The Journal of Supercomputing (2025) 81:171
https://doi.org/10.1007/s11227-024-06640-6

Clustering-based topic modeling for biomedical documents extractive text summarization

Nabil M. AbdelAziz[1] · Aliaa A. Ali[1] · Soaad M. Naguib[1] · Lamiaa S. Fayed[1]

Accepted: 21 October 2024

Abstract

The increasing volume of electronic text, especially in the biomedical domain, requires automatic text summarization (ATS) to help researchers navigate and find relevant information. This study proposes an unsupervised extractive ATS method to improve the quality of biomedical document summaries by focusing on subtopic diversity. The method integrates topic modeling and clustering with bidirectional encoder representation from transformers (BERT). To evaluate the effectiveness of the proposed study, it tested on a new corpus of 200 biomedical research papers from Biomed Central. The results were evaluated using the ROUGE metric and qualitative evaluation by medical experts. The ROUGE metric yielded scores of 0.4838 (Rouge-1), 0.2174 (Rouge-2), and 0.2206 (Rouge-L), while the qualitative evaluation achieved an average score of 4.10, 4.06, 3.55, 4.0, and 4.0 for completeness, relevance, conciseness, informativity, and readability, respectively. The results demonstrate the effectiveness of the proposed method in summarizing long medical documents.

Keywords Automatic text summarization · Extractive summarization · Topic modeling · Clustering · BERT

✉ Aliaa A. Ali
aliaam@fci.zu.edu.eg

Nabil M. AbdelAziz
nmabedelaziz@fci.zu.edu.eg

Soaad M. Naguib
smnagieb@fci.zu.edu.eg

Lamiaa S. Fayed
lsfayed@fci.zu.edu.eg

[1] Information Systems Department, College of Computers and Information, Zagazig University, Zagazig 44519, Egypt

Published online: 13 November 2024

1 Introduction

The number of digital texts in the medical field has increased rapidly due to the continuous change and development in the internet and technology [1–4]. The biomedical information and literature, which are critical and priceless in managing global diseases, are accessible in the form of scientific papers, clinical reports, patient health records, and web documents [1, 5, 6]. A wide range of sources provides information including electronic biomedical research databases, online medical reports, and electronic health record systems (EHRs) [1, 7]. For example, the number of biomedical papers published in the PubMed database exceeds 30 million, with an annual increase of over 1 million new papers [2]. Returning to COVID-19, the volume of research papers has increased significantly since November 2019, when the pandemic started [8]. In addition, in November 2019 the Medline database contained over 24 million articles from over 5500 biomedical journals [7]. Navigating this massive volume of information manually is impossible. Researchers and clinicians who search these sources, to understand current practices in a specific field, access recent works, capture new ideas, conduct experiments, and obtain results found it very challenging to retrieve and read all relevant information they are seeking for [4, 9].

Automatic text summarization (ATS) has been proposed to assist clinicians and researchers in extracting the relevant information from large collections of biomedical literature [10]. ATS reduces long textual documents into shorter versions while preserving their essential meaning and informational content [11]. ATS provides significant benefits in the biomedical field. It manages the overloaded information by providing concise summaries that enable researchers/healthcare professionals to capture the essential information (such as potential treatments, key trends, gaps, and findings) without reading the entire document. This optimizes time and improves productivity [7, 8, 12–15]. The generated summaries produced by ATS can be utilized in other fields of study such as information retrieval, text classification, and questions-answering [2]. Although abstracts are often included in scientific papers, there are several reasons to generate summaries from full-text sources. First, no ideal summary exists as it depends on each user with different information needs and domains. Second, a document's abstract may ignore important content in the full text. Finally, customized summaries can be valuable in question-answering systems where they provide personalized information [16].

ATS has two main methods which are extractive and abstractive summarization. The first is a method that identifies and retrieves important sentences from the source material to build a concise brief [17]. It is difficult to maintain coherence between sentences in extractive summarization and to simplify lengthy, complex sentences [18]. Instead, abstractive summarization rewrites essential parts of the source text into new sentences [19]. For summarizing biomedical research papers, the extractive method is more suitable than abstractive as it keeps the original terminology and vocabulary used by researchers; this guarantees the accuracy of facts represented in the source text. In contrast, clinical documents

and EHRs are better suited for abstractive summarization, which helps medical professionals understand patient reports more concisely and humanly [6, 20].

Traditional summarization techniques focused on simple term frequency approaches and diverse attributes like title relevance, sentence position, length, extracted keywords, and numerical content for extracting salient sentences from the source document. However, these generic features were less effective in summarizing biomedical documents [21]. Another critical issue is the duplication of sentences in the produced summary, with lack of coherence and semantic accuracy. In addition, summarizing lengthy documents that contain multiple subtopics is considered one of the most significant challenges when using ATS. This leads to a lack of diversity in the summarized content. Traditional summarization methods often provide biased and partial summaries because they cannot capture all subtopics represented in the source documents [22, 23]. Many studies try to solve the problem of redundancy and coherence by using techniques like attention mechanisms, coverage head models, and secondary encoders. While this to some extent eliminates repetition, it also makes summaries excessively biased toward certain subtopics covered in the text, especially in lengthy documents and multi-document summarization. Recent works used topic modeling and clustering to address this problem by picking an equal number of sentences according to the distribution of topics without ranking the sentences with the topic itself [24]. Therefore, this work attempted to focus on subtopics by using K-medoid to cluster sentence vectors and score sentences inside the subtopic itself.

Based on the above, we introduce an unsupervised methodology that combines topic modeling and clustering with bidirectional encoder representation from transformers (BERT) for an extractive summarization of a single document. The first stage in the proposed methodology is to improve text readability by preprocessing the source document. Next, latent Dirichlet allocation (LDA) is applied to identify hidden topics in the document, with the coherence measure employed to optimize the number of topics distributed in the document. The allocating topics are distributed over sentences, where every sentence is related to a specific topic, and sentences related to the same topic are grouped together. BERT is incorporated to transform the text into deep conceptualized embeddings for accurate sentence vectorization. The vectorized sentences are then fed to K-medoid clustering to extract the top representative sentences, which are finally used to construct the final summary.

The major contributions of this study are:

1. A new corpus comprising 200 biomedical research papers on knee osteoarthritis management was collected and introduced for extractive summarization of single documents.
2. An unsupervised methodology is introduced and effectively evaluated on the new corpus for single-document extractive summarization.
3. The coherence measure is integrated into the proposed methodology to allocate subtopics.
4. Different variants of BERT were tested and compared to select the best one.

5. The use of topic modeling and clustering in conjunction with BERT yielded better results when compared to prior efforts in developing topic-modeled ATS systems.
6. Two types of evaluation (qualitative and quantitative) were utilized.

This paper is structured as follows. Section 2 presents a comprehensive review of relevant studies. The methodology is represented in Sect. 3. In Sect. 4, results and discussion were provided. Finally, the closing comments and outline of potential areas for future work are presented in Sect. 5.

2 Related work

ATS is a branch of natural language processing (NLP) where the computer creates a summary of single/multiple documents, ensuring the summary aligns with the main topic and concept of the source text document [25]. Different factors can be considered to classify ATS as described in Fig. 1. For the input document, summarization can be done on a single or multi-document. A single document includes summarizing each document individually, while multi-document ones summarize more than one document together to produce one summary [26]. Also, the summarization method can be extractive, abstractive, and hybrid. The extractive method entails ranking and identifying the most significant sentences in the document to provide a short, representative summary [27]. In contrast, abstractive summarization redrafts the key ideas represented in the document instead of selecting individual sentences [28]. Hybrid summarization starts by selecting important sentences from text to generate an extractive summary, and then abstractive methods are applied to rewrite and convert the extractive summary to an abstractive one [29]. In addition, the content

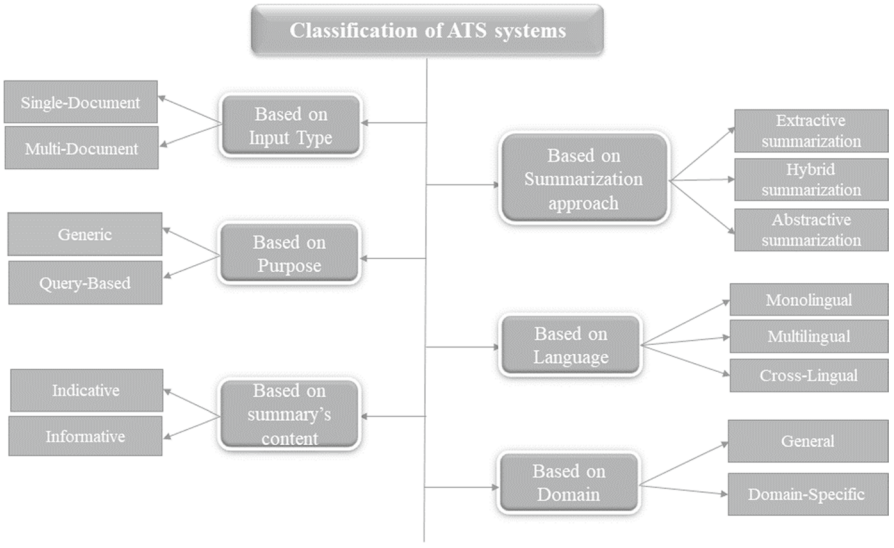

Fig. 1 Different factors for classify ATS systems [33]

of the summary is classified as indicative or informative. The indicative summary provides a brief idea about the topic and the issues presented in the document, the informative summary provides a complete and more information covered in the input document [1, 30]. The purpose of summarization is another factor for classification. It can be generic or query-based. Generic summarization summarizes the overall information content available in the source text, while query-based answers a specific user's query which focuses on providing information related to this query [31, 32].

