Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 7, No. 3, July 2025, pp. 0250301-01 ~ 0250301-010
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v7i3.1427

Hybrid Approaches for Advanced Medical Text Summarization: Combining TF-IDF, BERT, and Seq2Seq Models

Matimpati Chitra Rupa[1*], Kasarapu Ramani[2*]

[1]Department of CSE, JNTU Anantapur, Andhra Pradesh, India.

[2]Department of Data Science, Mohan babu University, Andhra Pradesh, India.

* cussely@gmail.com

Abstract. Clinicians, researchers, and healthcare professionals are confronted with the challenge of efficiently extracting relevant knowledge from vast amounts of textual data. Medical text summarization emerges as a crucial tool to address this challenge by condensing lengthy medical documents into concise, informative summaries. A comprehensive hybrid approach is proposed to address the challenges in medical text summarization by combining both extractive and abstractive methods, by integrating Term Frequency-Inverse Document Frequency (TF-IDF) of Natural Language Processing (NLP) and AutoModelForSeq2SeqLM of Large Language Model. The performance this proposed approach is compared with existing methods such as Bidirectional Encoder Representations from Transformers (BERT), Text Rank, K-means, face book BART-Large-CNN, GPT2 using ROUGE-1, ROUGE-2 and ROUGE-L metrics. The experimental results show that hybrid approach is outperforming other existing methods. Medical text summarization helps extract important information from large medical documents. This work combines two methods, TF-IDF and AutoModelForSeq2SeqLM, to create better summaries, performing better than existing techniques like BERT and GPT-2 based on ROUGE scores.

Keywords: Medical NLP, Hybrid Summarization, Text Mining, Extractive Summary, Abstractive summary, AutoModelForSeq2SeqLM, BERT, BART-Large-CNN, Text Rank.

(Received 2025-01-27, Revised 2025-03-25, Accepted 2025-03-26, Available Online by 2025-05-06)

1. Introduction

In the medical field, professionals struggle with a vast amount of research and clinical documentation, leading to information overload. Summarization addresses this challenge by distilling crucial insights from extensive documents, aiding healthcare professionals in staying updated without spending excessive time reading. These concise summaries support clinical decision-making processes by providing quick overviews of emerging treatments, research outcomes, and patient records, ultimately enhancing efficiency and effectiveness in healthcare.

The BERT and BART have revolutionized medical document summarization, improving efficiency by analyzing context and generating coherent summaries. Advanced models are needed to understand semantic relationships and handle medical terminology for more nuanced summaries. Summarizing medical texts aids researchers in quickly assessing relevance and navigating through diverse document types. Despite the variability in medical literature, summarization provides a valuable solution for extracting key insights from clinical notes, research articles, and patient records. NLP methods play a crucial role in the advancement of text processing tools [1], with various approaches proposed for effective document summarization [2]. These methods find application across a broad spectrum of biomedical domains. Though the basic BER and BART models are showing better results in text summarization, the Seq2Seq model learns to map input sequences to output sequences, capturing the underlying structure and semantics of the data. With advances in deep learning architectures and techniques, Seq2Seq models continue to be at the forefront of various natural language processing tasks. In this paper, TF-IDF combined with AutoModelForSeq2SeqLM used to generate the text summaries from medical texts is proposed. The hybrid approach combining extractive (TF-IDF) and abstractive (AutoModelForSeq2SeqLM) methods is necessary because medical text are often complex, containing critical information that must be accurately preserved while also being simplified for better understanding. This approach is particularly valuable in healthcare and medical fields, where quick and accurate summarization of clinical notes, research papers, and patient records can significantly improve decision-making. Two primary types of text summarization techniques exist: extractive and abstractive. The extractive approach involves identifying key sentences within the document that convey its contents and arranging them in a coherent manner. BART (Bidirectional and Auto-Regressive Transformers) and BERT (Bidirectional Encoder Representations from Transformers), two influential transformer models, excel in various natural language processing tasks, such as text summarization. However, they have distinct architectures and approaches to summarization BART is adept at text summarization, leveraging both auto-regressive and bidirectional training for coherent output. Its generative nature allows for flexible summary generation, with fine-tuning capabilities for task adaptation. However, its computational demands and complexity may hinder practicality in resource-constrained settings, leading to longer training times and challenges in understanding its inner workings. BERT, renowned for capturing bidirectional context, enhances summarization by understanding word relationships within sentences. Pre-trained on extensive corpora, it offers flexibility through fine-tuning for various tasks, including summarization, leveraging transfer learning for enhanced performance. However, BERT's token limit for input sequences poses challenges for summarizing longer documents, potentially impacting summary quality. Additionally, its resource-intensive implementation demands significant computational power and memory resources. [4-5].

