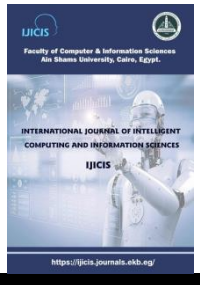
IJICIS, Vol.23, No.2, 50-61
DOI: 10.21608/ijicis.2023.190004.1252

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



A SYSTEMATIC REVIEW ON TEXT SUMMARIZATION OF MEDICAL RESEARCH ARTICLES

Alshimaa.M.Ibrahim
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
alshimaa.mohamed@cis.asu.edu.eg

Marco Alfonse
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Laboratoire Interdisciplinaire de l'Université Française d'Égypte (UFEID LAB), Université Française d'Égypte, Cairo, Egypt
marco_alfonse@cis.asu.edu.eg

Mostafa Mahmoud Aref
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
mostafa.araf@cis.asu.edu.eg

Received 2023-01-27; Revised 2023-03-17; Accepted 2023-03-20

***Abstract**: The term "Medical Text summarization" refers to the process of extracting or collecting more useful information from medical articles in a concise manner. Every day, the count of medical publications increases continuously, and applying text summarization techniques can minimize the time needed to manually transform medical papers into a summarized version. This study's goal is to present a summary of recent works in medical text summarization from 2018 to 2022. It includes 15 papers covering different methodologies such as Clinical Context-Aware (CCA), Prognosis Quality Recognition (PQR), Bidirectional Encoder Representations From Transformers (BERT), Generative Adversarial Networks (GAN), Recurrent Neural Network (RNN), and Sequence-To-Sequence (seq-2-seq) model. Also, the paper describes the newest datasets (PubMed, arXiv, SUMPUBMED, Evidence-Based Medicine Summarization, COVID-19 Open Research, BioMed Central, Clinical Context-Aware, Biomedical Relation Extraction Dataset, Semantic Scholar Open Research Corpus, and Prognosis Quality Recognition) and evaluation metrics (Recall-Oriented Understudy for Gisting Evaluation (ROUGE), F1 Metric, Bilingual Evaluation Understudy (BLEU), BERTScore (BS), and Accuracy) used in medical text summarization.*

***Keywords:** Text Summarization, Machine Learning, Natural Language Processing, Medical Papers*

1. Introduction

Nowadays, many fields have made extensive use of machine learning. It is a branch of computer science based on computational learning theory in artificial intelligence. "It is a field of study that gives computers the capability to learn without being explicitly programmed" [1]. It's used in several natural language processing applications, including machine translation and text summarization [2]. Text summarization is defined as "The process of extracting or collecting important information from the original text and

presenting that information in the form of a summary". Text summarization consists of two methods; extractive and abstractive summarization. Extractive summarization is a method for choosing sentences from the document depending on sentence and word features, then combining them to produce a summary. Abstractive summarization is a method for understanding the core ideas of a specific document and then presenting those ideas clearly and naturally. So several applications, such as medical reports, search engines, and industry reviews, now require text summarization. Summarization allows for obtaining the information needed in less time [3]. Text summarization is now a crucial tool for reducing and deciphering information from texts. Every day the publications in the field of medicine are increasing and using text summarization approaches help to reduce the time needed to manually convert medical papers to a summarized version. The recent datasets utilized in the text summarization include PubMed which contains 133000 documents, arXiv includes 215000 documents, SUMPUBMED consists of 26 million biomedical research papers, Evidence-Based Medicine Summarization has 2707 single-document, COVID-19 Open Research includes over 350000 full-text documents, BioMed Central includes more than 250 scientific journals, Clinical Context-Aware contains 173000 documents, Biomedical Relation Extraction Dataset has 600 documents, Semantic Scholar Open Research Corpus contains 81.1 million papers, and Prognosis Quality Recognition consists of 2686 documents.
The structure of this paper presents as follows: the second section discusses the different datasets, evaluation methods, metrics, and summarization models used in medical text summarization. The third section provides a discussion of the newest summarization models that are utilized in medical text summarization. The fourth section shows the conclusion and future work.