Extractive summarization of biomedical documents is usually preferred over abstractive summarization due to its ease of sentence extraction and higher accuracy [1]. Extractive methods are categorized to (1) statistical-based techniques, (2) concept-based techniques, (3) topic-based techniques, (4) graph-based techniques, and (5) machine learning methods [2, 32]. Recent studies in the biomedical domain used these methods to improve the quality of the produced summaries. An unsupervised method based on semantic similarity and keyword extraction was proposed for single- and/or multi-document extractive summarization. This method combined a concept map and RAKE approach to produce summaries for 1040 biomedical transcripts [34]. Many studies combined itemset mining with domain knowledge to create a concept model for summarizing biomedical documents [35, 36]. Moradi et al. [7] introduced a Bayesian summarizer that mapped the text into unified medical language system (UMLs) concepts, also six different features were incorporated to define the critical concepts. This method was evaluated on a medical corpus consisting of 400 biomedical documents. Moradi et al. [37] developed a graph-based method where Helmholtz principle was used to identify the essential concepts from text then a graph-based method was built based on these concepts to capture the essential sentences for the summary. Davoodijam et al. [2] proposed a multi-layer graph called MultiGBS that incorporates MultiRank algorithm for selecting sentences from multi-layers to produce an extractive summary of single biomedical document. Different studies leveraged a graph-based method that combines itemset mining and sentence clustering to enhance and improve biomedical text summarization tasks. For example, Rouane et al. [30] used UMLs to represent biomedical articles as a collection of concepts. Related sentences are grouped together using K-means clustering algorithm. Next, Apriori model was used to identify the common itemsets among the grouped sentences. Finally, the important sentences were picked from each group to generate an extractive summary. Azadani et al. [3] combined a minimum spanning tree-based clustering with frequent itemset mining for extractive summarization. CIBS summarizer was introduced in 2018 for extractive summarization. This method used itemset mining to extract the main topics and then clustered sentences with the same topic together. Finally, the extractive summary includes sentences from each cluster, ensuring that the summary covers all topics [38].

Recently, many studies incorporated pre-trained language models (PLMs) for extractive summarization of biomedical documents. For instance, Du et al. [39] proposed BioBERTSum, a PLM encoder which has been optimized and finetuned for biomedical extractive summarization tasks. Kanwal et al. [40] finetuned BERT on MIMIC-III dataset for extractive summarization of Digital Health Records.

Moradi et al. [21] utilized a hierarchical clustering method to group contextual embeddings of phrases using the BERT encoder. The most informative sentences from each group are selected to construct the final summary. Padmakumar et al. [41] introduced an unsupervised extractive summarization approach that encodes phrases using GPT-2 model and uses pointwise mutual information (PMI) to determine semantic similarity between texts. This approach was evaluated on a medical journal dataset. A new approach that combines graph-based and domain-specific word embedding BioBERT for summarizing biomedical articles was proposed by Moradi et al. [6]. Xie et al. [42] proposed a KeBioSum framework for biomedical extractive summarization tasks. It improved performance of PLMs by incorporating fine-grained domain knowledge (PICO components) and employing sophisticated training approaches. CovSumm is an unsupervised approach that leverages strengths of both transformer-based models and graph-based methods for summarizing COVID-19 literature [8]. Overall, there are many PLMs pre-trained specifically for biomedical texts, such as BioBERT [43], PubMed BERT [44], SciBERT [45], BlueBERT [46], ClinicalBERT [47], and ALBERT [48]. Meng et al. [49] suggested splitting the knowledge graph into subgraphs and injecting them with several PLMs like BioBERT, SciBERT, and PubMed BERT.

Topic modeling was first introduced in 2000 by [50]. Topic modeling plays a crucial role in enhancing text summarization, especially for long complex documents with topic diversity such as biomedical research papers. It identifies main topics or themes within a document or collection of documents, which guarantees that the generated summaries contain a comprehensive coverage of key concepts represented in the source document/documents [51]. To avoid topics biased on the summaries generated from long or multi-documents, topic modeling is incorporated to manage topics diversity [22]. Scarce studies applied topic modeling in summarizing medical documents. A study by [52] proposed a method for extractive summarization for long documents. The framework leverages topic information to capture dependencies in long documents by using a heterogeneous graph neural network. This method evaluated over three datasets PubMed, arXiv, and GovReport. Also Xie et al.[53] integrated domain-knowledge and graph-based topic modeling into transformer architecture for biomedical text summarization.

In fields other than biomedicine, Issam et al. [54] combined topic modeling and TextRank together for summarizing WikiHow dataset. Also, Liu et al. [55] combined topic modeling and statistical-based features for summarizing multi-documents. Srivastava et al. [22] examined a novel method that combines clustering with topic modeling for single-document extractive summarization and then evaluate it over three datasets: the DUC2002, WikiHow, and CNN-DailyMail.

The following conclusions were drawn from the previous extractive summarization review:

1. The benefits of topic modeling in summarizing documents with topic diversity are well known.
2. The benefits of organizing sentences to determine the most significant ones are well established.

3. PLMs that pre-trained on biomedical corpora showed their effectiveness in enhancing the quality of the extractive summary.
4. Although some studies have employed LDA in topic-modeling summarization, few studies have tried to fine-tune the model to identify an appropriate number of subtopics.

Based on these conclusions, this paper proposes a new methodology to summarize a single biomedical document based on an unsupervised extractive summarization. The proposed methodology employs LDA to find and assign topic to every sentence in the document. To enhance performance, we used coherence to find a suitable number of topics in the source document. Once each sentence is assigned to a specific topic, the sentences with the same topic are grouped together. For each topic, we tested variants of BERT (S-BERT, BlueBERT, SciBERT, PubMed BERT) to map the text in each group to its conceptualized embedding, then each group clustered by applying K-medoid clustering. Finally, the complete summary is constructed by taking the top n sentences from each group. Section 3 gives more details about the proposed method.

3 Proposed methodology

3.1 Summarization method

The proposed biomedical extractive summarization method consists of four phases: (1) document preprocessing, (2) topic assignment, (3) deep conceptualized embedding, and (4) clustering and sentence selection as shown in Fig. 2. The experiments and all tests were performed on Google Colaboratory with the following computational resources: NVIDIA T4 GPU, 12.7 GB RAM, 2 vCPUs, and approximately 107 GB of disk space.

3.1.1 Document preprocessing

The proposed summarization method begins with a preparatory step to prepare the document, as the content of the documents doesn't align with the phases of the proposed methodology. The preprocessing involves the following actions:

(a) Unnecessary sections in the document have been removed, such as (the references section, title, author's information, keywords, acknowledgments, competing interests, headers, sub-headers, citations, etc.).
(b) Figures and tables have been omitted since they aren't involved in the summarization process.
(c) Abbreviations and their expansions are extracted from the abbreviation section, and the occurrences of the abbreviations with the expansions are replaced throughout the document.