In [6] the proposed method integrates BERT with Group-Average Linkage Clustering for document clustering and summarization, leveraging BERT's semantic understanding and hierarchical structure. Parameter optimization is crucial for performance alignment with dataset characteristics. However, computational demands increase when applying BERT embeddings and clustering to large datasets, requiring significant resources. Additionally, selecting an appropriate similarity threshold for cluster merging is essential and may involve experimentation. In [7] a comparative analysis evaluates traditional machine learning methods and ChatGPT in generating intelligent summaries for medical documents on preliminary diagnoses of low-risk diseases and symptoms, highlighting that precise diagnosis isn't guaranteed. Deepika, et al. [8] proposed a method serving as an introductory overview of text summarization, with a specific focus on its importance, methods, application, and evaluation within the context of analyzing COVID-19 datasets. It begins by introducing the concept of text summarization, emphasizing its critical role in managing the growing volume of textual data sourced from various channels. The discussion then delves into the two primary methods of summarization: extractive and abstractive, with a particular emphasis on extractive summarization in the context of COVID-19 datasets. Surabhi Datta et al. [9] identified studies that described an NLP method for extracting specific cancer-related information from Electronic Health Records (EHR) sources obtained from PubMed, Google Scholar, ACL Anthology, and existing reviews using low-level extraction methods. Furthermore, it was suggested that given the extensive duplication in cancer NLP systems, the creation of a comprehensive resource comprising annotated cancer frames and corresponding NLP tools would be highly beneficial. Snehal Sameer Patil et al. [10] proposed the implementation of Natural Language Processing (NLP) methods for data processing, including classification, prediction, Word Sense

Disambiguation, segmentation, and word embedding. Deep learning methods, such as CNN, BERT, and LSTM, significantly enhanced the performance of various NLP tasks, including language understanding, text generation, classification, recognition of named entities and translation. However, challenges persisted in classification techniques, requiring further work to improve model accuracy. In their study, Yiheng Liu et al. [11] presented an extensive examination of ChatGPT and GPT-4, exploring their potential applications and notable advancements in natural language processing. The analysis covered crucial innovations such as large-scale pre-training, fine-tuning of instruction, and Reinforcement Learning from Human Feedback (RLHF), with the objective of augmenting the adaptability and efficacy of large language models (LLMs). Mayank Soni, et al. [12] proposed the study focused on assessing ChatGPT's performance in Abstractive Summarization, using automated metrics and human reviewers. Despite previous anecdotal evaluations, systematic research on ChatGPT was limited. Their findings revealed that although text classification algorithms effectively discerned real summaries from generated ones, humans encountered difficulties in making this distinction. From above literature survey combining extractive and abstractive techniques and leveraging advanced technologies like LLMs and NLP models, the methodology aims to deliver more holistic and refined summaries of medical texts. This approach promises enhanced comprehension and depth in summarization outcomes, addressing the limitations of current methodologies in medical text summarization. The paper provides a comprehensive review of recent advancements in extractive text summarization, highlighting the integration of NLP, LSTM, TF-IDF, and word vector embeddings to improve summary quality. It explores various hybrid models and techniques used to enhance summarization accuracy and efficiency [17]. The paper offers a comparative analysis between deep learning models and conventional techniques across various datasets, serving as a valuable resource for researchers and practitioners aiming to enhance text summarization methods [18]. The authors propose a hybrid approach that combines concept extraction with Term Frequency-Inverse Document Frequency (TF-IDF) and employs the Bidirectional Encoder Representations from Transformers (BERT) model to process electronic health records (EHRs). This method aims to distill essential patient information from daily progress notes, thereby enhancing the accuracy of LOS predictions [19]. The paper emphasizes the importance of integrating various summarization strategies to effectively handle the complexity and specificity of medical texts [20].