2. Medical Text Summarization

2.1. Different Datasets

The datasets utilized in the medical text summarizer are briefly abbreviated in the following section.

2.1.1 PubMed dataset

The PubMed dataset consists of Extensible Markup Language (XML) files that belong to the PubMed Central (PMC) repository's open-access collection. The dataset contains 133000 documents with abstracts. The average length of the abstract text is 214 words, and the average length of the full text is 3224 words [4].

2.1.2 arXiv dataset

The arXiv dataset consists of LATEX files that belong to the arXiv repository of electronic preprints. There are 215000 abstracted documents in the dataset. The average length of the full text is 6913 words, and the average length of the abstract is 292 words [4].

2.1.3 SUMPUBMED dataset

The SUMPUBMED dataset contains 26 million biomedical research papers extracted from PubMed. The papers come from various sources, including online books, MEDLINE, and life science journals. The dataset is separated into 3 categories: train (93%), test (3%), and validation (4%) [5].

2.1.4 EBMSummariserCorpus dataset

The Evidence-Based Medicine Summarization (EBMSummariserCorpus) dataset is a public dataset that contains 2707 single-document summaries. The dataset is a repository of data from the Journal of Family Practice (JFP). It has 1388 training records and 1319 evaluation records [6].

2.1.5 COVID-19 Open Research Dataset

The COVID-19 Open Research dataset is a free resource that contains over 1000000 scientific papers about COVID-19 and the coronavirus family of viruses, including over 350000 full-text documents. The dataset contains papers from 1970 to 2022 [7].

2.1.6 PQR dataset

The Prognosis Quality Recognition (PQR) dataset is collected from the scientific documents of the PubMed dataset that are delicate for summarization [8]. It contains 2686 documents, and 697 positive records (scientifically delicate).

2.1.7 CCA dataset

The National Center of Biotechnology Information's (NCBI) PubMed and the Biomedical Natural Language Processing (BioNLP) dataset is combined to create the Clinical Context-Aware (CCA) dataset. The dataset contains 173000 documents where 131000 documents come from BioNLP and 42000 documents from NCBI PubMed [8].

2.1.8 BMC dataset

BioMed Central (BMC) is an open-access publisher that provides over 250 scientific journals. It currently publishes online for all its journals. The first and biggest open data science publisher is BioMed Central. It was founded in 2000, now known as Springer Nature, and has owned it since 2008 [9].

2.1.9 BioRED dataset

The Biomedical Relation Extraction Dataset (BioRED) is an automated relation extraction from biomedical research papers. It is the first kind of biomedical relation extraction dataset, which includes a variety of entity types like (gene-protein, chemical, and disease) and relation pairings like (chemical-chemical and gene-disease) in the document. It is a set of 600 PubMed abstracts [10].

2.1.10 S2ORC dataset

The Semantic Scholar Open Research Corpus (S2ORC) dataset is the greatest collection of scholarly papers in English that are openly accessible and span many different academic fields. It contains 1.5 million LATEX source files, 8.1 million open-access PDFs, 81.1 million papers, and 380.5 million resolved citation links. The corpus has various academic fields including the biomedical and computer science fields [11].

2.2. Evaluation Methods

Text summarization model evaluation in Natural Language Processing (NLP) uses the following metrics.

2.2.1. ROUGE Metric

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric compares a summary's quality to that of other gold summaries generated by people to automatically determine its quality. It contains different measures such as ROUGE-N and ROUGE-L.

ROUGE-N determines how many overlapping units there are like word pairs, n-grams, and word sequences between the model-generated summary and the gold summaries generated by people to be reviewed [12]. The formula of ROUGE-N is shown in Eq. (1).

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferemceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{ReferemceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \quad (1)$$

where $\mathbf{gram_n}$ is the most n-grams that can co-occur in a candidate summary, The length of the n-gram is represented by n, and $\mathbf{Count_{match}(gram_n)}$ is a collection of reference summaries.

