Label-based Topic Modeling to Enhance Medical Triage for Medical Triage Robots

Jiayi FENG[1], Runtong ZHANG[1]*, Donghua CHEN[2], Lei SHI[3], Chenghao XIAO[4]

[1] Beijing Jiaotong University, 3 Shangyuan Village, Beijing, 100044, China
jyfeng@bjtu.edu.cn, rtzhang@bjtu.edu.cn (**Corresponding author*)

[2] University of International Business and Economics, 10 Huixin East Street, Beijing, 100029, China
dhchen@uibe.edu.cn

[3] Newcastle University, Urban Sciences Building, Newcastle upon Tyne, NE4 5TG, United Kingdom
lei.shi@newcastle.ac.uk

[4] Durham University, Stockton Road, Durham, DH1 3LE, United Kingdom
chenghao.xiao@durham.ac.uk

Abstract: Medical triage robots leverage natural language processing algorithms to provide accurate medical information and triage services, ultimately alleviating the strain on healthcare specialists. However, their effectiveness often hinges on the quality of topic assignment. This study proposes the Knowledge-Constrained Labeled Latent Dirichlet Allocation (KC-LLDA) method, which incorporates domain-specific knowledge constraints with LDA. KC-LLDA was compared with other existing similar topic extraction methods, which demonstrated that the proposed method is more suitable for topic modeling in the context of medical texts. In addition, this paper sets forth a novel hybrid method that combines supervised and unsupervised learning, leveraging the synergies between KC-LLDA and the BERT model, which results in a better learning of contextual information contained in medical texts, leading to the improvement of the classification accuracy. The obtained results highlight the fact that1 utilizing topic assignment can increase the efficiency of medical triage robots, ultimately improving the healthcare services provided to patients.

Keywords: Topic assignment, Medical triage robot, Question answering system, Triage, Domain knowledge.

1. Introduction

According to industry observations, patients coming to the hospital for their first visit or consultation encounter difficulties in identifying the most suitable doctor for their specific medical condition, primarily due to their lack of adequate medical knowledge (Pan et al., 2018). This issue is further exacerbated among underrepresented populations, including racial and ethnic minorities, low-income individuals, and residents of rural areas, who often face significant barriers in accessing appropriate medical knowledge (Tawfik et al., 2023). The triage process is further complicated by the allocation of limited resources, the time-consuming assessment, shortages of experts in specific fields, and the presence of cultural and language barriers (Alqaysi et al., 2022). Healthcare providers are increasingly under pressure to improve the efficiency of healthcare delivery (Lee et al., 2017), especially for customized services that tend to be more resource-intensive than standard healthcare services. One of the most prominent challenges in healthcare service is the glaring disparity between supply and demand, particularly concerning the availability of medical experts (Song et al., 2016). An intelligent robot capable of providing automated responses emerges as an imperative solution, alleviating the burden on the healthcare system and caregivers (Chaudhary et al., 2023; Zhu, 2023). Moreover, such technology can bridge the information gap experienced by vulnerable populations with limited health literacy, thereby promoting healthcare equity.

Using medical triage robots for patient assessment holds the potential to enhance care quality, diminish wait times, and alleviate the workload of triage specialists (Jen et al., 2021). The accurate and efficient response of medical service robots to user queries within the complex hospital environment, rich in data sources, is of paramount importance (Al-Taee et al., 2016). In the clinical context, medical service robots encompass any system or device (Karabegović & Doleček, 2017) capable of performing tasks, either partially or fully autonomously, to deliver a service beneficial to humans in a medical capacity. This includes tasks as diverse as serving a cup of tea to a patient or tailoring treatments to individual variations (Cingolani et al., 2023; Morsi, 2023). In addition, these robots have the capacity to improve psychological well-being and overall satisfaction (Suligoj et al., 2018). Given the frequent interaction between service robots

https://doi.org/10.24846/v32i4y202304

and users, a critical consideration lies in enabling seamless collaboration between humans and robots (Gbouna et al., 2021). In the case of medical triage robots, this involves translating natural human language into commands that the machine can understand, thereby ensuring that one receives accurate responses. One of the most important steps is identifying the correct topic that addresses the user's needs (Khan et al., 2023). Within a question answering system (QAS) based on user input, the implementation of topic assignment plays a crucial role in delivering precise service through a healthcare service robot. Automatic categorization of user-generated questions by topic, along with suggesting similar questions and answers, diminishes redundancy and augments the question answering capability of healthcare service robots. By offering analogous question recommendations within the same category, the search scope is greatly reduced, elevating retrieval efficiency and accuracy. These advancements hinge on the foundation of robust question classification techniques (Xue et al., 2008; Zhang et al., 2014). The aim of question classification is to reduce search space for selecting appropriate candidate answers during the answer processing stage (Wasim et al., 2019). Therefore, integrating topic assignment with established QAS techniques represent an opportunity to refine robot-assisted triage services within a hospital context, enabling the provision of similar recommendations for the same problem under the same category and further enhancing the capabilities of healthcare service robots in QAS.

This paper presents several noteworthy contributions. Firstly, it proposes a novel triage approach designed to facilitate patient triage for intelligent medical customization services. This new mechanism enables the translation of patients' needs expressed in natural language into medical terminology, ultimately reducing the burden on healthcare resources. Secondly, the Knowledge-Constrained Labeled LDA (KC-LLDA) method is proposed, which is grounded in knowledge constraints. Through a comparative analysis including existing similar approaches for topic extraction, it is demonstrated that the incorporation of medical knowledge constraints into the LDA method proves more adept for topic modeling of medical texts. Finally, an extensive dataset of consultation records sourced from the online medical community is leveraged for training a classifier that synergistically combines BERT and KC-LLDA. This classifier was engineered to surmount prevailing challenges by effectively discerning the precise intention of the patient and bridging the semantic gap between medical terminology and individual patient needs. This integrated approach holds promise in significantly enhancing the efficacy and precision of patient triage in intelligent medical customization services.

