiently extract key information from extensive medical reports, thus saving time and improving overall workflow. Moreover, the accurate summaries produced by our model can facilitate collaboration and knowledge sharing among medical experts, leading to enhanced accuracy in diagnoses and treatment plans.

İ. ANALYSIS AND FUTURE WORK

In conclusion, our study demonstrates the effective- ness of the Biobart-V2 model in summarizing medical documents. With a remarkable ROUGE-L score of 69.42, our model showcases its state-of-the-art performance and its potential to be an invaluable as- set for radiologists and other medical professionals.

The ability to generate accurate summaries efficiently can enhance productivity and decision-making in the healthcare industry, contributing to improved patient care and outcomes. Future research in this domain could focus on fine-tuning the model for specific medical specialties or exploring its integration into existing clinical workflows to further optimize its benefit.

References

[1] Q. Xie, Z. Luo, B. Wang, and S. Ananiadou, "A survey on biomedical text summarization with pre-trained language model," *arXiv preprint arXiv:2304.08763*, 2023.

[2] R. Jain, A. Jangra, S. Saha, and A. Jatowt, "A survey on medical document summarization," *arXiv preprint arXiv:2212.01669*, 2022.

[3] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," pp. 160–167, 2008.

[4] D. B. Johnson, R. K. Taira, A. F. Cardenas, and D. R. Aberle, "Extracting information from free text radiology reports," *International Journal on Digital Libraries*, vol. 1, pp. 297–308, 1997.

[5] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, 2022.

[6] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[7] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li *et al.*, "Pre-trained language models in biomedical domain: A systematic survey," 2021.

[8] B. Gundogdu, U. Pamuksuz, J. H. Chung, J. M. Telleria, P. Liu, F. Khan, and P. J. Chang, "Customized impression prediction from radiology reports using bert and lstms." IEEE, 2021.

[9] A. A. Syed, F. L. Gaol, and T. Matsuo, "A survey of the state-of-the-art models in neural abstractive text summarization," *IEEE Access*, vol. 9, pp. 13 248–13 265, 2021.

[10] G. Frisoni, G. Moro, and A. Carbonaro, "A survey on event extraction for natural language understanding: Riding the biomedical literature wave," vol. 9. IEEE, 2021, pp. 160 721–160 757.

[11] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[12] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," vol. 32, no. 1, 2018.

[13] P. Bose, S. Srinivasan, W. C. Sleeman IV, J. Palta, R. Kapoor, and P. Ghosh, "A survey on recent named entity recognition and relationship extraction techniques on clinical texts," *Applied Sciences*, vol. 11, no. 18, p. 8319, 2021.

[14] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert systems with applications*, vol. 165, p. 113679, 2021.

[15] S. Chopra, M. Auli, and A. M. Rush, *Abstractive sentence summarization with attentive recurrent neural networks*, 2016.

[16] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[17] A. Celikyilmaz, A. Bosselut, X. He, and Y. Choi, "Deep communicating agents for abstractive summarization," *arXiv preprint arXiv:1803.10357*, 2018.

[18] Y. Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.

[19] B. Goodrich, V. Rao, P. J. Liu, and M. Saleh, "Assessing the factual accuracy of generated text," pp. 166–175, 2019.

[20] Q. Grail, J. Perez, and E. Gaussier, "Globalizing bert-based transformer architectures for long document summarization," pp. 1792–1810, 2021.

[21] B. Gundogdu, U. Pamuksuz, J. H. Chung, J. M. Telleria, P. Liu, F. Khan, and P. J. Chang, "Customized impression prediction from radiology reports using bert

and lstms," *IEEE Transactions on Artificial Intelligence*, 2021.

[22] G. Hripcsak, J. H. Austin, P. O. Alderson, and J. Friedman, "Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports," *Radiology*, vol. 224, no. 1, pp. 157–163, 2002.

[23] S. Rothe, J. Maynez, and S. Narayan, "A thorough evaluation of task-specific pretraining for summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 140–145.

[24] A. H. Roudsari, J. Afshar, S. Lee, and W. Lee, "Comparison and analysis of embedding methods for patent documents," in *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2021, pp. 152–155.

[25] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–35, 2021.