2. Methods

In the current work a hybrid text summarization using extractive technique combined with abstractive technique is proposed. To demonstrate its performance, it is compared with other existing models as shown in Fig.1.

2.1. A Hybrid TF-IDF

combined with Auto Model ForSeq2SeqLM Text Summarization:

Read input medical dataset from The Kaggle repository containing a named dataset like clinical documents on syndromes and 493 diseases' medical reports according to patient health history. The pre-processing of the input dataset begins with tokenization using the word tokenizer from the NLTK library. Following this, stop-word removal is applied using the NLTK corpus to eliminate irrelevant words. The refined text then undergoes lemmatization using the WordNet lemmatize to convert words into their base forms. After pre-processing, feature extraction is performed using the TF-IDF vectorizer from the Sci-kit Learn library. Named entities are identified, and sentence scores are computed to determine the most relevant sentences. The top-scored sentence indices are extracted to retain the most informative content. For medical text summarization, the AutoModelForSeq2SeqLM class from the Hugging Face Transformers library is utilized, specifically employing the GPT-4-based Kaludi-chatgpt-gpt4 model. This pre-trained model is applied to the extracted key sentences, generating concise and meaningful text summaries.

2.2. Text Rank Summarization

Read a medical dataset from Kaggle, consisting of 493 clinical documents detailing patient health history and information on syndromes and diseases. Pre-processing the text by transforming it to lowercase, eliminating punctuation marks and stop-words, then segmenting it into separate sentences. Construct a similarity matrix by determining the Jaro-Winkler similarity score for each sentence pair. Save these scores in a matrix format, with each cell (i, j) denoting the similarity between the ith and jth sentences.

Construct a graph representation by generating a Network X graph, with each node representing a sentence. Assign the Jaro-Winkler similarity scores as edge weights between nodes. Utilize the PageRank algorithm on the graph to allocate importance scores to each sentence/node.

Generate the top N sentences with the highest PageRank scores to create the summary. Concatenate these selected sentences to produce the final summary text.

2.3. K-Means Summarization

Read a medical dataset from Kaggle, consisting of 493 clinical documents detailing patient health history and information on syndromes and diseases. To tokenize the text, employ NLTK's sent-tokenize function, which segments the text into individual sentences. Each cleaned sentence is tokenized, and stop words are eliminated using NLTK's English stop words' corpus. Utilize Word2Vec to produce word embeddings derived from the cleaned sentences. Compile a list containing all words present in the corpus for preparation. Perform K-Means clustering on the sentence vectors to group similar sentences together. Determine the optimal number of clusters and initialize centroids using the k-means algorithm. For every cluster centroid, locate the sentence whose vector exhibits the smallest Euclidean distance to the centroid. Display the closest sentences associated with each cluster centroid. This method groups similar sentences, offering a representative sentence for each cluster, and facilitating summarization.

2.4. Chatgpt-2 Summarization

Read a medical dataset from Kaggle, consisting of 493 clinical documents detailing patient health history and information on syndromes and diseases. Utilize the GPT2LMHeadModel and GPT2Tokenizer classes provided by the Transformers library to initialize the pre-trained GPT-2 model and tokenizer. Utilize the GPT-2 tokenizer to tokenize the input text. Transform the input text into tokenized IDs, which are formatted appropriately for input into the GPT-2 model. Utilize the GPT-2 model to generate a summary from the tokenized input text. Define parameters including the maximum length of the summary (max_length), the number of summaries to generate (num_return_sequences), and others as needed. Decode the generated summary from token IDs back into human-readable text format using the GPT-2 tokenize.