The remainder of this paper is as follows. Section 2 delves into the existing research works related to the topic of this work. Section 3 sets forth a novel triage method that aims to understand the needs of patients expressed in natural language and construct mappings based on the research methodology employed in this study. Section 4 discusses the performance of the proposed method and Section 5 includes the conclusion of this paper and possible future research directions.

2. Related Works

Extracting precise topics from short text documents such as large-scale medical question-and-answer conversations is a critical and challenging task (Rashid et al., 2019). Many question and answer systems for medical robots still use rule-based matching methods and string-matching algorithms to build domain dictionaries (Veisi & Shandi, 2020) for classifying and querying questions, which only provide fixed medical terms without considering the relevance and contextual semantics among medical terms for the user (Jiang et al., 2021). However, many intents can be elicited from one utterance depending on the context interpretation (Vatian et al., 2019). Linguistic features take on a significant role to develop an accurate question classifier (Yilmaz & Toklu, 2020). In order to correctly understand human needs as a question topic, plenty of topic modeling techniques in the text mining field are explored by researchers. Some of them, such as Rapid Automatic Keyword Extraction (RAKE) and TextRank algorithms, merely depend on the content of the text corpus, which makes them unable to obtain latent semantics in the future and especially makes it difficult for them to handle complicated decision-making in healthcare scenarios. Other methods that employ machine learning techniques, such as Latent Dirichlet Allocation (LDA)-based methods, can extract the semantics from a large text corpus and in many fields, and they have achieved acceptable results. The topic model in (Zhou et al., 2015) leverages LDA to extract potential topic information from problem descriptions, which is then used for problem classification. This model also addresses lexical gaps by conducting a concept-based computational similarity analysis to compare inter-topic similarities (Naderi et al., 2020). However, topic-based methods exhibit an optimal performance with small-scale data volumes. While they excel in improving recall, they may incur a trade-off in terms of precision. Question classification methods based on classical machine learning techniques face limitations in capturing the hidden relationships between features and may struggle to handle complex languages and very large-scale datasets (Dodevski et al., 2021). The main challenge is that the text representation is inherently high-dimensional and highly sparse, with weak feature representation, in addition to requiring manual feature engineering, which is costly.

With the advancement of deep learning algorithms in natural language processing, QASs increasingly rely on deep neural network-based methods for extracting nuanced user phrase features and classifying the underlying user intent (Liu et al., 2020). Deep learning algorithms like CNNs, RNNs, and Transformers excel in transforming the initial “low-level” feature representations into “high-level” feature representations by constructing multilayer artificial neural networks that iteratively extract and filter input information layer by layer (Cui et al., 2020; Ma et al., 2023; Ertuğrul & Abdullah, 2022). While Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and BERT-based approaches stand as the current state-of-the-art techniques in medical natural language processing (NLP) tasks, their effectiveness has been substantiated across a range of medical scenarios (Liao et al., 2020; Tavabi et al., 2023; Yang et al., 2022). However, in clinical text classification tasks, only very few words contribute to a particular label (Gao et al., 2021). Compressed BERT models reduce resource consumption but can potentially limit the extraction of deep information between phrases and sentences, potentially reducing classification accuracy (Yu et al., 2023). Deep learning-based methods, though capable of extracting semantics, still grapple with accurately predicting the true intents of patients and fall short of providing patients with the support of health-related expert knowledge. Therefore, it is believed that topic modeling techniques should incorporate additional properties of deep learning-based methods and expert knowledge sources within medical domains to enable the patients seeking health information in an untrusted online environment to find that information. This work aims to extract meaningful topics and create correlations between topics and expert knowledge, thereby enabling the patients' access to meaningful topics rather than to keyword-based search results.

3. Research Methodology

3.1 Overview of the Proposed Framework

Medical service robots may be used for various purposes. Here only the robot that is designed to facilitate human-robot interaction is considered. It is important for medical service robots to provide intelligent QA services (Wang et al., 2023). A medical QAS (MQAS) in a medical service robot provides user-friendly interfaces for individuals who aim to obtain robust answers for decision support (AFDS) in accordance with user questions and accordingly receive robot-assisted services. A AFDS refers to the answers provided by a QAS that can help users provide medical decision support based on existing medical data. The MQAS in a medical triage robot includes two models, namely question (*Q*) as input and *AFDS* as output. The narratives of *Q* are identified by a voice recognition module of the robot. The *AFDS* extends a simple answer that only includes text by introducing additional properties regarding the hospital context. When a *Q* is input, the system outputs the optimal *AFDS*; a mapping process between the *Q* and *AFDS* on the basis of domain knowledge (*DK*) in the medical field is established, which is defined as

$$Q \xrightarrow{\{DK\}} \{AFDS\} \tag{1}$$

DK in (1) is considered to provide meaningful answers for the robot so that is can implement them through a properly designed *MQAS*. For instance, if it is not possible to build customized decision-making models for each person, the proposed method can overcome this challenge and provide automated modeling of decision-making models related to *AFDS* using *DK*. *Q*, *AFDS*, and mapping process *Q* → *AFDS* indicate a basic process of question answering in the robot.

In the context of a medical triage robot, the task of predicting user intentions hinges on the analysis of the semantic meaning embedded within the user's query *Q*. The goal is to perform topic identification on the questions posed by patients to achieve the mapping *Q* → *AFDS*. The actual meaning of patient problems is highly context-based, making an approach based solely on word co-occurrence, like Latent Dirichlet Allocation (LDA), potentially inadequate. Considering these limitations, the aim is to leverage the context-dependent word representation BERT to perform word embedding on the input text and extract its semantic features. Subsequently, the topic features of the input text are enriched by utilizing trained LDA-based models and domain knowledge. The BERT vector for *Q* is then integrated with the LDA topic distribution to create a more comprehensive semantic representation. This concatenated representation serves as the input to the variational autoencoders (VAE) model for topic allocation, which is capable of learning medical triage tag representations in the latent space and mapping the input *Q* to the latent space. Combining topic models with BERT can significantly improve the performance of semantic similarity prediction, particularly when it involves domain-specific words (Peinelt et al., 2020). The process of constructing *AFDS* based on Algorithm BERT-KC-LLDA is exhibited in Figure 1.

