Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

Auto-Labeling Of Medical Arabic Texts

[1]Salma Mahmood Molhi, [2]Akram Al-Subari

[1,2] Ibb University, Ibb, Yemen

Email:[1] salma.molhi@ibbuniv.edu.ye, [2] akram.alsubari@ibbuniv.edu.ye

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Arabic title generation, Natural Language Processing (NLP), Transformer-based models, AI evaluation metrics, Dataset development and human evaluation.

Abstract

Title generation is a specialized area of natural language processing (NLP), and most studies in this area focus on English, while other languages, such as Arabic, are still in their infancy. Recent research has focused on developing title generation systems for Arabic texts, which can be challenging given the nature of the Arabic language and of datasets. Using artificial intelligence (AI), titles were created and information from lengthy texts was made easier to obtain. The titles and content of each title from verified websites were compiled into a dataset. The title-generating work was started using the Arabic T5 transformer-based title generation model. Following that, the final model was assessed using assessment techniques such as BERTScore, ROUGE, BLEU, and F-Score. On the BERTScore scale, the suggested model performed well, earning the highest evaluation score of 87, lso the model was evaluated by four independent, expert evaluators, with the arithmetic mean of the evaluations being 7.25/10, indicating good acceptance of the results. In order to improve the model's performance, the researchers propose applying human evaluation, changing the model's structure, and growing the dataset.

Introduction

The use of big data in healthcare has grown in recent years as a means of enhancing patient outcomes, operational effectiveness, and clinical decision-making. Electronic Health Records (EHRs) are a major source of big data in healthcare since they hold a wide range of patient data, such as medication records, treatment plans, diagnostic results, and medical histories [1]. Manually processing large amounts of text data has become increasingly challenging due to the sheer volume of information available. Automatic text summarization technology has emerged as a solution to address this problem [2]. The workflow of healthcare providers now includes interacting with computer-based systems and retrieving textual data. Six BTS is a biological application of Automatic Text Summarization (ATS) that uses computing to provide concise and pertinent representations of the source materials. It facilitates medical decision-making and improves healthcare quality by assisting healthcare providers in concentrating on the most important facts. The effectiveness of reading an automatically generated summary rather than the raw records was demonstrated by usability studies done with physicians for EHR summarization [3]. The domain of Natural Language Processing (NLP) has, transformed from a niche academic field to a cornerstone of modern technological applications. Central to this transformation has been the development and refinement of sequence-to-sequence (Seq2Seq) tasks. These tasks, pivotal in applications ranging from machine translation to automated chatbots, involve the conversion of one sequence, often a sentence or paragraph in one language, into

another sequence, possibly its translation in another language or a summarized version of the original [4]. Before the advent of transformers, natural language processing (NLP) relied heavily on recurrent neural networks (RNNs), such as long short-term memory (LSTM), which were used to manage sequential data. These models, while effective for short sequences [5], faced challenges such as vanishing gradients and poor parallelism, which limited their ability to accommodate long-term variables. Some models, such as Word2Vec and GloVe, enhanced semantic understanding, but lacked dynamic context. The attention mechanism introduced began to address these issues by allowing models to focus on relevant input segments. However, transformers took full advantage of attention by eliminating redundancy, leading to models like BERT that enabled bidirectional context processing [6]. Later models, such as AraBert, GPT, and T5, improved transformer design and scaled with larger datasets and computational resources, facilitating their emergence in natural language processing [7]. The main problem is the length of texts and articles, the difficulty of summarizing them efficiently and in a short time, and the difficulty of obtaining high-quality summaries that are easy for many to understand [8]. Therefore, this paper presents the Arabic T5 technique for summarizing texts and articles, specifically medical ones. Arabic T5 performs text extraction tasks using a transformer architecture that utilizes pre-training and fine-tuning. The Research Objectives are building an automated summarization model for medical articles, to evaluate how well the Arabic T5 model summarizes and effectively summarizes Arabic medical articles, to provide an Arabic language framework for AI applications in the medical domain.