2.5. BERT Summarization

Read medical dataset sourced from Kaggle, containing 493 clinical documents that offer comprehensive insights into patient health records, including details on various syndromes and diseases. Initialize the BERT tokenizer using the Bert Tokenizer from the pre-trained method. Employ the BERT tokenizer to tokenize the input text. Convert the input text into token IDs suitable for input to the BERT model. Initialize the summarization pipeline using the Hugging Face pipeline function. Specify the model and tokenizer to be used in the pipeline. Use the initialized summarization pipeline to generate a summary of the input text.

2.6. BART Summarization

Read Kaggle medical dataset comprising 493 clinical documents, providing thorough insights into patient health records, encompassing a range of syndromes and diseases.

- Initialize the BART tokenizer using the Bart Tokenizer from_ pretrained method.
- Define the reference medical text that needs to be summarized.
- Tokenize the reference medical text using the BART tokenizer.
- Convert the tokenized text into token IDs suitable for input to the BART model.
- Use the BART model to generate a summary of the medical text.
- Decode the summarized text from token IDs back into human-readable text format using the tokenizer.
- Generate Summary

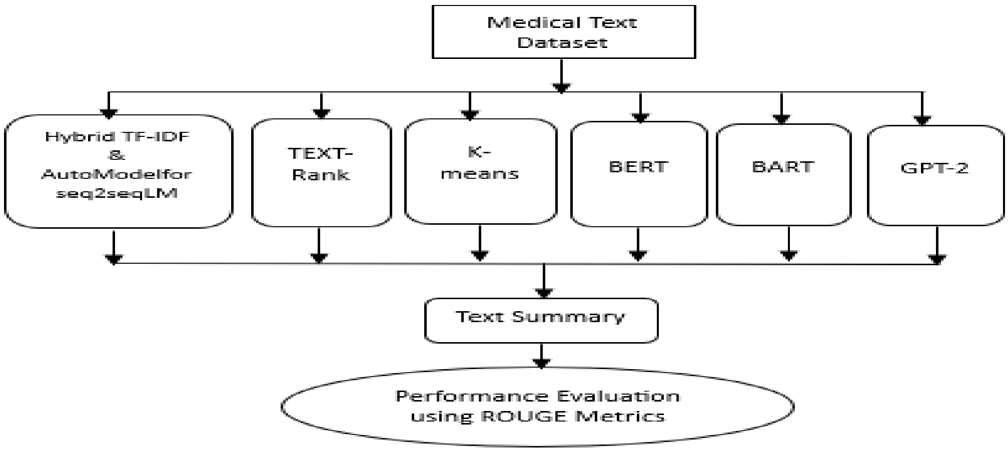

Figure 1. Block diagram of Text Summarization Models and Performance Evaluation

Table 1. Limitations of Each Model

Model	Limitation	Solution in Hybrid Approach
TF-IDF	Lacks contextual understanding	Uses AutoModelForSeq2SeqLM for coherence
BERT	Limited input length for long texts	TF-IDF pre-filters key content
TextRank	Produces redundant or disjointed summaries	Abstractive model improves flow
GPT-2	Struggles with medical terminology accuracy	TF-IDF ensures factual precision

3. Results and Discussion

3.1. Environment

The implementation of above-mentioned proposed methodology used python 3.11.0 version and encompassed relevant libraries: ChatGPT for integration, spacy for NLP, NumPy for numerical operations, nltk for additional NLP functionalities, Sumy for Text Rank implementation, and rouge for ROUGE metric calculation.

3.2. Experimental Results

The comparative study outlines various methods for text summarization, including Text Rank, K-means, BERT, BART [11] and GPT-2 with the proposed method. Text Rank utilizes network analysis techniques, employing the Jaro-Winkler similarity metric to establish similarity between sentences, which are then represented as nodes in a network. The PageRank algorithm determines the importance of each sentence based on its similarity to others, selecting top-ranked sentences for the summary. Kmean utilizes Word2Vec embeddings and K-means clustering to cluster sentences, providing insights into the main topics or themes of the text [12]. BERT and BART are implemented using pre-trained models for summarization, with BERT employing a pipeline for summary generation, and BART utilizing specific parameters for summary generation [13-14]. GPT-2, a Generative Pre-trained Transformer 2 model, operates through tokenization, embedding, and multiple layers of Transformer blocks to generate summaries [4]. It predicts the next token in the sequence iteratively, generating summaries and evaluated using Rouge scores.