3.2 Knowledge-constrained Topic Modeling

The purpose of this article is to help patients render their described disease symptoms from natural language into specialist medical terminology through a medical triage robot. To achieve this, employing NLP techniques becomes crucial for understanding the text that conveys the patient's needs. While existing NLP research typically requires large amounts of relevant data, the reality is that there is no large-scale publicly available

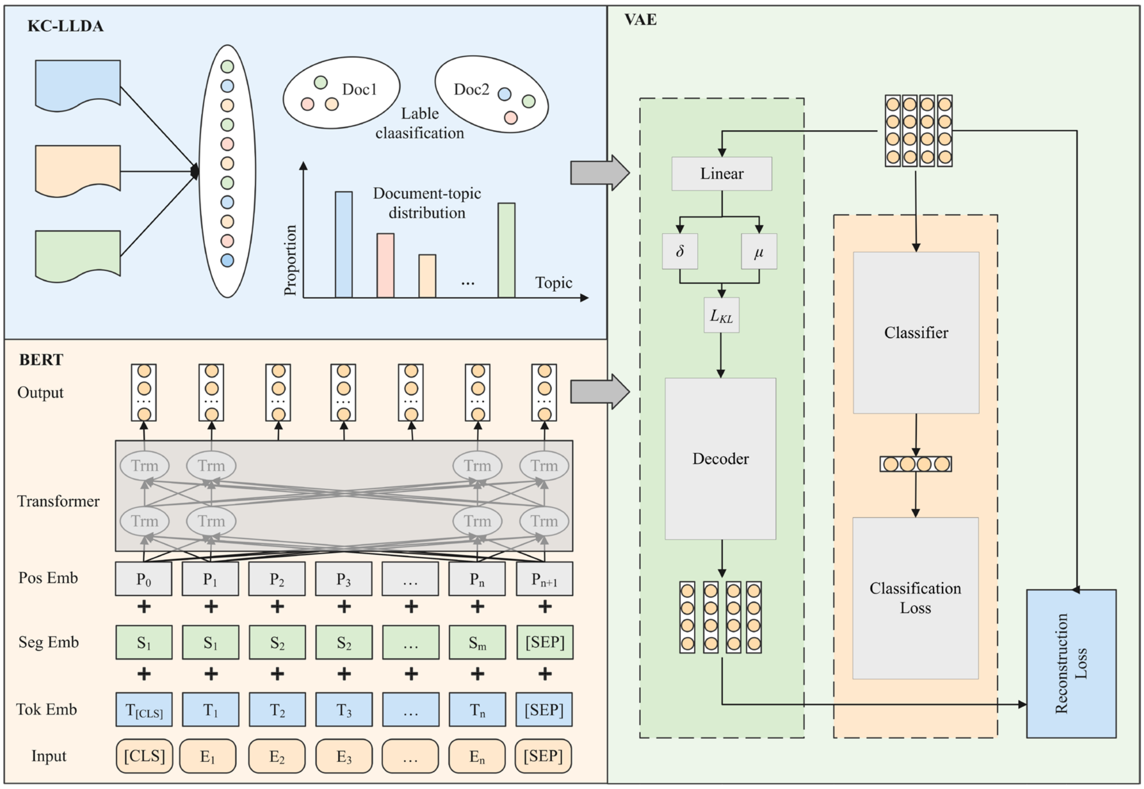


Figure 1. The overall flow chart of BERT-KC-LLDA

medical triage data. It has been recognized that the needs expressed by patients in free-form text are very similar to the questions asked by users in the online health community (OHC).

In fact, a large number of medical question-and-answer datasets exist in OHC containing a lot of labelled information. This includes instances where doctors label their responses based on their own understanding. Comparatively, LDA has excellent implicit semantic mining and data dimensionality reduction capabilities in relation to previous topic models like LSA and PLSA. However, LDA is only a data reduction and clustering algorithm, unable to effectively determine these external labels. This gap is addressed through the proposed Labeled LDA (L-LDA), specifically the application of domain knowledge resulting in the creation of Knowledge-Constrained Labeled LDA (KC-LLDA). This approach is employed to train and test appropriate topic models for questions that are based on existing medical question corpora. L-LDA (Ramage et al., 2009) is a supervised topic model for credit attribution in multilabeled corpora. For each document, it discerns the sections that must be attached specific labels, consequently learning accurate models of the words optimally associated with each label globally. L-LDA may outperform traditional LDA when analyzing questions having multiple labels, tags, or background information. However, developing and evaluating QASs remains a complex task (Marx et al., 2014). The L-LDA is defined as a probabilistic graphical model that represents a process for generating a labeled document collection. Similarly with LDA, L-LDA models each document as a mixture of underlying topics and generates each word based on these topics. The proposed model incorporates supervision by simply constraining the topic model to use only those topics corresponding to a document's labels or tags. The modeling process is outlined as follows.

It is assumed that a document in the training dataset in an OHC is made of a list of words $w^d = \left(w_1, \ldots, w_{N_d}\right)$ and a list of topics $w^d = \left(w_1, \ldots, w_{N_d}\right)$, where each $w_i \in V$ and each $l_k \in \{0,1\}$. Here V is the vocabulary size, K is total number of unique labels in the dataset and N^d is the number of words in a document. In contrast with LDA, L-LDA aims to restrict θ^d to be defined only over the topics that correspond to its label set t^d. Therefore, the word-topic assignments are drawn from this distribution, which ensures that all the topic assignments are limited to a document's labels. The training process uses Gibbs sampling, where the sampling probability for a topic for position i in a document d in L-LDA is given by:

$$P\left(z_i = j \mid z_{-i}\right) \propto \frac{n^{w_i}_{-i,j} + \eta_{w_i}}{n^{(\cdot)}_{-i,j} + \eta^T \mathrm{I}} \times \frac{n^{(d)}_{-i,j} + \alpha_j}{n^{(d)}_{-i,\cdot} + \alpha^T \mathrm{I}} \qquad (2)$$

where $n^{w_i}_{-i,j}$ is the count of word w_i in topic j, that does not include the current assignment z_i, and a missing subscript or superscript indicates a summation over that dimension, where l is a vector of l's appropriate dimension.