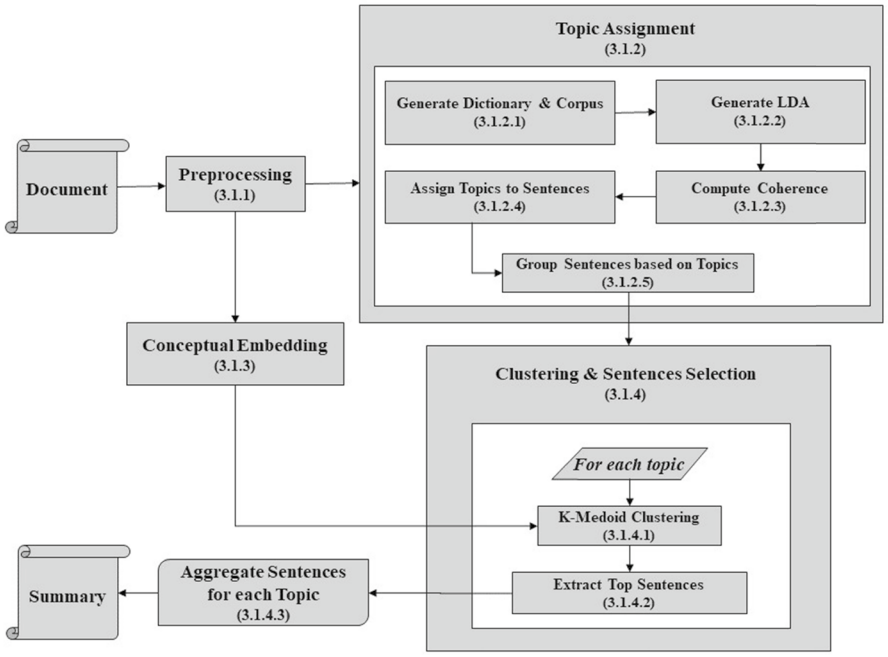

Fig. 2 Overview of our biomedical extractive text summarization method

(d) We separated the abstract from the document's body to use it as a reference summary for further evaluation.
(e) The text of the document's body is split into sentences using sent_tokenize from nltk tokenize.
(f) We utilized a list of stop words from Medline to remove the words that do not contribute to the identification of meaningful sentences.

3.1.2 Topic assignment using LDA

LDA is used to extract the main topics of the preprocessed document. It was introduced by Blei et al. [56]. LDA is an unsupervised generative probabilistic model that seeks to discover latent (hidden) topics in unstructured text. It consists of a Bayesian network structure with three levels, namely "document-topic-word" [57, 58]. Figure 3 describes the generative process of LDA, which supposes that any document is created by identifying the number of topics (θ). Then a number of words are selected for each topic. Therefore, it considers every document as a combination of topics. Both document-topic distribution (θ) and topic-word distribution (β) are controlled by (α) and (η) correspondingly. This process is repeated N times for each word in the document, and D times for each document in the collection [51, 59, 60].

Topic assignment phase includes several steps. First, the preprocessed document from the previous phase is used to create a corpus and dictionary, which are crucial steps for creating topic distributions in the document. Each word in the document is

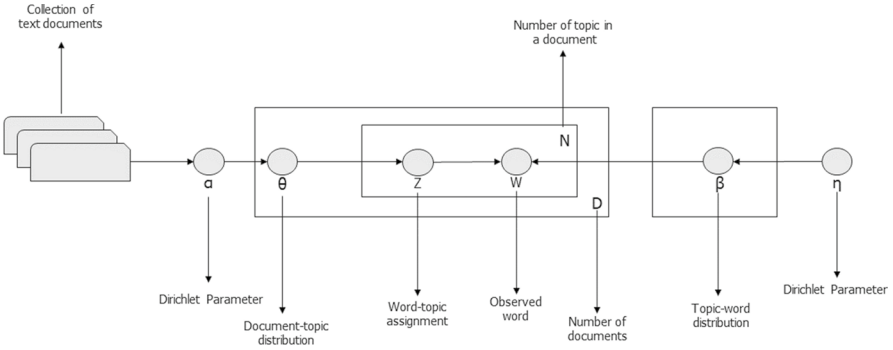

Fig. 3 Latent Dirichlet allocation [56]

mapped to its respective word ID using a dictionary. All words in the document are represented as a bag-of-words namely a corpus. The corpus consists of many tuples, where each tuple contains word ID and the corresponding count for each word in the document [54]. Second, after creating the corpus, the LDA model must be initialized and trained on that corpus. One of the most critical steps in initializing the LDA model is determining a suitable number of topics, for that we used the coherence score to determine the best number of topics. The coherence score is a qualitative measure used to assign a score for a single topic by measuring semantic similarity between high-scoring words in the topic [22]. Due to the semantic similarity of coherence, it better aligns and correlates with human interpretation unlike perplexity measure [61]. Topic modeling using LDA we used LDA model and the coherence model from the GENSIM library, and for the coherence score, we utilized the C_V measure due to its ability to effectively integrate various statistical methods such as sliding windows, pointwise mutual information (PMI), and cosine similarity. This measure assesses the degree of word co-occurrences within a specified context, which helps in understanding relationships between the dominant words of a topic; this feature makes it ideal for LDA-based model and considers both polysemy and synonymy. In addition, it is considered one of the most effective measures of coherence for topic modeling due to its strong correlation with human assessments of topic coherence. It is computed as the following equations [62]:

$$\text{C\_V} = \left(\frac{2}{N*(N-1)}\right)* \sum_{i<j} \text{NPMI(wi, wj)} \tag{1}$$

$$\text{NPMI(wi, wj)} = \frac{\text{PMI(wi, wj)}}{-\log(\text{p(wi, wj)})} \tag{2}$$

$$\text{PMI(wi, wj)} = \frac{\log(p(\text{wi, wj}))}{p(\text{wi})*p(\text{wj})} \tag{3}$$

where N is the number of top words in the topic, wi, wj are words in the top N, NPMI is the normalized pointwise mutual information, PMI is pointwise mutual information, p(wi, wj) is the probability of wi and wj happen simultaneously, *and* p(wi) and p(wj) are the individual specific probability of wi and wj.

Once the LDA is fitted on the corpus and the number of topics is determined, every sentence in the document is associated with a dominant topic. Finally, after assigning topics to sentences, the sentences with the same topic are grouped to facilitate the clustering task in the next phases.

3.1.3 Deep conceptualized embedding

In this phase, every sentence is associated with an n-dimensional vector composed of real values. The summarization methodology employs the BERT language model to generate contextualized embeddings for the sentences. BERT differs from other models by being bidirectional, which allows to concurrently evaluating context from both left and right. Also, It is built upon a transformer framework that employs a self-attention mechanism and encoder to create contextual representations of text [63]. Figure 4 illustrates BERT input Representation. For the proposed method, we tested various variants of BERT for mapping sentences to their conceptualized embeddings. These variants are S-BERT, PubMed BERT, BlueBERT, and SciBERT.

S-BERT (sentence BERT) is optimized for sentence-level comparison tasks by incorporating Siamese and triplet network architectures to learn semantic similarities and differences while processing pairs or triplets of sentences. SERT is suitable for handling large volumes of sentence comparisons quickly as it adapted BERT to reduce computational time and resources required [64] making S-BERT suitable for the proposed methodology. PubMed BERT is a specific domain language model that is pre-trained on PubMed abstracts and full articles using BERT architecture [44] which making it highly recommended for our biomedical dataset. For scientific text, SciBERT model was trained on large corpora including 1.14 Million papers from semantic scholars [45]. BlueBERT is also a domain-specific BERT model that is pre-trained on PubMed clinical notes and abstracts

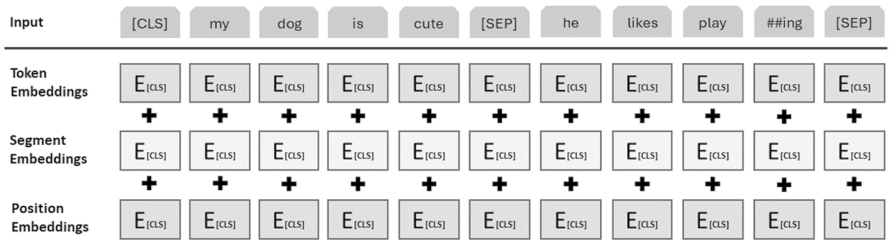

Fig. 4 BERT input representation [63]

Table 1 Comparison between BERT variants mentioned

Feature	S-BERT	PubMed BERT	SciBERT	BlueBERT
Domain	General	Biomedical	Scientific	Biomedical/clinical
Corpus	Wikipedia, news articles	PubMed abstracts, PMC full texts	Papers from semantic scholar	PubMed abstracts, MIMIC-III clinical notes
Training approach	Incorporates Siamese and triplet network architectures	Trained from scratch	Trained from scratch	Further pre-trained from general BERT
Year	2019	2020	2019	2019

from MIMIC-III [46]. Table 1 summarizes these variants based on different factors.