3.3. Performance Metrics

ROUGE score, is computed using the Rouge library. Finally, the generated summary and its associated ROUGE scores are printed to assess the quality of the summarization output. For experimentation, medical documents of three different lengths-short, medium and long are applied and based on the given reference summary and text summaries are generated. The obtained text accuracy is evaluated based on ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score. The Text summarization methods' performance is evaluated using the ROUGE score, which measures the similarity between the generated summary and the reference summary. The ROUGE-1, ROUGE-2 and ROUGE-L metrics w.r.to recall, precision and F-score as given in Eqn. (1) to (6).

Precision(p)=Crucial Sentences/Total Number of Summarized Sentences.......(1)

Recall(r)=Total number of crucial sentences retrieved/Total number of significant sentences found ....(2)

F-Score(f)=2X ((Precision X Recall))/((Precision+Recall))------------------(3)

ROUGE-1= Count of overlapping Unigrams/Total Unigrams in Reference ....(4)

ROUGE-2= Count of overlapping Bigrams/Total Bigrams in Reference ....(5)

ROUGE-L= Length of Longest Common Subsequence/ Total words in Reference.... (6)

$$Precision(p) = \frac{Crucial\ Sentences}{Total\ Number\ of\ Summarized\ Sentences} \text{.......(1)}$$

$$Recall(r) = \frac{Total\ number\ of\ crucial\ sentences\ retrieved}{Total\ number\ of\ significant\ sentences\ found} \text{..........(2)}$$

$$F-Score(f) = 2X\,\frac{(Precision\ X\ Recall)}{(Precision+Recall)} \text{------------------(3)}$$

ROUGE-1= Count of overlapping Unigrams/Total Unigrams in Reference ....(4)

ROUGE-2= Count of overlapping Bigrams/Total Bigrams in Reference ....(5)

ROUGE-L= Length of Longest Common Subsequence/ Total words in Reference.... (6)

Table 2. Comparison of summarization models for medical summary of ROUGE Score.

Performance metrics	BERT	GPT2	Text Rank	BART	K-Means	Hybrid Model
ROUGE-1r	14.55	83.33	27.54	80.00	44.97	76.47
ROUGE-1p	76.66	14.97	90.19	72.72	100	100
ROUGE-1f	24.44	25.38	42.20	76.19	62.04	86.66
ROUGE-2r	07.11	61.11	18.75	61.11	35.95	67.00
ROUGE-2p	52.11	09.16	70.31	57.89	97.75	100
ROUGE-2f	12.54	01.59	29.60	59.45	52.75	86.36
ROUGE-Lr	14.55	83.33	27.54	80.00	44.97	76.47
ROUGE-Lp	76.66	14.33	90.19	72.72	100	100
ROUGE-Lf	24.46	25.38	42.20	76.19	62.04	86.66

The above Table. 2 presents performance evaluation of various medical text summarization models using ROUGE metrics.

Table 3. Performance Analysis of Text Summarization Models

Model Name	**ROUGE 1**	**ROUGE 2**	**ROUGE L**
BERT	38.55	23.92	38.55
GPT2	41.22	23.95	41.22
Text Rank	53.31	39.55	53.31
K-Means	69.00	62.15	69.00
BART	76.30	59.48	76.30
Hybrid Model	87.71	87.45	87.71

Table 3. describes the performance of summarization models w.r.to Rouge-1, Rouge-2 and Rouge-L metrics. The Fig.2 represents the ROUGE 1 performance of six distinct models for medical text summarization, Fig. 3 represents the comparison of ROUGE 2 score and Fig. 4 displays performance comparison of summarization methods according to ROUGE L score.