In summary, the question is transformed into a relational network of terms that are expanded later on the basis of the knowledge source in the medical field. The topic distribution θ^d provided by trained KC-LLDA models in (2) provides a prediction of the range of knowledge in relation to Q.

3.3 Combining KC-LLDA with BERT

Patients' medical problems can be typically rendered as a less standardized free text that must be mapped into continuous, dense word vectors in a computer-processable format. Traditional word vector models such as one-hot and TF-IDF feature high dimensionality and sparsity, and are also computationally expensive and unable to characterize the text well. Neural network-based models like Word2Vec provide an improvement, yet they cannot encode word meaning within a given context or distinguish between words with multiple meanings. As such, a word vector was needed that could represent words differently in different contexts. BERT is a fully bidirectional language model that proposes a new task called the Masked Language Model (MLM) for training a truly bidirectional encoding model for a supervised task. MLM is a technique for training a two-way language model by replacing a number of words with mask or another random word with a low probability. This is followed by fine-tuning through an additional output layer that can pre-train a deep bidirectional representation by jointly modulating the context across all layers.

The patient's questions were first encoded into a suitable form for computer processing - a sequence of characters separated by spaces $\{E_1, E_2, ..., E_n\}$, which were numerically mapped using BERT's own vocabulary, with E_n denoting the *n*-th character in the vocabulary. The special character CLS indicates the start of a sentence, and the special symbol SEP is inserted between different sentences to differentiate them. The input sequence is then converted into a list of low-dimensional dense vectors, token sequence T_n, segment sequence S_m, and position sequence P_n. Token Embedding T_n is used to convert the individual words of a sequence into a 768-dimensional vector, which represents the information related to that character. Segment Embedding S_m is used to convert the sequence type into a vector, which represents the information that differs from one sequence type to another one. Position Embedding P_n is used to convert the position of the sequence into a vector, which represents the position information for the input sequence. Thus, the embedding representation of topics of Q is defined as:

$$BERT\ Emb(S) = Tok\ Emb\ (T_n) + Seg\ Emb\ (S_m) + Pos\ Emb\ (P_n) \tag{3}$$

Then, in order to obtain the deep semantic properties of the input sequence, a 12-layer transformer network with a BERT encoding layer is trained so that the semantic representation of the text output by the proposed model can characterize the semantic features of *Q* and fine-tune it to obtain the semantic vector C_i corresponding to the *i*-th character. The token vector C_{CLS}, which represents the sequence of characters in *Q*, is then combined with the topic distribution θ_d obtained using the topic model KC-LLDA. All topic vectors are normalized to create the optimized topic vector $T(C_{CLS} + \theta^d)$. During the training process, the hyperparameters of BERT-KC-LLDA were optimized, including the learning rates, batch size, and training epochs.

3.4 Answer for Decision Support in Medical Triage Robots

To integrate the topic probabilities obtained from BERT and KC-LLDA, a variational autoencoder was utilized. The use of VAE is proposed to learn a low-dimensional representation of the topic probability distribution. This will allow one to perform more efficient computations and improve the accuracy of the triage process. VAE is a generative model that consists of an encoder network and a decoder network. The encoder network maps the input topic probabilities into a low-dimensional latent variable space, and the decoder network maps the latent variables back into the original space. VAE is trained to minimize a loss function that consists of a reconstruction loss and a *KL* divergence loss.

The reconstruction loss measures the difference between the input topic probabilities and the reconstructed probabilities generated by the decoder network. It is defined as follows:

$$L_{recon} = -\textstyle\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{4}$$

where *N* is the number of topics, y_i is the true probability of topic *i*, and $\hat{y}_i$ is the reconstructed probability of topic *i*.

KL divergence loss measures the difference between the learned latent variable distribution and a prior distribution. It is defined as follows:

$$L_{KL} = -\frac{1}{2}\textstyle\sum_{j=1}^{J}\left(1 + \log\delta_j^2 - \mu_j^2 - \delta_j^2\right) \tag{5}$$

where *J* is the dimension of the latent variable space, and μ_j and δ_j are the mean and standard deviation of the learned latent variable distribution, respectively.

The total loss function is defined as follows:

$$L = L_{recon} + \beta L_{KL} \tag{6}$$

where β is a hyperparameter that controls the weight of the *KL* divergence loss.

The topic distribution $\theta^{d'}$ obtained from the KC-LLDA algorithm is transformed into a 1×768-dimensional vector through VAE encoding, which is concatenated with the sentence vector C_{CLS} obtained from BERT. The concatenated vector integrates semantic information from both *Q* and the topic. The dimensions of the feature vectors for *Q* and *T* are kept consistent to ensure equal weighting in classification. A feedforward neural network (FNN) is added to the concatenated matrix to output a probability matrix of *Q* belonging to each topic category.

To summarize, KC-LLDA transforms *Q* into a relational network of terms, which is further expanded based on knowledge sources in the medical field. The topic of *Q*, as predicted by BERT-KC-LLDA, provides an indication of the range of knowledge that is relevant to *Q*. As a result, VAE can be used to accurately represent the user's intention and predict the probable tags associated with *Q*. This allows for the determination of *AFDS*.

4. Results and Discussion

4.1 Experimental Setup

To validate the feasibility of the proposed method, over 325,000 user question posts were collected from three different online health platforms: HealthTap, QuestionDoctors, and WebMD, with approximately 140,000 of these posts including tags. Notably, data from the Question Doctors website was used in the form of expert-validated manually annotated tags, as this website had responses provided by actual physicians for tag assignment. Table 1 displays the extracted raw data obtained from these three online healthcare websites.

To evaluate the performance of the proposed machine learning algorithm in predicting the outcomes for a complex classification problem, a range of well-established metrics was utilized, including precision (P), recall (R), and F1 score (F1). These metrics enabled a comprehensive evaluation of the performance of the proposed machine learning algorithm and made it possible to compare it with other state-of-the-art models in the field.