3.1.4 Clustering and sentence selection

After assigning topics to sentences, sentences with the same topic are gathered into a single list for clustering. The objective of our summarization method is to choose representative sentences from each topic to create the summary. To achieve this, we employ the K-medoid clustering in conjunction with an Euclidean-distance (ED) metric. Compared to K-means clustering, the K-medoid algorithm demonstrates a faster rate of convergence and is less susceptible to noise and outliers [22]. The ED metric evaluates the degree of correspondence between objects in a space of vectors by taking into account the size of vectors in various dimensions. Consider two vectors $X = \{\times 1, \times 2, \ldots, xN\}$ and $Y = \{y1, y2, \ldots, y_N\}$ represent contextualized representations of two sentences, the ED between them is computed as follows:

$$\text{Euclidean distance } (X,\ Y) = \sqrt{\sum_{i=1}^{N} (xi - yi)^2} \tag{4}$$

For sentence selection, the centroid of each cluster is extracted as it represents the most representative sentence for the topic among the other sentences. The sentence that represents the centroid combined from each cluster to form the final extractive summary.

Algorithm 1 provides the pseudo-code outlining the clustering and sentence selection process employed by the proposed method.

1. Two empty sets are initialized for clusters C and representative sentences TRS. The set C is created to store the collection of clusters obtained from K-medoid clustering, while TRS is created to store the top representative sentence selected from each cluster that exist in clusters C.
2. For each topic Ni contains a set of sentences representing this topic, if the number of these sentences is large than 1, we fitted K-medoid on sentences' vectors V obtained from BERT variant and number of clusters K (*lines 3,4,5*).
3. If number of sentences for each Ni is equal 1, the sentence directly added to representative sentences TRS (*lines 7,8*).
4. After fitting K-medoid, the created clusters Ci are added to clusters C set (*line 6*).
5. Iterating over C set, we captured the centroid of each cluster Ci as this centroid represent the most representative sentence for this topic (*lines 11, 12*).
6. The centroids from each cluster Ci are then added to TRS, where each centroid is a sentence (*line 13*).

7. Finally, the sentences in TRS are grouped together to build the final extractive summary.

Algorithm 1 Sentences clustering and selection algorithm employed by the proposed method

Input: Set of Topics N // a set of topics where each topic has associated sentences
Set of Vectors V // a set of vectors for sentences
Number of clusters K // the desired number of clusters
Output: Set of clusters C // the resulting clusters of sentences
Set of Top Representative Sentences TRS

1. *C* = ∅ // initialize an empty set for clusters
2. *TRS* = ∅ // initialize an empty set for top representative sentences
3. **For all** *Ni* in *N* **do** // iterate over each topic in N
4. **If** size of sentences assigned to *Ni* is greater than 1 **then**
5. Fit K-Medoid on *V, K*
6. Create Ci and add it to *C*
7. **Else do**
8. Add sentence to *TRS* //if there are one sentence in the topic add it to TRS
9. **End if**
10. **End for**
11. **For all** *Ci* in *C* **do** // iterate over each resulting cluster ci
12. Get *Ci* centroid // calculate the cluster centroid for each cluster in clusters C
13. Add *Ci* centroid to *TRS* // add centroid (representative sentence) to TRS
14. **End for**
15. **Return** *TRS* // return the final set of top representative sentences

3.2 Evaluation method

3.2.1 Evaluation corpus

There is currently no standardized, manually annotated corpus that can be used to assess this form of biomedical summarization. We followed the same method used by previous studies [7, 30]. We created an evaluation corpus by randomly retrieving 200 biomedical documents from the BioMed Central (Biomed) repository using the search keyword "Knee Osteoarthritis Management" to gather the relevant articles and research papers. Each document of 200 documents was preprocessed to remove the irrelevant content as described in Sect. 3.1.1. Then, each document was split into two parts (abstract and full text). The abstracts were saved in a folder named "Abstracts" where each abstract is stored as a separate txt file. Similarly, the full texts were saved in a folder named "Full-Texts" where each full document was stored as an individual txt file. The size of this corpus is sufficient to guarantee that the results are significant according to [65]. Table 2 describes the statistical characteristics of our corpus.

Table 2 Detailed statistical analysis of the introducing corpus characteristics

	No. of sentences			No. of words		
	Min.	Max.	Avg.	Min.	Max.	Avg.
Full texts	53	358	155	1374	12,105	4378
Abstracts	5	21	13	167	549	357

3.2.2 Quantitative analysis method

In this study, we used Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric for evaluating the quality of the generated extractive summary created by the introduced methodology. ROUGE evaluates the generated summary by quantifying the shared terms (Rouge-N) or longest common subsequence (Rouge-L) between the generated and reference summary [26]. In our experiments, we utilized three variants of ROUGE which are ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L). R-1 calculates number of unigrams (individual words) that appear in both the generated and the reference summaries, while R-2 considers number of bigrams (pairs of consecutive words) that appear in both the generated and the reference summaries [35]. For Rouge-L, it utilizes Longest common subsequence (LCS), which represents longest subsequence of words that appear in both generated and reference summaries in the same order, but not necessarily continuously [66]. For each variant, we calculated recall, precision, and f-score. We utilized the python rouge library to compute the mean f-score, precision, and recall for R-1, R-2, and R-L.

3.2.3 Qualitative analysis method

In addition to ROUGE metrics, we assessed the generated summaries using human evaluations. We asked eight orthopedic surgeons in the orthopedic surgery department (Faculty of Medicine, Egypt) to review and evaluate the generated summaries. For evaluation, 50 random summaries generated by BlueBERT were randomly selected. First, we collected different criteria from previous studies to evaluate the quality of the generated summaries as explained in Table 3. These criteria were provided to the eight reviewers to select the most suitable criteria for evaluation. Based on the reviewers' experience, the following criteria were chosen to evaluate the generated summaries of the medical documents: completeness, relevance, conciseness, informativity, and readability. After defining these criteria, all reviewers assessed the total 50 summaries and their corresponding reference summaries independently and were asked to rate each summary using a 5-point Likert scale, where (1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent) for each of the five previous criteria.

Table 3 Different criteria used for human evaluation

Criteria	Description	References
Coherence	Describes how well the summary' sentences well organized and connected to present the main information and ideas	[67]
Consistency	Represents to what extent the facts supplied in the source text are presented in the produced summary	[67]
Relevance	Describes how well the summary contains only the relevant facts	[67, 68]
Informativity	Defines how well the generated summary extracts important points from the reference summary	[69]
Conciseness	Represents how well the generated summary is clear, short, and captures the essential information without unnecessary details	[69, 70]
Readability	Describe how the summary is easy to read and understand	[69–71]
Completeness	Represents how well the summary covers all the key points and essential information from the reference summary	[68, 71]
Repetition	Represents how well the produced summary avoids repeating the same ideas in different sentences	[68]
Contradiction	Describe if there are any conflicting ideas or information represented in the summary	[68]

To ensure accuracy and consistency in the evaluation process, the reviewers were given a questionnaire containing the mentioned criteria and detailed guidelines explaining each criterion and what factors determine the various rating levels (Table 11 in Appendix). They were also instructed to read both generated and reference summaries carefully before providing their ratings. Once all reviewers completed the questionnaire, we calculated the mean scores for each criterion across the eight reviewers as shown in Fig. 6.

Also, we used Fleiss' Kappa to measure the inter-rater agreement among reviewers across multiple criteria, as it is suitable when multiple raters are involved. The values of Fleiss' Kappa range from -1 to 1, where higher values indicate stronger agreement. A Kappa value of 0 suggests no agreement beyond what could be expected by chance, while values above 0 indicate varying levels of agreement. Table 10 presents the inter-rater agreement among reviewers across multiple criteria.

4 Results and discussion

The proposed methodology was evaluated on the medical corpus presented in Sect. 3.2.1 to ensure its efficiency. For topic modeling and sentence clustering, the LDA and K-medoid implementations in scikit-learn were used. All data points were used simultaneously without any splits for training, testing, or validation since the suggested methodology is unsupervised.

To optimize the performance, we conducted experiments on the parameters specified in our algorithm. The optimization of LDA hyperparameters was conducted by evaluating different combinations of their values. We experimented (symmetric, asymmetric, auto) for α, and for β we utilized (auto). Consequently, we examined a different number of topics (between 2:10) and a fixed number of passes at 20 to balance the complexity and training time of the mode during the optimization process. For each combination, we trained LDA model, and then the coherence score was calculated using the Coherence Model from Gensim Library for every combination. Therefore, the combination of hyperparameters that result from the highest score of coherence measure was considered the best combination of LDA.