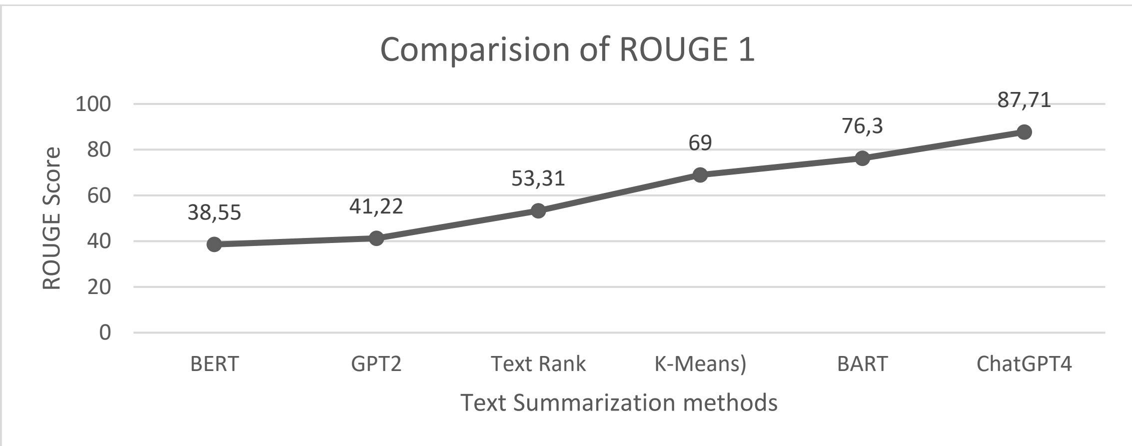

Figure 2. Comparison of ROUGE 1 score

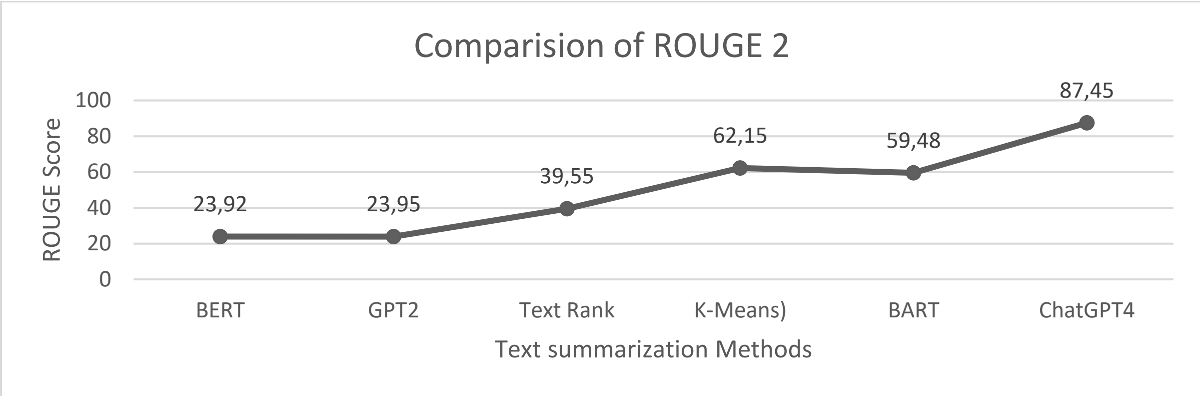

Figure 3. Comparison of ROUGE 2 score

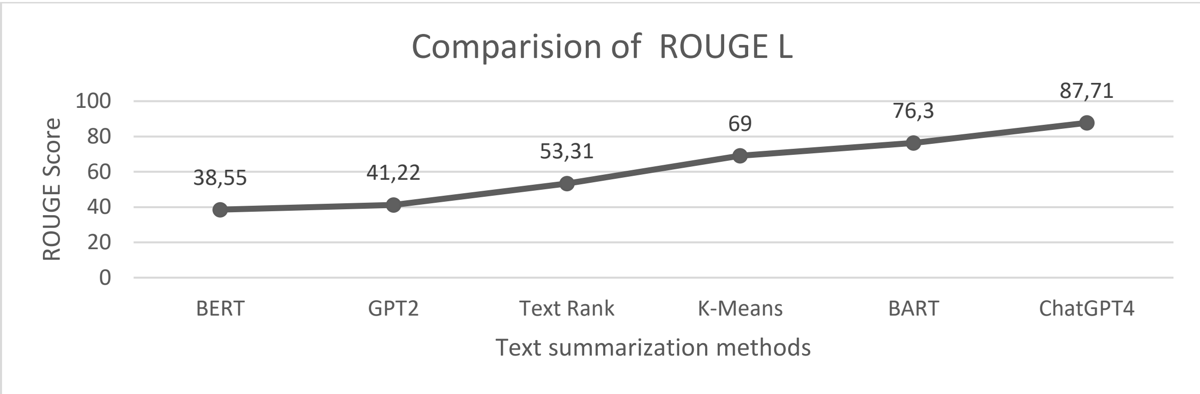

Figure 4. Comparison of ROUGE L score for medical text Summarization Models.

Based on the experimental findings, it is observed that both K-means and Text Rank outperform than BERT and GPT2 models, whereas the hybrid method TF-IDF with AutoModelSeq2Seq outperforms all mentioned models, and emerges as the top performer, showcasing exceptional summarization capabilities.

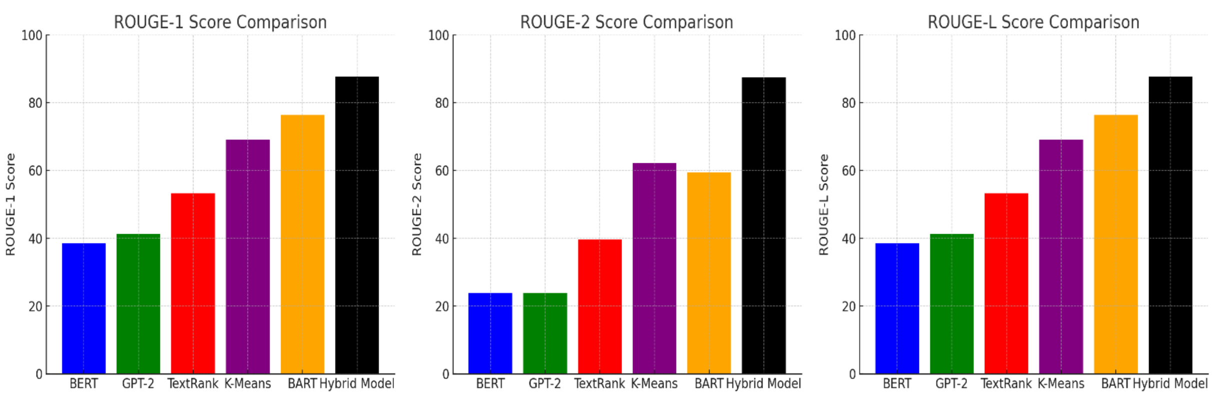

Figure 5. Detailed comparison of the performance metrics across the different models and the hybrid approach.

4. Conclusions

This research work addresses the limitations of existing approaches in medical text summarization by presenting a novel hybrid summarization methodology that integrates both extractive and abstractive methods. Leveraging advanced language models our approach ensures a more comprehensive and nuanced summarization of medical texts. The comparison of various text summarization models based on their performance across different ROUGE metrics reveals distinct patterns and strengths among each model. These results underline the varying degrees of effectiveness among the models, with TF-IDF with AutoModelSeq2Seq standing out as the top performer, followed closely by K-Means and Text Rank, while BERT and GPT-2 exhibit more moderate performance in comparison. Fine-tuning existing models, investigating domain adaptation strategies, and refining evaluation metrics are also promising directions for improving summarization performance in future enhancement. This method can help doctors, researchers, and healthcare workers by providing short and clear summaries of medical documents, making it easier to understand important information quickly.Future improvements could include training the model with more medical data, using AI to improve summary quality, and applying it to real-time medical report summarization. Additionally, adding images, medical scans, and knowledge graphs could help the model better understand complex medical information.