Table 1. Examples of Questions on Online Medical Websites

Websites	Medical questions	Tags
Health-Tap	Zirconium dental implants. How common is it used now? Is there any advantages or benefits over titanium implants. cons & pros please. Thanks.	(dentistry)
Question Doctors	Is my anti hiv test conclusive or need retest?	(hiv test)
WebMD	My son has add and mild autism. He has been successfully on concerta for 6+ years. Can you help with his weight loss.	(autism, weight loss)

4.2 Comparison with Other Topic Modeling Methods

Here, the proposed KC-LLDA method was evaluated in comparison with other existing topic modeling techniques from the literature in terms of obtaining useful information from health-related questions. The experiments were conducted using a health-related question-answering dataset, where the effectiveness of a knowledge-based technique was analysed through P, R, and F1 comparisons between L-LDA and KC-LLDA.

First, the influence of healthcare knowledge on trained LDA-based models was examined. As it is depicted in Figure 2, three LDA-based methods using different text processing techniques were compared: a conventional LDA method (Baseline), an LDA-based method incorporating pos filtering (Part-of-speech), and a LDA-based method integrated with healthcare knowledge (Knowledge). A medical question dataset containing medical questions and their associated tags was used for analysis. The figure shows the change in perplexity over the number of topics set during LDA model training, where a lower perplexity represents a better model. Notably, the baseline method exhibits a higher perplexity in comparison with the other two methods. While both the baseline and the part-of-speech filtering-based methods tends to generate higher perplexity with an increasing number of topics, the knowledge-based method demonstrates a decrease in perplexity with a higher number of topics. These results indicate that the trained knowledge-based LDA models, with lower perplexities, offer greater confidence in fitting the topic distribution for the tested dataset.

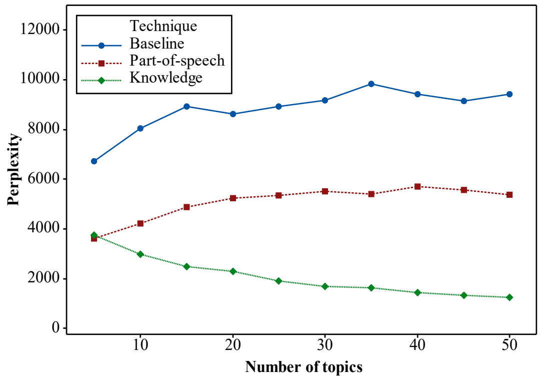


Figure 2. Comparison of change in perplexity among baseline, part-of-speech-based and knowledge-involved methods over the number of topics

Next, a comparison of estimated mean values of P, R, and F1 using ROUGE metrics is presented for the proposed method and for other methods, including Word Frequency (Freq)-based method (Beliga, 2014), TF-IDF-based method (Robertson, 2004), TextRank, RAKE (Rose et al., 2010), LDA, and L-LDA, as summarized in Table 2. The efficiency of ROUGE metrics is well-established and they have demonstrated their ability to make more precise assessments in cases where algorithms may struggle to recognize synonyms. Three types of measures were used: ROUGE-1, ROUGE-2, and ROUGE-L. The estimated values show the differences in mean estimates for P, R, and F1 across different similarity metrics (1-gram and 2-gram). As it can be seen in Table 2 and Figure 3, the proposed KC-LLDA outperforms the other methods. Specifically, when using ROUGE-2, the mean precision, recall, and F1 scores for KC-LLDA are 0.15, 0.35, and 0.19, respectively, an improvement of 4%, 3%, and 4%, respectively over L-LDA. Moreover, as per Table 2, the conventional L-LDA also outperforms LDA in this experiment. Figure 3 further supports these findings, demonstrating that LDA, without considering knowledge related to the labels of the dataset, has the poorest performance among all methods, while KC-LLDA achieves the best performance.

Table 2. Comparison of Estimated Mean Values for P, R and F1

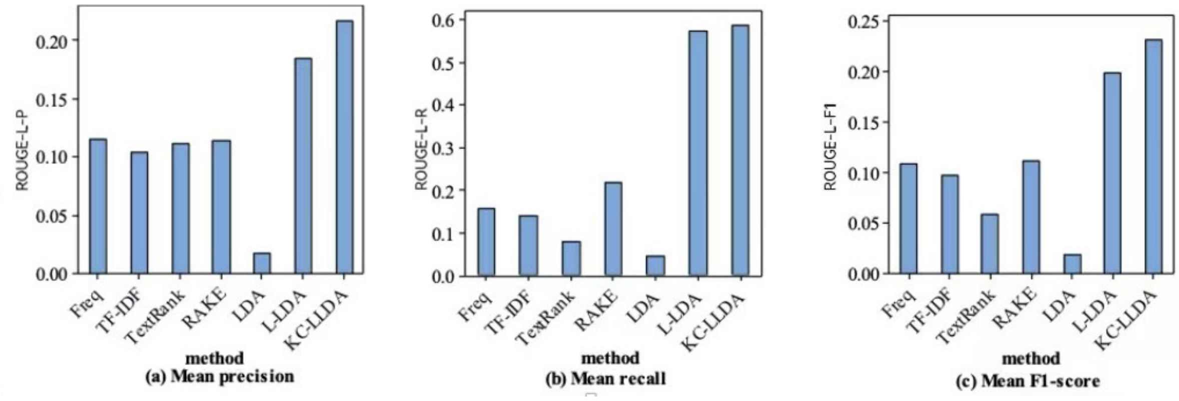
Method	ROUGE-1			ROUGE-2		
	P	R	F1	P	R	F1
Freq	0.12	0.16	0.11	0.01	0.01	0.01
TF-IDF	0.10	0.14	0.10	0.01	0.00	0.00
TextRank	0.11	0.08	0.06	0.01	0.02	0.01
RAKE	0.11	0.22	0.11	0.03	0.04	0.03
LDA	0.02	0.04	0.02	0.00	0.00	0.00
L-LDA	0.18	0.57	0.20	0.11	0.32	0.15
KC-LLDA	0.22	0.59	0.23	0.15	0.35	0.19

Subsequently, the impact of parameters α and β on KC-LLDA was examined, as it is shown in Figure 4. Figure 4(a) shows that ROUGE metrics feature slow increases with $\alpha = [0, 0.5]$, and then stabilize when $\alpha = [0.5, 1.5]$. By contrast, Figure 4(b) shows a gradual decrease in the values of ROUGE metrics as β increases. These results empirically indicate a change of the precision, recall, and F1-scores over the parameters of KC-LLDA.