Related Works

Title generation is the process of summarizing long articles into short, understandable texts. This process aims to produce a shortened version of a longer text while retaining and main meanings of the information. Thus, users can quickly obtain important information without having to read the entire document [9]. Training a machine learning model to perform natural language processing (NLP) tasks often requires that the model can process text in a way that is amenable to downstream learning. This can be loosely viewed as developing general-purpose knowledge that allows the model to "understand" text [10]. The framework that converts all text based language problems into a text-to-text format presented through the T5 model is attractive. In addition to its simplicity, this approach is also effective since it allows knowledge from high-resource tasks to transfer to low-resource tasks without the need for changes in model architecture. Unlike models such as BERT [11]. Which are based on encoders only, the T5 model is also an encoder-decoder that can be naturally used not only for natural language understanding tasks, but language generation as well [12]. Techniques are extractive and abstractive approaches. To communicate the main ideas of the original text, abstractive summarization builds new terms [13]. Abstractive text summarization has attracted much attention since it is capable of generating novel words using language generation models conditioned on representation of source documents [8]. Extractive summarization has an important position in Natural Language Processing (NLP). Its primary function is to condense lengthy texts into concise summaries. The importance of sentences depends on the linguistic and statistical characteristics of sentences [14]. Previous work has been done on generating titles for medical reports, much of which relies on the use of pre-trained models. In Helwan, Azar, & Ozsahin (2023), the T5-small model was used to summarize medical reports. The researchers used 3,800 samples from the Indiana database, evaluated using the ROUGE criterion (ROUGE-1 = 84.24, ROUGE-2 = 59.58, ROUGE-3 = 75.85, and ROUGE-4 = 78.78). The results showed that long reports were summarized well. The limitations were the small data size, the limited length of the input (512) and output (150), and the use of only the Findings and Impressions fields [15]. In the study by Nishio, M., Matsunaga, (2023), four T5 models were built and the results were compared using the Wilcoxon test on MIMIC-CXR and JMID data. The models showed the following results (ROUGE-1 = 57.75 ± 30.99 in MIMIC-CXR and 50.00 ± 29.24 in JMID), with limitations of relying on only two data sets and only two languages [16]. Rehman, T., and Sanyal (2024), pre-trained models, such as T5-base, BART-base, and PEGASUS-large, were compared to generate research titles on CSPubSum and LREC-COLING 2024ـ datasets. ChatGPT 3.5 was tested in zero-shot mode. PEGASUS-large performed best, with ROUGE-1 = 49.85 and ROUGE-2 = 30.51 on LREC-COLING-2024, and achieved high performance on CSPubSum (ROUGE-1 = 98.13 and ROUGE-2 = 98.08). Limitations include the evaluation being limited to two datasets, relying heavily on automated metrics such as ROUGE and BLEU, and very little human evaluation [17].

Methodology

The steps that followed to evaluate summarization using Arabic T5 model is shown in **Figure 1**.

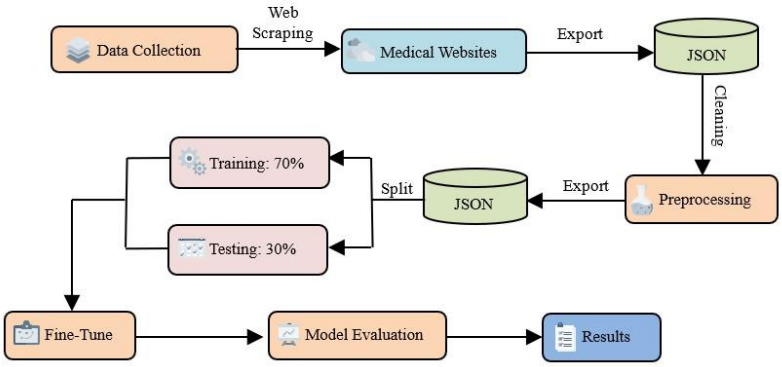

Figure 1. The Steps Of Methodology.

A. Data Collection

Using web scraping technique from multiple websites, a dataset was created. This technique uses Python libraries, including BeautifulSoup[1] and Request[2]. Scraping was used for several websites such as WebTeb, Altibbi, Daily Medical Info, Sehatok and MedArabic, which were used so that many medical articles were extracted into a JSON file, each article containing a title and content. where the total number of saved articles 233,188 covering various medical fields such as health, skin, children, nutrition, hair, heart diseases, and other fields.

B. Data Preprocessing

The dataset underwent several preprocessing steps, including removing duplicate and empty articles containing only titles, clearing the article content of dates and links, and removing non-Arabic characters and symbols. Using several libraries, including Pandas[3] and Remover[4], the data size was reduced to 45,938 JSON files.