The ROUGE-L value represents the similarity between two sequences. It depends on the Longest Common Subsequence (LCS). The LCS issue selects the longest co-occurrence in sequence n-grams by considering sentence-level structure consistency. The ROUGE-L formula is represented in Eq.(2).

$$\text{ROUGE-L} = \frac{(1+\beta^2)\,\text{Recall} * \text{Precision}}{\text{Recall} + \beta^2 * \text{Precision}} \quad (2)$$

where β adjusts the significance of recall and precision concerning each other and is set to a high value. Eq.(3) and Eq.(4) represent Recall, and Precision respectively.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

where FN is a False Negative.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

where TP is a True Positive and FP is a False Positive.

2.2.2 F1 Metric

F1 Metric is defined as the symmetric mean of recall and precision. The formula for the F1 metric is represented in Eq.(5).

$$\text{F1 Metric} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

2.2.3 Accuracy

Accuracy is the percentage of correctly classified items to all dataset values. Its formula is represented in Eq.(6).

$$\text{Accuracy} = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ records} \quad (6)$$

2.2.4 BLEU Metric

The Bilingual Evaluation Understudy (BLEU) metric compares the candidate's n-grams to the reference translation's n-grams and counts the number of matches [13]. The BLEU metric formula is in Eq.(7).

$$\text{BLUE} = \left(\int_{e^{(1-r/c)}}^{1} \begin{matrix} if\ c > r \\ if\ c \le r \end{matrix}\right) .\ \boldsymbol{exp}\left(\sum_{n=1}^{N} \boldsymbol{w_n}\ \boldsymbol{log}\ \boldsymbol{p_n}\right) \quad (7)$$

where r is the length of the effective reference corpus, and c is the candidate translation's length, $\boldsymbol{p_n}$ is an n-gram precision, employing n-grams up to length N, and positive weights $\boldsymbol{w_n}$.

2.2.5 BERT Score (BS)

The BERT Score (BS) is a text-generation evaluation metric. It calculates a similarity score in tokens of the candidate sentence and tokens of the reference sentence as a sum of the cosine similarities between their token embeddings [14]. The BS evaluation metric formula is in Eq.(8).

$$\text{RecallBERT} = \frac{1}{|x|}\sum_{x_i \in x} max_{y_j \in y}\, x_i y_j \ ,\ \text{PrecisionBERT} = \frac{1}{|y|}\sum_{y_i \in y} max_{x_j \in x}\, y_i x_j$$
$$\text{F1-BERT} = 2 * \frac{\text{PrecisionBERT} * \text{RecallBERT}}{\text{PrecisionBERT} + \text{RecallBERT}} \quad (8)$$

where x is the reference sentence, and y is the candidate sentence.

2.3 Summarization Models

Gidiotis et al [4] present a Divide-ANd-ConquER (DANCER) approach for summarization of long-documents. They combine the DANCER approach with multiple summarization models like seq2seq, Recurrent Neural Networks (RNNs), and transformers. They generated three models called DANCER Long Short-Term Memory (LSTM), DANCER Rotational Unit of Memory (RUM), and DANCER Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence (PEGASUS). The authors used PubMed and arXiv datasets. For evaluation, they used ROUGE metrics. The output of the DANCER LSTM, DANCER RUM, and DANCER PEGASUS models is displayed in Table 1.

Table 1. The result of the DANCER LSTM, DANCER RUM, and DANCER PEGASUS models (adapted from [4])

adModels	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
DANCER LSTM	PubMed	44.09	17.69	40.27
	arXiv	41.87	15.92	37.61
DANCER RUM	PubMed	43.98	17.65	40.25
	arXiv	42.7	16.54	38.44
DANCER PEGASUS	PubMed	46.34	19.97	42.42
	arXiv	45.01	17.60	40.56

Gupta et al [5] created a new dataset called SUMPUBMED. The dataset was generated using scientific publications from the PubMed archive. The authors split the dataset into 3 parts: train (93%), test (3%), and validation (4%). The seq2seq model and the Coverage (+cov) mechanism were used to evaluate the dataset. They used ROUGE metrics for evaluation. ROUGE-1's score is 40.13, ROUGE-2's is 13.77, and ROUGE-score L's is 36.73.