Lastly, the performance of training topic models was compared for KC-LLDA and L-LDA. Figure 5 illustrates the change in perplexity over the number of iterations. In Figure 5(a)-(b), it is evident that in both experiments, the perplexity for both methods decreases significantly between approximately 2 and 10 iterations, reaching its lowest perplexity around Iteration=20. Subsequently, the perplexity for both methods gradually increases as the iteration count rises. Additionally, Figure 5(c) compares the perplexity for the two methods with the same input parameter, highlighting that the initial perplexity of KC-LLDA is lower than that of L-LDA. KC-LLDA also reaches its lowest perplexity point more quickly than L-LDA. Thus, KC-LLDA, with knowledge-constrained conditions, outperforms the conventional L-LDA in terms of training iterations.

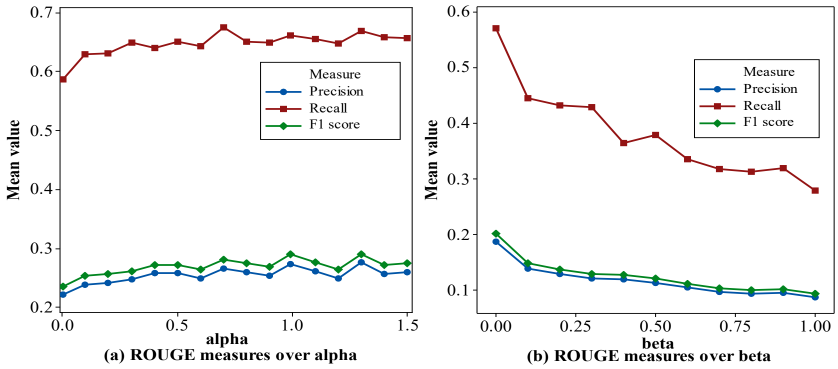
4.3 Comparison with State-of-the-art Methods

A comparative analysis of the proposed method was conducted against several prominent text representation methods, including SVM (El Adlouni et al., 2019), KNN (Srba & Bielikova, 2016), ALBERT (Lan et al., 2019), TBERT (Peinelt et al., 2020), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019). To assess the performance of the analysed classifiers, metrics such as macro precision, recall, and F1-score were employed, which were widely adopted for



(a) Mean precision (b) Mean recall (c) Mean F1-score

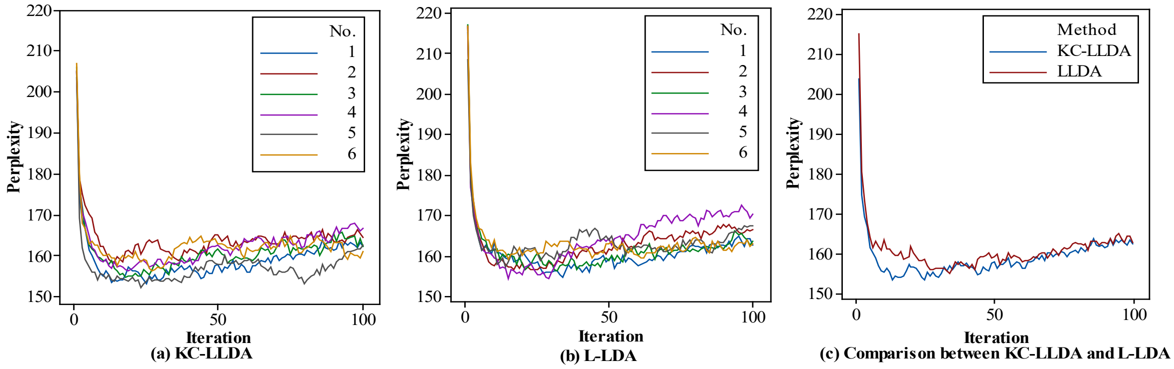
Figure 3. Comparison of ROUGE-L values for the KC-LLDA and other existing methods in the literature



(a) ROUGE measures over alpha

(b) ROUGE measures over beta

Figure 4. Comparison of the values of ROUGE metrics over *α* and *β,* respectively



(a) KC-LLDA

(b) L-LDA

(c) Comparison between KC-LLDA and L-LDA

Figure 5. Comparison of the perplexity values over iterations between KC-LLDA and L-LDA

evaluating classifier quality. Notably, in this study, a greater emphasis was placed on precision, recall, and f-measure over accuracy.

In the training process, the human-transcribed data was randomly split into the train (80%), validation (10%), and test (10%) datasets. To achieve a fair comparison, the hyperparameters of all models were tuned. Specifically, for the proposed model, these were empirically set as βv = 0.01, and α = 50/K. Hyperparameter ρ was set as the mean of all the word embeddings, δ as the number of dimensions in the word embeddings, and ψ as an identity matrix. Given the proposed model's support for latent sub-topics within a given label, Kl = 2 was chosen for labels. The training process was carried out in parallel using PyTorch on an NVIDIA RTX 3090. Table 3 shows a comparison of topic classification performances for the proposed model and the state-of-the-art approaches. As it can be observed from Table 3, the proposed method outperforms other models. These experimental results indicate the superiority of Transformer-based pre-trained language models over traditional machine learning models, particularly in terms of classification accuracy. ALBERT, TBERT, and RoBERTa have demonstrated impressive performance by leveraging pre-training on large-scale corpora and fine-tuning for specific tasks.