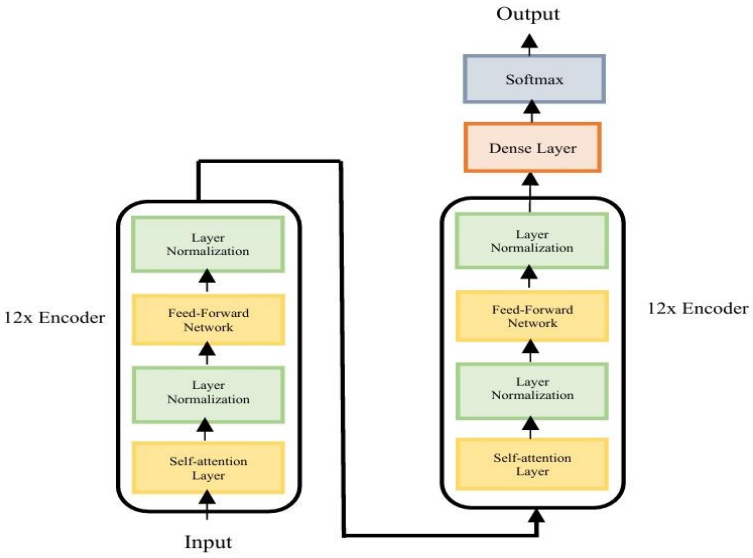

Figure 2. T5 Architecture

C. Training Data and Testing Data

Data was collected by scraping, as previously explained, and then pre-processed to match the selected model. The data was divided into 70% for training and 30% for testing. A sample of 15,000 data was taken for the model training process, with 10,500 articles for training and 4,500 articles for testing.

D. Arabic T5 Model

The encoder-decoder transformer concept known as the Text to-Text Transfer Transformer (T5). The T5 model's primary concept is to create a single framework for many NLP tasks by utilizing transfer learning. The architecture of the T5 is comparable to that of the original transformer [23]. T5 works on a simple but powerful idea: it turns all NLP difficulties into text-to-text challenges. Similar to Transformer-based sequence-to-sequence models, the model has an encoder decoder architecture. It functions by: Text-to-Text Task Formulation: It reformulates each NLP task into a text-based input and output rather than handling them independently. T5 Architecture is shown in **Figure 2**.

The input is encoded: Sentence Piece is used to tokenize the input text before it is transmitted through the encoder, which creates a contextual representation. Decoding the Output: The output text is produced autoregressively by the decoder using the encoded representation. Model Training: T5 is pre trained with a denoising objective, which teaches the model to recreate masked text passages. After that, it is adjusted for different jobs [24].

E. Arabic T5 Model Finetuned

The Arabic T5 model was used to perform the title generation task and was trained on the dataset according to the following: The title generation model was downloaded from the Hugging Face website, and parameters were changed to match the model in terms of text length, number of batches, etc. The model was trained on the collected dataset, teaching the model to generate titles from existing content. The model has 12 layers in the encoder and 12 layers in the decoder. This makes a total of 24 layers. The vocabulary size used in the model is 110,0805.

Results And Discussion

Several parameters were used in the training phase, with a learning rate of 5e-5 to determine the number of parameter changes per update step. The number of training epochs was 3, and the batch size was 4. After the model was trained, it was saved and evaluated by adding content to it and then generating a title for it. This was done in the following steps: Content from articles outside the training material was used, and appropriate titles were assigned to each article entered into the model. Several metrics were used to evaluate the model's performance, and the results shown in **Table 1** were obtained. It was found that the model understood the context of the content and generated an appropriate title for it.