For the K parameter, representing the number of clusters required for sentence clustering, we tested several values ranging from 2 to 5 to select the appropriate number. For summary evaluation, we used three matrices of ROUGE as mentioned before R-1, R-2, and R-L. For each one of the Rouge matrices, the recall, precision, and F-score were calculated. By varying the parameter K and employing different BERT variants for mapping sentences to their conceptualized embeddings, as described in Sect. 3.1.3, the results obtained by our extractive method on the medical corpus are presented in Tables 4, 5, 6, 7 and 8. For the BERT variant models, Tables 4, 5, and 6 represent the recall, precision, and F-score values of R-1, R-2, and R-L.

As observed from Tables 4 and 5, when the number of clusters K increased, the recall scores for all variants models increased. This indicates that the generated

summary captured the essential contents represented in the reference summary. At $K=5$, the SciBERT gave the best results for recall scores compared to the other models. On the other hand, precision scores decreased when increasing the number of clusters, the lower values of precision indicate that the generated summary contains more irrelevant information compared to the reference summary. For precision, S-BERT resulted from the highest scores at $K=2$.

We considered F-score values to choose the K number of clusters and Bert model that led to a robust evaluation of the summarization system. F-score considers both recall and precision values, which provides a balanced evaluation metric that considers both the completeness and accuracy of the generated summaries. As shown in Table 6, the optimal number of clusters K varies across different Bert-based models. Generally, $K=3$ seems to provide the best scores for most models, in both R-1 and R-2 metrics, S-BERT and SciBERT showed strong f-score while BlueBERT outperformed the other models at $K=3$. However, increasing the number of clusters K beyond 3 tends to return a less relevant information. This suggested that using a moderate number of clusters K enables the models to capture relevant information and maintain summary coherence.

Additionally, we compared our extractive summarization method with two other similar methods, (1) a method proposed by Srivastava et al. [22]. This study applied word2vec vector representation in conjunction with topic modeling using LDA and K-medoid clustering for extractive summarization, evaluated on the WikiHow dataset, CNN-DailyMail dataset, and DUC 2002 dataset. (2) the second method was developed by Issam et al. [54], the method applied LDA for topic modeling but instead of using clustering, it used TextRank for summary generation. To evaluate the relevance of topic modeling using latent Dirichlet allocation (LDA) in the proposed methodology, we eliminate LDA from the framework and instead focus on summarization primarily based on K-medoid clustering and BlueBERT. The rationale behind this adjustment is to center the study on the impacts of LDA in measuring the performance and effectiveness of producing relevant summaries. Table 7 presents the Rouge scores achieved by the proposed method and other methods previously mentioned. As shown, the proposed model achieved R-1 score of 0.4838, significantly higher than word2vec method (0.4215), the TextRank method (0.3866), and clustering with BlueBERT method (0.3933), this suggested that the BlueBERT with LDA is more effective in capturing and generating relevant unigrams compared to the other techniques. R-2 which measures the overlap of bigrams, the proposed method also led to the best score of 0.2174. This improvement in R-2 compared to other methods indicates that the BlueBERT with LDA better captured the contextual relationships between pairs of words, leading to more coherent summaries. When comparing the proposed method with other mentioned studies, it gave the highest score of R-L (0.2206). This indicates that our method better captures the LCS between the generated and reference summary. Overall, the proposed method achieves the highest performance of R-1, R-2, and R-L scores.

Also, we calculated the average time taken by the suggested methodology and the comparison methods for summarizing our biomedical corpus which contains 200

Table 4 Recall scores achieved using various variants of BERT model and different values of K on the proposed corpus

K	S-BERT			PubMed BERT			BlueBERT			SciBERT		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
2	0.5029	0.2165	0.2252	0.5356	0.2338	0.2400	0.5391	0.2423	0.2439	0.5449	0.2440	0.2468
3	0.6024	0.2722	0.2676	0.6329	0.2907	0.2808	0.6370	0.2998	0.2861	0.6462	0.3034	0.2903
4	0.6680	0.3147	0.2995	0.6954	0.3363	0.3131	0.6989	0.3431	0.3206	0.7052	0.3439	0.3214
5	**0.7094**	**0.3451**	**0.3233**	**0.7361**	**0.3682**	**0.3373**	**0.7398**	**0.3748**	**0.3435**	**0.7482**	**0.3769**	**0.3458**

The highest score in each column is shown in bold and underline type

Table 5 Precision scores achieved using various variants of BERT model and different values of K on the proposed corpus

K	S-BERT			PubMed BERT			BlueBERT			SciBERT		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
2	**0.5099**	**0.2225**	**0.2389**	**0.4818**	**0.2133**	**0.2226**	**0.4944**	**0.2234**	**0.2289**	**0.4830**	**0.2176**	**0.2249**
3	0.4458	0.1990	0.2015	0.4191	0.1917	0.1873	0.4257	0.1981	0.1914	0.4159	0.1943	0.1880
4	0.3968	0.1835	0.1779	0.3711	0.1763	0.1657	0.3727	0.1793	0.1692	0.3659	0.1760	0.1655
5	0.3562	0.1696	0.1615	0.3341	0.1633	0.1507	0.3348	0.1664	0.1531	0.3268	0.1616	0.1482

The highest score in each column is shown in bold and underline type

Table 6 *F*-scores scores achieved using various variants of BERT model and different values of *K* on the proposed corpus

K	S-BERT			PubMed BERT			BlueBERT			SciBERT		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
2	0.4731	0.2043	0.2155	0.4766	0.2091	**0.2163**	0.4821	**0.2251**	0.2161	0.4801	0.2153	0.2159
3	**0.4812**	**0.2155**	0.2153	**0.4766**	0.2179	0.2118	**0.4838**	0.2174	**0.2206**	**0.4805**	**0.2245**	**0.2199**
4	0.4698	0.2185	0.2102	0.4598	0.2196	0.2056	0.4623	0.2238	0.2103	0.4595	0.2219	0.2081
5	0.4512	0.2163	**0.2460**	0.4391	**0.2162**	0.1989	0.4405	0.2202	0.2022	0.4364	0.2169	0.1989

The highest score in each column is shown in bold and underline type

Table 7 The Rouge values achieved by the proposed methodology and other methods.The highest score in each column is shown in bold type.

References	Method	Rouge-1	Rouge-2	Rouge-L
[22]	Topic modeling using word2vec	0.4215	0.1745	0.1828
[54]	Topic modeling and TextRank	0.3866	0.2126	0.1940
Current study	Clustering with BlueBERT	0.3933	0.1840	0.2180
Proposed Model	**Topic modeling using BLUE BERT**	**0.4838**	**0.2174**	**0.2206**

The highest score in each column is shown in bold

Table 8 The average time in seconds taken by each method for summarizing 200 biomedical documents

Method	Taken time
Topic modeling using word2vec	48.07
Topic modeling and TextRank	49.32
Clustering with BlueBERT	35.55
Topic modeling using BLUE BERT	**55.82**

The highest score in each column is shown in bold

biomedical documents. Table 8 represents the average time taken by each method. As illustrated, the proposed method took the longest time (55.82 s) compared to word2vec, TextRank and clustering with BlueBERT methods. This is due to the complexity of the BlueBERT model which leveraged deep learning and advanced language processing capabilities that require more computational resources and time and the iterative process of LDA that needs multiple passes through the data to assign topics and find the best hyperparameters for LDA model.

Moreover, we analyzed the number of topics captured across various document lengths to explore the impact of document length (number of words) on topic selection. This clarifies how different numbers of words influence the diversity of the identified topics. The observation showed that as the number of words increased, the number of topics decreased as illustrated in Table 9. In addition, Fig. 5 shows an inverse relationship between document length and the number of topics. Where short documents capture more topics than longer ones. The value of R^2 (0.699) suggests a moderately strong correlation between document length and the number of topics.