Table 3. Performance comparison between the proposed method and the state-of-the-art approaches

Method	HealthTap			WebMD QA		
	P	R	F1	P	R	F1
SVM	0.58	0.49	0.53	0.61	0.52	0.56
KNN	0.52	0.45	0.48	0.55	0.47	0.51
ALBERT	0.77	0.70	0.73	0.72	0.67	0.69
TBERT	0.78	0.65	0.71	0.76	0.69	0.72
RoBERTa	0.82	0.79	0.81	0.81	0.75	0.78
GPT-2	0.74	0.69	0.71	0.75	0.68	0.71
BERT+KC-LLDA	0.87	0.80	0.83	0.87	0.78	0.82

These models possess the ability to adapt to different contextual word embeddings, capture rich contextual information, and exhibit robust generalization capabilities. This suggests their effectiveness in health-related question-answering tasks. However, GPT-2 may not be as well-suited for topic classification tasks, as its primary strength lies in generating coherent text rather than performing classification.

To assess the stability of the proposed model, a robustness analysis was conducted. Specifically, the impact of different LDA methods on the performance of the proposed model was examined. This entailed a comparison among three combinations of the BERT model, namely with LDA, LLDA, and

KC-LLDA. To ensure a fair comparison, the word embedding size was standardized to 768 for all three methods. The results, presented in Table 4, reveal that the proposed BERT-KC-LLDA model attains the highest performance, leveraging external labels derived from user-asked question posts. This also suggests the effectiveness of using external labels obtained through KC-LLDA in enhancing the accuracy of topic classification.

Table 4. Comparison between the proposed method and the BERT-LDA and BERT-LLDA

Method	HealthTap			Webmd QA		
	P	R	F1	P	R	F1
BERT+LDA	0.76	0.67	0.71	0.74	0.61	0.67
BERT+LLDA	0.79	0.77	0.78	0.77	0.72	0.74
BERT+KC-LLDA	0.87	0.80	0.83	0.87	0.78	0.82

5. Conclusion

The objective of this study is to optimize the effectiveness of medical triage robots so that they alleviate the workload of triage specialists. To achieve this, patient triage questions were mapped into topic-labeled classification questions and a domain knowledge-based LLDA approach combined with BERT and VAE was proposed to infer labels for unlabeled topic labels. The proposed method integrates external knowledge sources and existing medical data stored in medical information systems for enhancing QAS of a medical triage robot. The presented BERT-KC-LLDA method extends the limited information extracted from narrative questions and enriches explicit knowledge in the question with implicit knowledge found in DK. This method provides a high-quality inference for further implementation of the delivery of domain-specific answers. Notably, this method bypasses the need for extensive training and testing on a massive corpus of medical questions and answers. In addition, a tag-based dataset was curated from three online healthcare websites to support fine-grained patient problem classification. The obtained experimental results indicate that the proposed method outperforms existing approaches in the literature in providing more useful and meaningful answers to patients.

However, it's important to acknowledge certain limitations related to this analysis. Firstly, the focus was solely on QA pairs from patient-reported posts in the OHC, overlooking potential alternative sources, such as patient profiles and medical images, that could be constructing topic models. Secondly, this knowledge-constrained approach relied solely on disease definitions from ICD-11 to build a corpus for KC-LLDA, thus neglecting other potential sources like UMLS and SNOMED, as well as the intricate semantic relationships between medical concepts within these sources. Future research should delve into a more comprehensive exploration of these knowledge resources to refine the precision and recall of the proposed method. Lastly, it should be noted that the aim of this study is not to supplant the decision-making process of medical experts, but rather to enhance the technical capabilities of OHCs, thereby reducing the pressure on the health services of medical institutions.

Acknowledgements

This work was supported partially by the Fundamental Research Funds for the Central Universities under grant number 2018YJS072, the National Natural Science Foundation of China under grant number 62173025, and a major project of the National Social Science Foundation of China under grant number 18ZDA086.

REFERENCES

Alqaysi, M. E., Albahri, A. S. & Hamid, R. A. (2022) Diagnosis-based hybridization of multimedical tests and sociodemographic characteristics of autism spectrum disorder using artificial intelligence and machine learning techniques: a systematic review. *International Journal of Telemedicine and Applications*. 2022, 3551528. doi: 10.1155/2022/3551 528.

Al-Taee, M. A., Al-Nuaimy, W., Muhsin, Z. J. & Al-Ataby, A. (2016) Robot assistant in management of diabetes in children based on the internet of things. *IEEE Internet of Things Journal*. 4(2), 437-445. doi: 10.1109/JIOT.2016.2623767.

Beliga, S. (2014) Keyword extraction: a review of methods and approaches. To be published at the University of Rijeka. [Preprint] https://www.inf.uniri.hr/images/datoteke/web/2014/beliga_kvalifikacijski_final.pdf [Accessed 16th October 2023].

Chaudhary, A., Mishra, R., Gupta, H. P. & Shukla, K. K. (2023) Jointly prediction of activities, locations, and starting times for isolated elderly people. *IEEE Journal of Biomedical and Health Informatics*. 27(5), 2288-2295. doi: 10.1109/JBHI.2021.3121296.

Cingolani, M., Scendoni, R., Fedeli, P. & Cembrani, F. (2023) Artificial intelligence and digital medicine for integrated home care services in Italy: Opportunities and limits. *Frontiers in Public Health*. 10, 1095001. doi: 10.3389/fpubh.2022.1095001.

Cui, F., Cui, Q. & Song, Y. (2020) A survey on learning-based approaches for modeling and classification of human–machine dialog systems. *IEEE Transactions on Neural Networks and Learning Systems*. 32(4), 1418-1432. doi: 10.1109/TNNLS.2020.2985588.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. To be published in *Proceedings of NAACL-HLT 2019*. [Preprint] https://doi.org/10.48550/arXiv.1810.04805 [Accessed 16th October 2023].

Dodevski, Z., Filiposka, S., Mishev, A. & Trajkovikj, V. (2021) Real time availability and consistency of health-related information across multiple stakeholders: A blockchain based approach. *Computer Science and Information Systems.* 18(3), 927-955. doi: 10.2298/CSIS200426017D.