Table 1. *Samples from testing the model that was trained.*

Content	Title
.أخذ قسط من الراحة والنوم العميق من أجل الصحة عموماً ومن أجل صحة عينيك على وجه الخصوص تجنب أشعة الشمس، ودرجات الحرارة .تجنب مسببات الحساسية خصوصاً حبوب اللقاح والأتربة والغبار وضع كمادات المياه الباردة على .العالية، والأماكن التي تساعد على ظهور الحساسية، والملوثات الهوائية .العين لعدة مرات يومياً	كيف يمكن تنظيف العين من الحساسية؟
عدوى الجهاز التنفسي السفلي التي تؤثّر على الرئة أو المسالك الهوائية السفلية، مثل: (التهاب الشّعب الهوائية نوبة من حالة طويلة .الحاد أو الالتهاب الرئوي) أو الحساسية، مثل: (التهاب الأنف التحسُّسي أو حُمى القش) استنشاق الغبار أو .الأمد، مثل: (الربو، مرض الانسداد الرئوي المزمن، أو التهاب الشعب الهوائية المُزمن) .الدخان	عدوى الجهاز التنفسي السفلي
تنتقل الملاريا عن طريق البعوض، ويتسلل هذا الطفيلي داخل كريات الدم الحمراء في جسم الإنسان فيدمرها. تراوح فترة الحضانة للمرض بين 7- 30 يومًا. تنتقل الملاريا بين البشر من خلال لدغات أجناس بعوضة الانوفيليس الحاملة لها.	الملاريا
يتساقط الشعر في دورة متكررة، و من الطبيعي أن نفقد ما بين 50 و100 شعرة يوميًّا غالبًا دون أن نلاحظ ذلك، ولكن في بعض الأحيان يمكن أن يكون علامة على وجود حالة طبية.	تساقط الشعر
تُعتبر الأظافر أحد أهم مؤشرات الصحة العامة والتي يتم تحديدها من خلال لونها ومدى نعومتها، حيثُ تتكون الأظافر من طبقات متعددة من الكيراتين، وهو البروتين البنائي للأظافر و الذي يعطيها القوة ، لذلك يجب الحفاظ على نظافتها دوماً	العناية بالأظافر

As shown in **Table 2** several metrics were used to calculate the model's performance efficiency, including ROUGE, BLEU BERTScore and F0.5 Score. Where each of which has sub-metrics, ROUGE stands for "Recall-Oriented Understudy of Gisting Evaluation," and refers to a collection of criteria for assessing automatically generated texts. ROUGE measures can be calculated in a variety of methods, based on the different granularity. The following are the most commonly used: 1. ROUGE-N refers to the overlapping of N-grams (unigram, bigram , trigram and so on) between the system summary and the reference summary; 2. ROUGE-L

measures the longest common word sequence, computed by the Longest Common Subsequence (LCS) algorithm; 3. ROUGE-S refers to a couple of words in an ordered sentence, that allows some gaps. Sometimes this measure is also called skip-gram; 4. ROUGE-SU is a weighted mean between ROUGE-S and ROUGE-L [25]. ROUGE-1, ROUGE-2, and ROUGE-L are the most commonly used metrics, since they reflect the granularity of the studied texts.BLEU Scale: BLEU (Bilingual Evaluation Study) is a standard for measuring the quality of a translation or text created objective of BLEU is to compare n-grams of the candidate translation with n-grams of the reference translation and count the number of matches; the more the matches, the better the candidate translation, BLEU scale gives a score between 0 and 1, where high values indicate high quality of the translation or generated text [26]. Neural networks trained using BERT (Bidirectional Encoder Representations of Transformers) provide the basis for the BERTScore measure. Depictions. To gauge how similar sentences and texts are to one another, BERTScore use BERT representations. ERTScore computes the similarity between reference and created sentences after converting sentences into BERT representations. High scores on the BERTScore scale, which ranges from 0 to 1, denote high sentence similarity [27]. The model proposed in this study was compared with the Deep Seek model. Both models were evaluated by four expert evaluators. The proposed model achieved an average score of 7.25 out of 10, while Deep Seek received 7.75. These results indicate that the performance of the two models is very close.

Table 2. *Performance of the model.*

ROUGE	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE- Lsum
	0.012	0.010	0.012	0.012
BLEU	BLEU		Precisions	
	0.27403		0.5260	
BERTScore	Precisions		Recall	
	0.8728		0.8486	
F Score	F1 Score		F0.5 Score	
	0.8577		0.8679	
Training loss	1.8928			

Conclusion

In this research paper, a model was created to generate Arabic titles based on a dataset collected from verified medical websites. The T5 Arabic model was used, and the model was trained to generate titles so that the generator generates text directly from the context of the content. Results demonstrate that the model is capable of generating titles for any input content, achieving a BERTScore of 0.87 and a good F0.5 score of 0.86. However, the results of the other two metrics, ROUGE and BLEU, were not satisfactory, and therefore the model requires further improvement by researchers and those interested in Arabic language generation.

Future Work

- On the ROUGE and BLEU metrics, which showed poor results compared to BERTScore and F0.5.
- A more comprehensive model could be developed by using a dataset with wider Arabic medical.
- In addition to evaluating the model with human factors.
- Other models that support Arabic should also be tested and compared to compare the results.

References

[1] O. Elijah, “Healthcare big data in electronic health records (EHR),” ResearchGate (2025).

[2] C. Setyawan, N. Benarkah, and V. R. Prasetyo, “Automatic text summarization berdasarkan pendekatan statistika pada dokumen berbahasa Indonesia,” Keluwih: Jurnal Sains dan Teknologi 2(1), 9–15 (2021).