References

[1] Sudarshan, R., Sasikala, D. and Kalavathi, S., “Advancing Clinical Text Summarization through Extractive Methods using BERT-Based Models on the NBME Dataset”in 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), IEEE, pp. 1288-1294, 2023.

[2] Tristam Dacayan, Daniel Ojeda, “Summarizing Behavioral Electronic Health Records using a Natural Language Processing Pipeline”, International Conference on Computational Science and Computational Intelligence (CSCI),pp,1635-1639, IEEE, 2022.

[3] Kasarapu Ramani, E.Lalitha, Text Summarization of Medical Documents using Abstractive Techniques,2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC),IEEE,2023.

[4] Deepika, S. and Shridevi, S., “Extractive Text Summarization for COVID-19 Medical Records” in innovations in Power and Advanced Computing Technologies (i-PACT), IEEE,2021.

[5] Mukesh Kumar Rohil, Varun Magotra, “An exploratory study of automatic text summarization in biomedical and healthcare domain”, Healthcare Analytics, Elsevier, 2022.

[6] Abhishek Kuber, Soham Kulthe, “Extensive Study of Automatic Text Summarization on Biomedical Texts”, 6th International Conference on Computing, Communication, Control and Automation (ICCUBEA), IEEE, 2022.

[7] Loredana Caruccio, Stefano Cirillo, “Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot”, Expert Systems with Applications, Elvesier, 2023.

[8] Deepika S, Lakshmi Krishna N, “Extractive Text Summarization for COVID-19 Medical Records”, Innovations in Power and Advanced Computing Technologies (i-Pact), IEEE, 2021.

[9] Surabhi Dattaa, Elmer V. Bernstam, “A frame semantic overview of NLP-based information extraction for cancer related EHR notes”, Journal of Biomedical Informatics, Elsevier, 2019.

[10] Snehal Sameer Patil, Vaishnavi Moorthy, “Extraction of Unstructured Electronic Healthcare Records using Natural Language Processing”, International Conference on Networking and Communication (ICNWC), IEEE, 2023.

[11] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge, “Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models”, Shanghai, China, IEEE 2023.

[12] Mayank Soni, Vincent Wade, "Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms", ADAPT Centre, Trinity College Dublin, ResearchGate, 2023.

[13] Zhu, Y., Yang, X., Wu, Y. and Zhang, W, "Leveraging Summary Guidance on Medical Report Summarization. IEEE,2023.

[14] Kasarapu Ramani, K. Bhavana," Extractive Text Summarization through k-means, LSA, and Text Rank, International Conference on Wireless Communications Signal Processing and Networking (WiSPNET),IEEE,2023.

[15] Pooja Vinod, Seema Safar Fine-tuning the BERTSUMEXT model for Clinical Report Summarization2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020.

[16] Zhimeng Luo, Yuelyu Ji, "Towards Accurate and Clinically Meaningful Summarization of Electronic Health Record Notes: A Guide Approach", EMBS International Conference on biomedical and Health Informatics (BHI), IEEE, 2023.

[17] Azam, M., Khalid, S., Almutairi, S., Khattak, H. A., Namoun, A., Ali, A., & Bilal, H. S. M. (2025). Current Trends and Advances in Extractive Text Summarization: A Comprehensive Review. *IEEE Access*.

[18] Shahade, A. K., & Deshmukh, P. V. (2025). A Unified Approach to Text Summarization: Classical, Machine Learning, and Deep Learning Methods. *Ingénierie des Systèmes d'Information*, *30*(1).

[19] Rhazzafe, S., Caraffini, F., Colreavy-Donnelly, S., Dhassi, Y., Kuhn, S., & Nikolov, N. S. (2024). Hybrid summarization of medical records for predicting length of stay in the intensive care unit. *Applied Sciences*, *14*(13), 5809.

[20] Agilandeeswari, L., Dagar, A., Deepthi, A., & Arangasakthivel, R. (2023, December). Automatic Text Summarization for Medical Dataset-An Analysis. In *International Conference on Intelligent Systems Design and Applications* (pp. 336-352). Cham: Springer Nature Switzerland.