Afzal et al [8] provided a solution for the restrictions of automatic text summarization by collecting accurate data from published biomedical resources. The authors created a Clinical Context-Aware (CCA) classifier and Prognosis Quality Recognition (PQR) model as a bidirectional long-short-term memory recurrent neural network. The authors used PQR and CCA datasets which are large datasets collected from the PubMed dataset. They divided datasets into 90 % training and 10 % testing. They used F1-metric and accuracy for evaluation. The PQR model had a 95.41 percent accuracy rate and the F1- metric =

96.93. A 93 percent accuracy rate was achieved with the CCA model and the F1- metric = 94.

Mahsa et al [15] improved the summarization system performance using a combination of the coreference resolution method, and Recurrent Neural Network (RNN). The authors used the COVID-19 open research dataset which contains over 59000 scientific papers that have been published, with over 47000 studies on COVID-19. They evaluated the system using the ROUGE metric. The model result is ROUGE-1= 0.53.

Moradi [16] extracted biomedical topics from the given documents where the goal was to find the main topics using an itemset mining algorithm and using a clustering algorithm. For text mining research, they employed a single-document corpus of 400 scientific biomedical articles from the BMC corpus. The evaluation is obtained by the ROUGE metric with a different number of Final Clusters (FCs). The scores gained by FC values 3, 2, and 4 are significantly higher than those of other values of FCs in single or multi documents (using FC value of 3 gives a score of ROUGE-2 = 0.3475 for single-document and ROUGE-2 = 0.2791 for multi-document, FC value of 2 gives a score of ROUGE-2 = 0.3392 for single-document and ROUGE-2 = 0.2654 for multi-document and FC value of 4 gives a score of ROUGE-2 = 0.3321 for single-document and ROUGE-2 = 0.2730 for multi-document).

Kieuvongngam et al [17] reduced the gap between researchers and the continuously increasing number of publications that use BERT, pre-trained NLP models, and an Open-source Artificial Intelligence Generative Pre-trained Transformer - 2 (OpenAI GPT-2). They used the COVID-19 open research dataset. This dataset includes more than 59000 research publications, including more than 47000 full-text documents about COVID-19 or associated diseases. The authors used ROUGE Metrics for evaluation. The ROUGE revealed that the 60% abstractive group is higher than the 40% group.

Milošević et al [18] provided and compared machine learning-based such as (Naive Bayes, DistilBERT, Random Forests, T5, PubMedBERT, and SciFive-based models) and rule-based methods to enhance the performance of biomedical text summarization. They utilized balanced and unbalanced datasets such as BioRED. The dataset was divided into 10% for testing and 90% for training. The PubMedBERT-based and distilBERT-based models obtained the best result with F1-score equal to 0.92 and 0.89.

Sarker et al [19] created a simple and fast text summarization system. The authors applied a word embedding model on a publicly available dataset for evidence-based medicine called EBMSummariserCorpus, which contains 2707 single-document summaries. The dataset has 1,388 records for training, and 1,319 for evaluation. The authors used the F1 metric for evaluation and the word-phrase embedding model result is F1 = 0.166.

Cai et al [20] proposed a new model based on the SciBERT-base model called the COVIDSum model. They collected salient sentences, created word co-occurrence graphs, and utilized a SciBERT-based sequence encoder. The authors used the COVID-19 Open Research dataset. The dataset was divided into training (114415), validation (6477), and testing (6356). The authors evaluated the model using ROUGE metrics where ROUGE-1 = 44.56, ROUGE-2= 18.89, and ROUGE-L = 36.53.