Rouge metric suffers from four main limitations: (1) it relies on n-gram overlap and discard the semantic meaning of the summary, (2) it lacks coherence and readability of the summary, (3) it requires human-written reference summaries for evaluation, (4) it cannot determine if the information represented in the summary is correct or not [72–74]. Due to these limitations, eight orthopedic surgeons were asked to evaluate 50 random summaries generated by the best-performance algorithm (BlueBERT). Eight orthopedic surgeons evaluated the summaries based on five criteria:

Table 9 The impact of document length (number of words) on topics selection

Doc	Number of words	Best number of topics	Coherence score
1	1374	10	0.4553
2	2550	10	0.3972
3	3731	9	0.4208
4	3751	8	0.4756
5	4054	8	0.4489
6	4194	7	0.4954
7	5200	7	0.3943
8	5574	7	0.4008
9	5934	6	0.4086
10	6263	5	0.5417
11	7379	5	0.4039
12	7335	3	0.4161
13	7393	4	0.3648
14	8346	3	0.5615
15	8786	3	0.4255
16	8802	2	0.3844
17	9038	2	0.3986
18	9971	2	0.5069

completeness, relevance, conciseness, informativity, and readability, as described in Sect. 3.2.3. The average scores for each criterion are calculated and shown in Fig. 6.

From results shown in Fig. 6, the summaries generated by our method meet the five criteria we established for evaluation. While comparing reference summaries

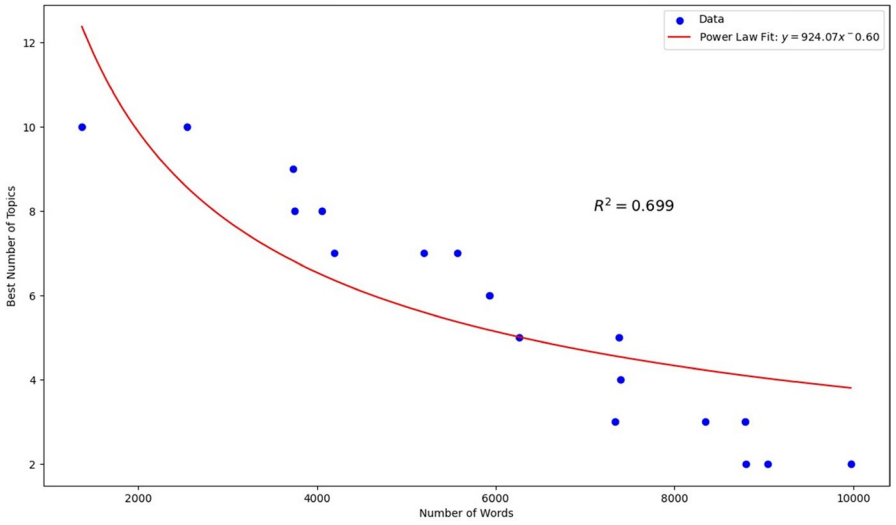

Fig. 5 Relationship between number of words and best number of topics

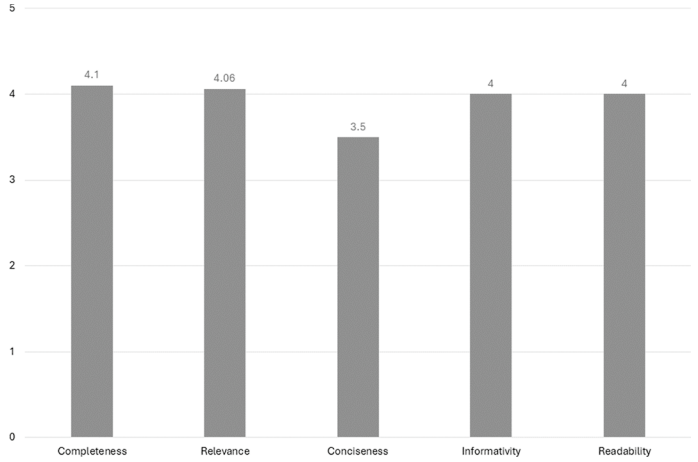

Fig. 6 Qualitative analysis results

Table 10 Inter-rater agreement among reviewers across multiple criteria

Criteria	Fleiss-Kappa	No. of raters	No. of summaries
Relevance	0.682	8	50
Completeness	0.68	8	50
Informativity	0.58	8	50
Readability	0.69	8	50
Conciseness	0.703	8	50

and generated summaries, we notice strong performance across all criteria. The average score of completeness is 4.10 which indicates that our generated summaries successfully cover the main points and important information represented in the reference summaries. The relevance score of 4.06 shows how well the salient features from the reference summaries have been captured in the generated summaries. Informativity and readability both received a score of 4.0 indicating that our generated summaries can encapsulate important information, easy to read and understand. Conciseness has a moderately high score of 3.55 which points out how concise, but not much more than that, the generated summaries are. The whole results indicate that our summarization method works effectively across all criteria.

The results presented in Table 10 provide confidence that, in most criteria, the evaluations were conducted reliably, reducing the likelihood that the outcomes were influenced by chance or bias. Relevance (0.682), completeness (0.680), readability (0.690), and conciseness (0.703) all demonstrated substantial agreement, meaning that the reviewers generally aligned in their assessments, while informativity (0.58) indicates moderate agreement. Example of the generated summary by the proposed methodology and its associated reference summary (paper's abstract) are shown in Figs. 7 and 8 in Appendix.

5 Conclusion

This study effectively proposed a new methodology for summarizing biomedical research papers. The proposed method combined topic modeling, clustering, and BERT. The methodology utilizes the online variant of LDA and K-medoids clustering. For vector representation, we tested and compared four variants of BERT (S-BERT, PubMed BERT, BlueBERT, SciBERT). Additionally, this study applied coherence in finding the most suitable number of topics and allocating them at the sentence-level rather than the word-level. The effectiveness of the proposed method was evaluated using two types of evaluation, (1) quantitative evaluation using metrics R-1, R-2, and R-L and (2) qualitative evaluation where eight orthopedic surgeons evaluated the generated summaries based on five criteria: completeness, relevance, conciseness, informativity, and readability. The evaluation was conducted on 200 biomedical research papers about knee osteoarthritis management collected randomly from BioMed central. The results showed that the proposed method outperformed the other models with 0.4838 (R-1), 0.2174 (R-2), and 0.2206 (R-L) for f-score measurement. Also, it meets the five criteria effectively with an average score of 4.10 (completeness), 4.06 (relevance), 4.0 (informativity and readability), and 3.55 (conciseness). Also, we applied Fleiss' Kappa to measure the inter-rater agreement among reviewers, and the results demonstrated substantial agreement, meaning that the reviewers generally aligned in their assessments as illustrated in Table 10. Furthermore, we conducted additional experiments to ascertain the significance of topic modeling using LDA within the proposed methodology by omitting the LDA and instead relying on K-medoid clustering and BERT-based models. The results presented in Table 8 indicate that the proposed methodology surpasses the other methods. Results indicate the efficacy of the topic-modeling method and K-medoids clustering in conjunction with BERT vectorizations over the Ranking algorithm (TextRank), static embedding (word2vec) and clustering with BERT-based. Moreover, we explored the impact of document length (number of words) on topic selection through analyzing the number of topics captured across various document length. The observation showed that as the number of words increased, the number of topics decreased, which explains the inverse relationship between the number of words and the number of topics, as illustrated in Table 9 and Fig. 5. The encouraging results of this study provide a strong basis for additional research. In the future, we can apply the proposed method for multi-document summarization. Given the poor R-2 scores and the bigram overlap in the generated summaries, future research may enhance its values. Finally, to improve the sentence selection phase, we can investigate variations in performance using different clustering, and embedding algorithms. In addition, we can investigate another criterion to select the top sentences from each cluster to build the final summary.

Appendix

See Figs. 7 and 8 and Table 11.

Generated summary

Female respondents with knee Osteoarthritis were found to have lower scores in most of the Quality of life dimensions with a significant difference seen in physical functioning ($p = 0.03$) even after adjusting for other factors ($p = 0.024$). One study looked at factors that were associated with physical functioning in symptomatic knee Osteoarthritis and the other analyzed factors influencing physical function in patients with knee and hip Osteoarthritis attending outpatient clinics. We have found that patients with knee Osteoarthritis attending primary care have relatively poor quality of life pertaining to the physical health components but there was less impact on the mental health of the patients. Abbreviations BMI: Body mass index; Bodily pain: Bodily pain; Dartmouth Primary Care Cooperative Information Project/World Organization of National Colleges, Academies, and Academic Associations of General Practice/Family Physicians HRQoL: Health-related quality of life; Osteoarthritis: Osteoarthritis; Physical health: Physical health; Quality of life: Quality of life; Rheumatoid arthritis: Rheumatoid arthritis; Role-emotion RP: Role-emotion; Role-physical: Role-physical; Social functioning: Social functioning; SF-36: 36-item short form; SIP: Sickness Impact Profile; Vitality function: Vitality function; Western Ontario McMaster Universities Arthritis Index: Western Ontario and McMaster Universities Osteoarthritis Index. Although there is no exact figure of patients with knee Osteoarthritis, the Community Orientated Program for Control of Rheumatic Diseases (COPCORD) study showed that 9.3% of adult Malaysians complained of knee pain with a sharp increase in pain rate to 23% in those over 55 years of age and 39% in those over 65 years. This is in agreement with a previous study that showed comorbidity had no association with pain or physical functioning. The affective and cognitive meaning of information that initially was experienced as threatening may be changed to make the present situation more acceptable. Patients who needed hospital admission or those with any other forms of lower limb immobility or abnormality such as paraplegia were also excluded. Health-related quality of life (HRQoL) is increasingly acknowledged as a valid health indicator in many diseases.