[3] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, and J. Mostafa, “A systematic review of automatic text summarization for biomedical literature and EHRs,” J. Am. Med. Inform. Assoc. 28(10), 2287–2297 (2021).

[4] J. Zhu, "Comparative study of sequence-to-sequence models: From RNNs to transformers," Appl. Comput. Eng. 42(67), 2755–2721 (2023).

[5] Nagajayant Nagamani. (2022). Explainability and Robustness Trade-offs: Ensuring Safety and Fairness in Large-Scale AI Deployments. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 484–494..

[6] N. Quach, Q. Wang, Z. Gao, Q. Sun, B. Guan, and L. Floyd, "Reinforcement learning approach for integrating compressed contexts into knowledge graphs," in Proc. 5th Int. Conf. Computer Vision, Image and Deep Learning (CVIDL) (IEEE, 2024), pp. 862–866.

[7] T. Wu, Y. Wang, and N. Quach, "Advancements in natural language processing: Exploring transformer-based architectures for text understanding," in Proc. 5th Int. Conf. Artificial Intelligence and Industrial Technology Applications (AIITA) (IEEE, 2025), pp. 1384–1388.

[8] A. R. Lubis, H. R. Safitri, M. Lubis, M. L. Hamzah, A. K. Al-Khowarizmi, and O. Nugroho, "Enhancing text summarization with a T5 model and Bayesian optimization," Rev. Intell. Artif. 37(5), 1213 (2023).

[9] M. W. B. D. Satya, A. Luthfiarta, and M. N. Althoff, "Comparative analysis of T5 model performance for Indonesian abstractive text summarization," SISTEMASI 14(3), 1092–1106 (2025).

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res. 21, 1–67 (2020).

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT (2019), pp. 4171–4186.

[12] B. Khan, M. Usman, I. Khan, J. Khan, D. Hussain, and Y. H. Gu, "Next-generation text summarization: A T5-LSTM FusionNet hybrid approach for psychological data," IEEE Access (2025).

[13] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural abstractive text summarization with sequence-to-sequence models," ACM Trans. Data Sci. 2(1), 1–37 (2021).

[14] M. Azam, S. Khalid, S. Almutairi, H. A. Khattak, A. Namoun, A. Ali, and H. S. M. Bilal, "Current trends and advances in extractive text summarization: A comprehensive review," IEEE Access (2025).

[15] A. Helwan, D. Azar, and D. U. Ozsahin, "Medical reports summarization using text-to-text transformer," in Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET) (IEEE, 2023), pp. 01–04.

[16] M. Nishio, T. Matsunaga, H. Matsuo, M. Nogami, Y. Kurata, K. Fujimoto, and T. Murakami, "Fully automatic summarization of radiology reports using natural language processing with language models," medRxiv (2023).

[17] T. Rehman, D. K. Sanyal, and S. Chattopadhyay, "Can pre-trained language models generate titles for research papers?" in Asian Digital Libraries (Springer, Singapore, 2024), pp. 154–170.

[18] WebTeb, "WebTeb – Trusted health information," https://www.webteb.com/ (2025).

[19] Altibbi, "Altibbi – Trusted Arabic health platform," https://altibbi.com/ (n.d.).

[20] Daily Medical Info, "Daily Medical Info – Trusted medical information," https://dailymedicalinfo.com/ (2025).

[21] Sehatok, "Sehatok – Credible Arabic health information," https://www.sehatok.com/ (2025).

[22] MedArabic, "MedArabic – Arabic medical sciences and translation community," https://medarabic.com/ (2014–2022).

[23] M. Al-Qaraghuli and O. A. Jaafar, "Arabic soft spelling correction with T5," Jordanian J. Comput. Inf. Technol. 10(1) (2024).

[24] "T5: Text-to-text transfer transformer," https://www.geeksforgeeks.org/nlp/t5-text-to-text-transfer-transformer/.

[25] M. Barbella and G. Tortora, "ROUGE metric evaluation for text summarization techniques," SSRN 4120317 (2022).

[26] H. Saadany and C. Orasan, "BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text," arXiv:2109.14250 (2021).

[27] H. A. Balla, M. Llorens Salvador, and S. J. Delany, "Arabic medical community question answering using ON-LSTM and CNN," in Proc. Int. Conf. Machine Learning and Computing (2022), pp. 298–302.