Xie et al [21] utilized multiple Pre-trained Language Models (PLMs) such as (BERT, RoBERTa, BioBERT, and PubMedBERT) to create a knowledge infusion training framework called KeBioSum for the challenge of extracting summarization of biomedical papers. They used PubMed, COVID-19, and S2ORC. The datasets were split into 75% for the training, 15% for the validation, and 10% for the test. The authors utilized ROUGE metrics and BERT score (BS) for evaluation. The result showed in Table 2**.**

Table 2. The result of the BERT, RoBERTa, BioBERT, and PubMedBERT models (adapted from [21])

Models	Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BS
BERT	PubMed	35.12	14.54	31.80	-
	COVID-19	30.79	10.37	25.13	-
	S2ORC	33.27	14.33	30.29	-
RoBERTa	PubMed	35.08	14.69	31.78	-
	COVID-19	30.10	10.72	27.81	-
	S2ORC	33.57	15.59	30.54	
BioBERT	PubMed	35.09	14.62	31.82	-
	COVID-19	31.11	10.74	27.82	-
	S2ORC	34.47	15.62	31.51	
PubMedBERT	PubMed	36.39	16.27	33.28	59.96
	COVID-19	32.04	12.61	29.10	53.56
	S2ORC	37.44	16.72	34.08	56.81

Moravvej et al [22] proposed a supervised extractive summary method depending on Generative Adversarial Networks Summarization (GAN-Sum) and Embedding Generative Adversarial Networks Summarization (E-GAN-Sum). The authors used 500 medical articles chosen from PubMed, and the database was split into three sets, each having 134, 100, and 266 samples, for the testing, validation, and training samples. The authors evaluated the result using ROUGE metrics. The GAN-Sum model's output is ROUGE-1= 40.86 and ROUGE-2= 24.59 and the E-GAN-Sum model's result is ROUGE-1=43.78 and ROUGE-2=26.73.

Moradi et al [23] demonstrated how contextualized embeddings generated by a deep bidirectional language model might be utilized to measure useful information in biomedical text summarization. The authors also showed that employing a Bidirectional Encoder Representations from Transformers (BERT)-based summarizer can increase the biomedical summarization performance. The authors utilized both the BERT model and a clustering method. They created a new corpus that randomly collected from BioMed Central (BMC) 3000 articles and generated a development corpus including 1000 articles that used the abstracts as model summaries. They used the ROUGE metric for the evaluation process and the result is ROUGE-1=0.7504, ROUGE-2=0.3312.

Song and Yongbin [24] combined the Sequence-To-Sequence (seq2seq) model with a classical keywords extraction method and the attention mechanism for detecting a summarization of medical papers. The authors used the COVID-19 Open Research dataset that contains 38937 title-abstract. The dataset was separated into training (36257) and testing (2680). The authors evaluated the model using BLEU and ROUGE metrics where ROUGE-1 = 30.16, ROUGE-2= 7.73, ROUGE-L = 28.40, and BLEU = 12.62.

Turky et al [25] proposed an abstractive summary for Covid-19 papers. The authors used LSTM and the glove model to improve the summary performance. The dataset utilized in the experiment is named COVID-19 dataset. Rouge metrics evaluated the experiment. The result of Rouge-1 = 43.6, Rouge-2=36.7 ,and Rouge-L =43.6.

Davoodijam et al [26] proposed a domain-specific method based on a multi-layer graph using the MultiRank algorithm. The authors utilized 450 biomedical papers from BioMed Central. The model was evaluated by ROUGE and BERTScore. The result of Rouge-1 = 0.164 , Rouge-2 = 0.052 , Rouge-L =0.146 ,and BERTScore = 0.806.

3. Discussion

This section will discuss the newest publications that have been shown in the paper. Table 3 shows a comparison between publications that have been recently utilized for data preprocessing, datasets, methodology, and evaluation while summarizing medical texts.

The basic functions of data preprocessing are stemming, stop word removal, sequence marker, tokenization, lemmatization, feature creation, cleaning, and vector generation. The methodologies were used to summarize several types of biomedical text, including single document, long document, and multi-document summarization.