Fig. 7 Example of the generated summary by the proposed methodology

Reference summary

Measurement of health-related quality of life (Health related quality of life;) among patients with osteoarthritis (Osteoarthritis;) helps the health care provider to understand the impact of the disease in the patients' own perspective and make health services more patient-centered. The main aim of this study was to measure the quality of life among patients with symptomatic knee Osteoarthritis; attending primary care clinic. We also aimed to ascertain the association between socio-demographic and medical status of patients with knee Osteoarthritis; and their quality of life. A clinic based, cross sectional study using the Short Form-36 (Social functioning;-36) questionnaire was conducted in two primary care health clinics in Hulu Langat, Selangor, Malaysia over a period of 8 months. The nurses and medical assistants were involved in recruiting the patients while the family physicians conducted the interview. A total 151 respondents were recruited. The mean age was 65.6 ± 10.8 years with females constituted 119 (78.8%) of the patients. The mean duration of knee pain was 4.07 ± 2.96 years. Half of the patients were overweight and majority, 138 (91.4%), had at least one co-morbidity, the commonest being hypertension. The physical health status showed lower score as compared to mental health component. The domain concerning mental health components showed positive correlation with age. There was a significant negative correlation between age and physical functioning ($p < 0.0005$) which indicated the deterioration of this domain as patients became older. Male respondents had better scores in most of the Quality of life; dimensions especially in the physical functioning domain ($p = 0.03$). There was no significant association between Quality of life; with different education levels, employment status and marital status. Patients with higher body mass index (Body mass index;) and existence co-morbidities scored lower in most of the Quality of life; domains. This study has shown that patients with knee Osteoarthritis; attending primary care clinics have relatively poor quality of life pertaining to the physical health components but less impact was seen on the patients' mental health.

Fig. 8 The corresponding reference summary for the previous generated one

Table 11 Basics criteria and questions for qualitative measurement

Criterion	Question	Weight				
		Poor	Fair	Good	Very good	Excellent
		1	2	3	4	5
Completeness	How well the summary covers all the key points and essential information from the reference summary?					
Relevance	How well the summary contains only the relevant facts from reference summary?					
Conciseness	How well the generated summary is clear, short, and captures the essential information without unnecessary details?					
Informativity	How well the generated summary extracts important points from the reference summary?					
Readability	How the summary is easy to read and understand?					

Acknowledgements The authors would like to express their deep gratitude to Dr. Mohammad k. Saleh, Dr. Sayed Selim, Dr Mohammad El Eisawy, Dr. Amr Shwail Dr Mohammad Maged, Dr. Sameh Holyel, Dr. Fahmey Sameer, and Dr. Mohammad Osama (Orthopedic department at faculty of Medicine, Zagazig University, Egypt) for their support and help in evaluating the generated summaries.

Author contributions AA and SM participated in all concerning starting from the idea till preparing, writing, and enhancing the final manuscript, while NM and LS prepared required figures and tables and revised the manuscript.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability The online version of questionnaire used for human evaluation is available at https://forms.gle/6EngNMnCL7RCPwVB9. For BERT variants models used in this experiment, we utilized models from Hugging Face https://huggingface.co/models. The biomedical corpus created and introduced in this study is available at https://github.com/alyaaahmed21/Biomedical-Corpus. The stop words list used in the preprocessing phase is available at https://pubmed.ncbi.nlm.nih.gov/help/.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval We confirm that this manuscript has not been published elsewhere and is not under consideration by another journal. All authors have approved the manuscript and agree with its submission to Supercomputing Journal.

References

1. Givchi A, Ramezani R, Baraani-Dastjerdi A (2022) Graph-based abstractive biomedical text summarization. J Biomed Inform 132:104099. https://doi.org/10.1016/j.jbi.2022.104099
2. Davoodijam E et al (2021) MultiGBS: a multi-layer graph approach to biomedical summarization. J Biomed Inform 116:103706. https://doi.org/10.1016/j.jbi.2021.103706
3. Azadani MN, Ghadiri N, Davoodijam E (2018) Graph-based biomedical text summarization: an itemset mining and sentence clustering approach. J Biomed Inform 84:42–58. https://doi.org/10.1016/j.jbi.2018.06.005
4. Mishra R et al (2014) Text summarization in the biomedical domain: a systematic review of recent research. J Biomed Inform 52:457–467. https://doi.org/10.1016/j.jbi.2014.06.009
5. Du Y et al (2022) UGDAS: unsupervised graph-network based denoiser for abstractive summarization in biomedical domain. Methods 203:160–166. https://doi.org/10.1016/j.ymeth.2022.03.012
6. Moradi M, Dashti M, Samwald M (2020) Summarization of biomedical articles using domain-specific word embeddings and graph ranking. J Biomed Inform 107:103452. https://doi.org/10.1016/j.jbi.2020.103452
7. Moradi M, Ghadiri N (2018) Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artif Intell Med 84:101–116. https://doi.org/10.1016/j.artmed.2017.11.004

8. Karotia A, Susan S (2023) CovSumm: an unsupervised transformer-cum-graph-based hybrid document summarization model for CORD-19. J Supercomput 79(14):16328–16350. https://doi.org/10.1007/s11227-023-05291-3
9. Plaza L (2014) Comparing different knowledge sources for the automatic summarization of biomedical literature. J Biomed Inform 52:319–328. https://doi.org/10.1016/j.jbi.2014.07.014
10. Zhang H et al (2011) Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform 44(5):830–838. https://doi.org/10.1016/j.jbi.2011.05.001
11. Kirmani M et al (2024) Biomedical semantic text summarizer. BMC Bioinform 25(1):152. https://doi.org/10.1186/s12859-024-05712-x
12. Chaves A, Kesiku C, Garcia-Zapirain B (2022) Automatic text summarization of biomedical text data: a systematic review. Information 13(8):393. https://doi.org/10.3390/info13080393
13. Plaza L, Stevenson M, Díaz A (2012) Resolving ambiguity in biomedical text to improve summarization. Inf Process Manage 48(4):755–766. https://doi.org/10.1016/j.ipm.2011.09.005
14. Sarker A et al (2020) A light-weight text summarization system for fast access to medical evidence. Front Digit Health 2:585559. https://doi.org/10.3389/fdgth.2020.585559
15. Wang M et al (2021) A systematic review of automatic text summarization for biomedical literature and EHRs. J Am Med Inform Assoc 28(10):2287–2297. https://doi.org/10.1093/jamia/ocab143
16. Reeve LH, Han H, Brooks AD (2007) The use of domain-specific concepts in biomedical text summarization. Inf Process Manage 43(6):1765–1776. https://doi.org/10.1016/j.ipm.2007.01.026
17. Joshi A et al (2022) RankSum—an unsupervised extractive text summarization based on rank fusion. Expert Syst Appl 200:116846. https://doi.org/10.1016/j.eswa.2022.116846
18. Muniraj P, Sabarmathi K, Leelavathi R (2023) HNTSumm: hybrid text summarization of transliterated news articles. Int J Intell Netw 4:53–61. https://doi.org/10.1016/j.ijin.2023.03.001
19. Bani-Almarjeh M, Kurdy M-B (2023) Arabic abstractive text summarization using RNN-based and transformer-based architectures. Inf Process Manage 60(2):103227. https://doi.org/10.1016/j.ipm.2022.103227
20. Rohil MK, Magotra V (2022) An exploratory study of automatic text summarization in biomedical and healthcare domain. Healthc Anal 2:100058. https://doi.org/10.1016/j.health.2022.100058
21. Moradi M, Dorffner G, Samwald M (2020) Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. Comput Methods Programs Biomed 184:105117. https://doi.org/10.1016/j.cmpb.2019.105117
22. Srivastava R et al (2022) A topic modeled unsupervised approach to single document extractive text summarization. Knowl-Based Syst 246:108636. https://doi.org/10.1016/j.knosys.2022.108636
23. Tohalino JV, Amancio DR (2018) Extractive multi-document summarization using multilayer networks. Phys A 503:526–539. https://doi.org/10.1016/j.physa.2018.03.013
24. Suleiman D, Awajan A (2020) Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. Math Probl Eng 1:9365340. https://doi.org/10.1155/2020/9365340
25. Wang T et al (2024) A study of extractive summarization of long documents incorporating local topic and hierarchical information. Sci Rep 14(1):10140. https://doi.org/10.1038/s41598-024-60779-z
26. Singh S, Singh JP, Deepak A (2024) Supervised weight learning-based PSO framework for single document extractive summarization. Appl Soft Comput 161:111678. https://doi.org/10.1016/j.asoc.2024.111678
27. Onan A, Alhumyani HA (2024) FuzzyTP-BERT: enhancing extractive text summarization with fuzzy topic modeling and transformer networks. J King Saud Univ Comput Inform Sci 2024:102080. https://doi.org/10.1016/j.jksuci.2024.102080
28. Jiang X, Dreyer M (2024) CCSUM: a large-scale and high-quality dataset for abstractive news summarization. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1. https://doi.org/10.18653/v1/2024.naacl-long.406
29. Zhang H, Yu PS, Zhang J (2024) A systematic survey of text summarization: from statistical methods to large language models. arXiv preprint arXiv 2406:11289. https://doi.org/10.48550/arXiv.2406.11289
30. Rouane O, Belhadef H, Bouakkaz M (2019) Combine clustering and frequent itemsets mining to enhance biomedical text summarization. Expert Syst Appl 135:362–373. https://doi.org/10.1016/j.eswa.2019.06.002