Table 3. Comparison between most recent papers.

Reference	Data preprocessing	Methodology	Dataset	Evaluation
Gidiotis et al, 2020 [4]	NA	seq2seq, RNNs, transformers DANCER LSTM DANCER RUM DANCER PEGASUS	PubMed and arXiv	Results in Table 1
Gupta et al, 2021 [5]	Remove non-textual content.	Sequence-To-Sequence (seq2seq) model + the coverage (+cov) mechanism	SUMPUBMED	ROUGE-1 = 40.13 ROUGE-2 = 13.77 ROUGE-L= 36.73
Afzal et al, 2020 [8]	1- Sequence marker 2- Cleaning 3- Tokenization 4- Vector generation 5- Feature creation	1- CCA 2- PQR	CCA PQR	F1- metric = 94 Accuracy = 93% F1- metric = 96.93 Accuracy = 95.41%
Mahsa et al, 2021 [15]	1- Sentence Splitting 2- Tokenization 3- Stemming 4- Stop words removal	1-RNN 2-Coreference resolution method	COVID-19 Open Research dataset	ROUGE-1= 0.53
Moradi, 2018 [16]	1- Separate sentences 2- Tokenization 3- Remove unnecessary parts	Clustering and Itemset mining based Biomedical Summarizer (CIBS)	BMC	single-document ROUGE-2 = 0.3475 multi-document ROUGE-2 = 0.2791
Kieuvongngam et al, 2020 [17]	Token classification	1- BERT model 2-Abstractive Summarization GPT-2 model	COVID-19 open research dataset	The abstractive group is 60% larger than the 40% group
Milošević et al 2023 [18]	NA	DistilBERT PubMedBERT	BioRED	F1-score = 0.92 F1-score = 0.89
Sarker et al, 2020 [19]	NA	word/phrase embedding model	EBMSummariser Corpus	F1 = 0.166

Cai et al 2022 [20]	Remove wrong papers	COVIDSum	COVID-19 Open Research Dataset	ROUGE-1 = 44.56 ROUGE-2= 18.89 ROUGE-L = 36.53
Xie et al 2022 [21]	NA	BERT RoBERTa BioBERT PubMedBERT	PubMed COVID-19 S2ORC	Results in Table 2
Moravvej et al, 2021 [22]	1-Stop word removal 2-Stemming	1- GAN-Sum 2-E-GAN Sum	PubMed	GAN-Sum ROUGE-1=40.86 ROUGE-2=24.59 E-GAN-Sum ROUGE-1=43.78 ROUGE-2=26.73
Moradi et al, 2020 [23]	The section and subsection headings, figures, tables, and other non-major items are removed from the articles.	1-BERT 2-Sentence Clustering	BMC	ROUGE-1=0.7504 ROUGE-2=0.3312
Song and Yongbin, 2020 [24]	Spacy tokenizer	Sequence-To-Sequence (seq2seq) model	COVID-19 Open Research Dataset	ROUGE-1=30.16 ROUGE-2=7.73 ROUGE-L=28.40 BLEU =12.62
Turky et al 2021 [25]	1-Data cleaning 2-Tokenization 3- Padding	LSTM glove model	COVID-19 Open Research Dataset	Rouge-1= 43.6 Rouge-2=36.7 Rouge-L =43.6
Davoodijam et al 2021 [26]	NA	MultiRank algorithm	BMC	Rouge-1 = 0.164 Rouge-2=0.052 Rouge-L= 0.146 BERTScore=0.806

Gidiotis et al [4] used a basic seq-to-seq RNN model and PEGASUS model with PubMed and arXiv datasets and achieved good results. They combined RUM units and LSTM inside the seq-to-seq model and demonstrated the benefits of those combinations. They showed that putting RUM units into the model's decoder makes the train more stable. This method's flaw is that the authors ignored the created model's complexity and instead concentrated on its effectiveness. Gupta et al [5] generated a dataset called SUMPUBMED. The authors compared the SUMPUBMED dataset with CNN-Daily Mail (CNN-DM) and DUC 2001 (DUC) datasets by applying the seq-2-seq model. The SUMPUBMED dataset achieved a high score compared to others. Afzal et al [8] proved that the modern models of deep neural networks get high accuracy in automatic text summarization against traditional approaches. The weakness of the generated model is the output summary consists of sentences from the original document without any additional processing to extract statistical data to make the article easier to understand. Moradi [16] proposed an automatic summarization system by using a clustering algorithm with an itemset mining algorithm that achieves the highest results. The results demonstrated that the topic-based sentence clustering approach boosted the useful content of the summary while decreasing the redundant details. This method has a weakness in that it could be unable to capture the overall structure of a document,

which can generate an unsuitable summary. Kieuvongngam et al [17] used pre-trained BERT and OpenAI GPT-2 models. They evaluated the result by using the stochastic sampling method. The weakness of the study is the limitation of computation power which is using the GPU instead of the DistilGPT2 version. Milošević et al [18] proved that using the PubMedBERT and BERT-based models improves the performance rather than the DistilBERT and T5-based models. Sarker et al [19] generated a simple and fast medical summarization system using a word-phrase embedding model that acquired a 94.3 percent accuracy when compared with another system which is QSpec where the accuracy is 96.8 %. The system is faster than QSpec since QSpec involves sentence categorization, query, and the creation of semantic types in the Unified-Medical Language System (UMLS). Cai et al [20] combined a Graph Attention Networks-based and SciBERT-based encoder to improve an abstractive summary for scientific papers. The weakness of this research is limited resources for training and evaluation which is using GeForce GTX 1080 Ti GPU instead of using newer graphics cards, like the RTX 3060.

Xie et al [21] proposed a new framework depending on PLMs called KeBioSum.Moravvej et al [22] provide a new model for medical summarization depending on the conditional generative adversarial network that outperformed previous models. Moradi et al [23] utilized a clustering method and a deep bidirectional language model to improve medical text summarization results without the requirement for computationally intensive domain-specific pretraining or knowledge bases. The main limitation of this approach is the rareness of available datasets with its gold summarization. Song and Yongbin [24] applied an innovative approach from the seq-2-seq model and achieve high results against the traditional seq-2-seq model. Turky et al [25] developed a model based on the glove model to transform input into vectors, which were passed across LSTM to generate the summary. The model was only allowed to provide single-sentence titles, instead of producing multi-sentence summaries. Davoodijam et al [26] proposed that utilizing the MultiRank algorithm with the features such as word, semantic, and co-reference similarity improves the summary results. The weakness of the following approaches [4,18,19,21,22,24,26] is that they did not utilize modern pre-processing methods like POS tagging to improve the outcome.

4. Conclusion and future work

This survey gives a general overview of text summarization models and the newest work in the field of medical summarization from 2018 to 2022. It includes 15 scientific publications and demonstrates how medical summarization allows medical researchers to capture more useful information in less time. The paper describes evaluation metrics and the newest datasets applied in text summarization. The future work is to propose an efficient and accurate model for medical summarization that overcomes the issues of the current models and applies it to different datasets.

References

1. Wang, Hua, Cuiqin Ma, and Lijuan Zhou. "A brief review of machine learning and its application." In 2009 international conference on information engineering and computer science, (2009) ,pp. 1-4. IEEE.
2. Yadav, Divakar, Jalpa Desai, and Arun Kumar Yadav. "Automatic Text Summarization Methods: A

Comprehensive Review." arXiv preprint arXiv:2204.01849 (2022).