31. Alanzi E, Alballaa S (2023) Query-focused multi-document summarization survey. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2023.0140688
32. Sharma G, Sharma D (2022) Automatic text summarization methods: a comprehensive review. SN Comput Sci 4(1):33. https://doi.org/10.1007/s42979-022-01446-w
33. El-Kassas WS et al (2021) Automatic text summarization: a comprehensive survey. Expert Syst Appl 165:113679. https://doi.org/10.1016/j.eswa.2020.113679
34. Bedi PPS, Bala M, Sharma K (2022) Extractive summarization using concept-space and keyword phrase. Expert Syst 39(10):e13110. https://doi.org/10.1111/exsy.13110
35. Moradi M, Ghadiri N (2017) Quantifying the informativeness for biomedical literature summarization: an itemset mining method. Comput Methods Programs Biomed 146:77–89. https://doi.org/10.1016/j.cmpb.2017.05.011
36. Moradi M (2018) Frequent itemsets as meaningful events in graphs for summarizing biomedical texts. In: 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE. pp 1–6. https://doi.org/10.1109/ICCKE.2018.8566651
37. Moradi M (2019) Small-world networks for summarization of biomedical articles. arXiv preprint arXiv: 1903:02861. https://doi.org/10.48550/arXiv.1903.02861
38. Moradi M (2018) CIBS: a biomedical text summarizer using topic-based sentence clustering. J Biomed Inform 88:53–61. https://doi.org/10.1016/j.jbi.2018.11.006
39. Du Y et al (2020) Biomedical-domain pre-trained language model for extractive summarization. Knowl Based Syst 199:105964. https://doi.org/10.1016/j.knosys.2020.105964
40. Kanwal N, Rizzo G (2022) Attention-based clinical note summarization. In: Proceedings of the 37th ACM/SIGAPP symposium on applied computing. https://doi.org/10.1145/3477314.3507256
41. Padmakumar V, He H (2021) Unsupervised extractive summarization using pointwise mutual information. arXiv preprint arXiv:2102.06272. https://doi.org/10.18653/v1/2021.eacl-main.213
42. Xie Q et al (2022) Pre-trained language models with domain knowledge for biomedical extractive summarization. Knowl Based Syst 252:109460. https://doi.org/10.1016/j.knosys.2022.109460
43. Lee J et al (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240. https://doi.org/10.48550/arXiv.1901.08746
44. Gu Y et al (2021) Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 3(1):1–23. https://doi.org/10.48550/arXiv.2007.15779
45. Beltagy I, Lo K, Cohan A (2019) SciBERT: a pretrained language model for scientific text. arXiv preprint arXiv:1903.10676. https://doi.org/10.48550/arXiv.1903.10676
46. Peng Y, Yan S, Lu Z (2019) Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474. https://doi.org/10.18653/v1/W19-5006
47. Huang K, Altosaar J, Ranganath R (2019) Clinicalbert: modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342. https://doi.org/10.48550/arXiv.1904.05342
48. Chen YP et al (2020) Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med Inform 8(4):e17787. https://doi.org/10.2196/17787
49. Meng Z et al (2021) Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT. arXiv preprint arXiv:2109.04810. https://doi.org/10.48550/arXiv.2109.04810
50. Lin CY, Hovy E (2000) The automated acquisition of topic signatures for text summarization. In: COLING 2000 volume 1: the 18th International Conference on Computational Linguistics. https://aclanthology.org/C00-1072
51. Jelodar H et al (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multim Tools Appl 78:15169–15211. https://doi.org/10.1007/s11042-018-6894-4
52. Han C, Feng J, Qi H (2024) Topic model for long document extractive summarization with sentence-level features and dynamic memory unit. Expert Syst Appl 238:121873. https://doi.org/10.1016/j.eswa.2023.121873
53. Xie Q, Tiwari P, Ananiadou S (2023) Knowledge-enhanced graph topic transformer for explainable biomedical text summarization. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2023.3308064
54. Issam KAR, Patel S (2021) Topic modeling based extractive text summarization. arXiv preprint arXiv:2106.15313. https://doi.org/10.48550/arXiv.2106.15313

55. Liu N et al (2014) Topic-sensitive multi-document summarization algorithm. In: 2014 sixth international symposium on parallel architectures, algorithms and programming. IEEE. https://doi.org/10.1109/PAAP.2014.22
56. Blei DM, Ng AY, Jordan I (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
57. Barcos-Redín L et al (2025) Topic-based engagement analysis: focusing on hotel industry Twitter accounts. Tour Manage. https://doi.org/10.1016/j.tourman.2024.104981
58. Xue Y et al (2024) A LDA-based social media data mining framework for plastic circular economy. Int J Comput Intell Syst 17(1):8. https://doi.org/10.1007/s44196-023-00375-7
59. Chen X et al (2024) Exploring hot topics and evolutionary paths in the diagnosis-related groups (DRGs) field: a comparative study using LDA modeling. BMC Health Serv Res 24(1):756. https://doi.org/10.1186/s12913-024-11209-3
60. Tong Z, Zhang H (2016) A text mining research based on LDA topic modelling. In: International Conference on Computer Science, Engineering and Information Technology. https://doi.org/10.5121/csit.2016.60616
61. Mimno D et al (2011) Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. https://aclanthology.org/D11-1024
62. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. https://doi.org/10.1145/2684822.2685324
63. Devlin J et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805
64. Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. https://doi.org/10.48550/arXiv.1908.10084
65. Lin CY (2004) Looking for a few good metrics: automatic summarization evaluation-how many samples are enough?. In: NTCIR. https://api.semanticscholar.org/CorpusID:11314673
66. Rhazzafe S et al (2024) Hybrid summarization of medical records for predicting length of stay in the intensive care unit. Appl Sci 14(13):5809. https://doi.org/10.3390/app14135809
67. Searle T et al (2023) Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. J Biomed Inform 141:104358. https://doi.org/10.1016/j.jbi.2023.104358
68. Zhang L et al (2021) Leveraging pretrained models for automatic summarization of doctor–patient conversations. arXiv preprint arXiv:2109.12174. https://doi.org/10.48550/arXiv.2109.12174
69. Zhang N et al (2020) Summarizing Chinese medical answer with graph convolution networks and question-focused dual attention. In: Findings of the association for computational linguistics: EMNLP 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.2
70. Hu J et al (2021) Word graph guided summarization for radiology findings. arXiv preprint arXiv:2112.09925. https://doi.org/10.18653/v1/2021.findings-acl.441
71. Hu J et al (2022) Graph enhanced contrastive learning for radiology findings summarization. arXiv preprint arXiv:2204.00203. https://doi.org/10.48550/arXiv.2204.00203
72. Lin CY (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out
73. Schluter N (2017) The limits of automatic summarisation according to rouge. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics
74. Ng JP, Abrecht V (2015) Better summarization evaluation with word embeddings for ROUGE. arXiv preprint arXiv:1508.06034. https://doi.org/10.18653/v1/D15-1222