3. Gaikwad, Deepali K., and C. Namrata Mahender. "A review paper on text summarization." International Journal of Advanced Research in Computer and Communication Engineering 5, no. 3 (2016), pp. 154-160.
4. Gidiotis, Alexios, and Grigorios Tsoumakas. "A divide-and-conquer approach to the summarization of long documents." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 3029-3040.
5. Gupta, Vivek, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. "SUMPUBMED: Summarization Dataset of PubMed Scientific Articles." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop,(2021), pp. 292-303.
6. Mollá, Diego, and Maria Elena Santiago-Martinez: "Development of a corpus for evidence-based medicine summarization". In: Proceedings of the Australasian Language Technology Association Workshop, (2011), pp. 86–94.
7. Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, et al. "Cord-19: The covid-19 open research dataset." (2020) ArXiv .
8. Afzal, Muhammad, Fakhare Alam, Khalid Mahmood Malik, and Ghaus M. Malik. "Clinical context-aware biomedical text summarization using deep neural network: model development and validation." Journal of medical Internet research 22, no. 10 (2020): e19810.
9. Gupta, Supriya, Aakanksha Sharaff, and Naresh Kumar Nagwani. "Biomedical Text Summarization Based on the Itemset Mining Approach." In New Opportunities for Sentiment Analysis and Information Processing, pp. 140-152. IGI Global, (2021).
10. Luo, Ling, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. "BioRED: a rich biomedical relation extraction dataset." Briefings in Bioinformatics 23, no. 5 (2022): bbac282.
11. Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. "S2ORC: The semantic scholar open research corpus." arXiv preprint arXiv:1911.02782 (2019).
12. Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In-Text summarization branches out,(2004), pp. 74-81.
13. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics,(2002), pp. 311-318.
14. Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).
15. Afsharizadeh Mahsa, Ebrahimpour-Komleh H, Bagheri A. Automatic Text Summarization of COVID-19 Research Articles Using Recurrent Neural Networks and Coreference Resolution. Frontiers Biomed Technol. 2020;7(4):236-248.

16. Moradi, Milad. "CIBS: A biomedical text summarizer using topic-based sentence clustering." Journal of biomedical informatics 88 (2018), pp. 53-61.

17. Kieuvongngam, Virapat, Bowen Tan, and Yiming Niu. "Automatic text summarization of covid-19 medical research articles using BERT and GPT-2." (2020),arXiv preprint arXiv:2006.01997 .

18. Milošević, Nikola, and Wolfgang Thielemann. "Comparison of biomedical relationship extraction methods and models for knowledge graph creation." Journal of Web Semantics 75 (2023): 100756.

19. Sarker, Abeed, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Aamir Abbas. "A light-weight text summarization system for fast access to medical evidence." Frontiers in digital health (2020): 45.

20. Cai, Xiaoyan, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. "COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers." Journal of Biomedical Informatics 127 (2022): 103999.

21. Xie, Qianqian, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. "Pre-trained language models with domain knowledge for biomedical extractive summarization." Knowledge-Based Systems 252 (2022): 109460.

22. Moravvej, Seyed Vahid, Abdolreza Mirzaei, and Mehran Safayani. "Biomedical text summarization using Conditional Generative Adversarial Network (CGAN)." (2021),arXiv preprint arXiv:2110.11870 .

23. Moradi, Milad, Georg Dorffner, and Matthias Samwald. "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization." Computer methods and programs in biomedicine 184 (2020), pp. 105-117.

24. Song, Guohui, and Yongbin Wang. "A hybrid model for medical paper summarization based on COVID-19 open research dataset." In 2020 4th International Conference on Computer Science and Artificial Intelligence,(2020), pp. 52-56.

25. Turky, Saja Naeem, Ahmed Sabah Ahmed Al-Jumaili, and Rajaa K. Hasoun. "Abstractive Text Summary of COVID-19 Documents based on LSTM Method and Word Embedding." Webology 18, no. 2 (2021).

26. Davoodijam, Ensieh, Nasser Ghadiri, Maryam Lotfi Shahreza, and Fabio Rinaldi. "MultiGBS: a multi-layer graph approach to biomedical summarization." Journal of Biomedical Informatics 116 (2021): 103706.