El Adlouni, Y., Rodríguez, H., Meknassi, M., El Alaoui, S. O., & En-nahnahi, N. (2019) A multi-approach to community question answering. *Expert Systems with Applications.* 137, 432-442. doi: 10.1016/j.eswa.2019.07.024.

Ertuğrul, D. Ç. & Abdullah, S. A. (2022) A Decision-Making Tool for Early Detection of Breast Cancer on Mammographic Images. *Tehnički Vjesnik*. 29(5), 1528-1536. doi: 10.17559/TV-20211221131838.

Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., Wu, X. C., Durbin, E. B., Doherty, J., Stroup, A., Coyle, L. & Tourassi, G. (2021) Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*. 25(9), 3596-3607. doi: 10.1109/JBHI.2021.3062322.

Gbouna, Z. V., Pang, G., Yang, G., Hou, Z., Lv, H., Yu, Z. & Pang, Z. (2021) User-interactive robot skin with large-area scalability for safer and natural human-robot collaboration in future telehealthcare. *IEEE journal of Biomedical and Health Informatics*. 25(12), 4276-4288. doi: 10.1109/JBHI.2021.3082563.

Jen, M., Goubert, R., Toohey, S., Zuabi, N. & Wray, A. (2021) Triage physicians in an academic emergency department: Impact on resident education. *AEM Education and Training*. 5(3), e10567. doi: 10.1002/aet2.10567.

Jiang, Z., Chi, C. & Zhan, Y. (2021) Research on medical question answering system based on knowledge graph. *IEEE Access*. 9, 21094-21101. doi: 10.1109/ACCESS.2021.3055371.

Karabegović, I. & Doleček, V. (2017) The role of service robots and robotic systems in the treatment of patients in medical institutions. In: Hadžikadić, M. & Avdaković, S. (eds.) *Lecture Notes in Networks and Systems, vol. 3 (Advanced Technologies, Systems, and Applications).* Cham, Switzerland, Springer, pp. 9-25.

Khan, A. W., Al-Obeidat, F., Khalid, A., Amin, A. & Moreira, F. (2023) Sentence embedding approach using LSTM auto-encoder for discussion threads summarization. *Computer Science and Information Systems*. 20(4), 1367-1387. doi: 10.2298/CSIS221210055K.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019) Albert: A lite bert for self-supervised learning of language representations. To be published in *arXiv*. [Preprint] https://arxiv.org/abs/1909.11942 [Accessed 16th October 2023].

Lee, M. L., Park, I., Park, D. U. & Park, C. J. (2017) Constrained ranking and selection for operations of an emergency department. *International Journal of Simulation Modelling*. 16(4), 563-575. doi: 10.2507/IJSIMM16(4)1.388.

Liao, Z., Liu, L., Wu, Q., Teney, D., Shen, C., Van den Hengel, A. & Verjans, J. (2020) Medical data inquiry using a question answering model. In: *Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging, ISBI 2020, 3-7 April 2020, Iowa City, USA*. IEEE. pp. 1490-1493.

Liu, H., Liu, Y., Wong, L. P., Lee, L. K. & Hao, T. (2020) A hybrid neural network BERT-Cap based on pre-trained language model and capsule network for user intent classification. *Complexity*. 2020, 8858852. doi: 10.1155/2020/8858852.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019) RoBERTa: A robustly optimized BERT pretraining approach. To be published in *arXiv.* [Preprint] https://arxiv.org/abs/1907.11692 [Accessed 16th October 2023].

Ma, Y., Liu, H., Zhang, D., Gao, C. & Liu, Y. (2023) A Named Entity Recognition Method Enhanced with Lexicon Information and Text Local Feature. *Tehnički Vjesnik*. 30(3), 899-906. doi: 0.17559/TV-20230121000257.

Marx, E., Usbeck, R., Ngomo, A. C. N., Höffner, K., Lehmann, J. & Auer, S. (2014) Towards an open question answering architecture. In: *Proceedings of the 10th International Conference on Semantic Systems, SEM`14, 4-5 September 2004*, *Leipzig, Germany.* New York, USA, Association for Computing Machinery. pp. 57-60.

Morsi, S. (2023) Artificial intelligence in electronic commerce: investigating the customers' acceptance of using chatbots. *Journal of System and Management Sciences*. 13(3), 156-176.

Naderi, H., Madani, S., Kiani, B. & Etminani, K. (2020) Similarity of medical concepts in question and answering of health communities. *Health Informatics Journal*. 26(2), 1443-1454. doi: 10.1177/146045821988133.

Pan, X., Song, J. & Zhang, F. (2018) Dynamic recommendation of physician assortment with patient preference learning. *IEEE Transactions on Automation Science and Engineering.* 16(1), 115-126. doi: 10.1109/TASE.2018.2839651.

Peinelt, N., Nguyen, D. & Liakata, M. (2020) tBERT: Topic models and BERT joining forces for semantic similarity detection. In: Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J. (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *ACL 2020, Online, 5-10 July 2020*. Kerrville, TX, USA, Association for Computational Linguistics. pp. 7047-7055.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019) Language models are unsupervised multitask learners. To be published in *OpenAI blog*. [Preprint] https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-RadfordWu/9405cc0d6169988371b2755e573cc28650d14dfe [Accessed 16th October 2023].

Ramage, D., Hall, D., Nallapati, R. & Manning, C. D. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, *EMNLP 2009, 6-7 August 2009, Singapore.* pp. 248-256.

Rashid, J., Shah, S. M. A. & Irtaza, A. (2019) Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management*. 56(6), 102060. doi: 10.1016/j.ipm.2019.102060.

Robertson, S. (2004) Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*. 60(5), 503-520. doi: 10.1108/00220410410560582.

Rose, S., Engel, D., Cramer, N. & Cowley, W. (2010) Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*. Hoboken, NJ, USA, John Wiley & Sons, Ltd.

Song, J., Qiu, Y. & Liu, Z. (2016) A real-time access control of patient service in the outpatient clinic. *IEEE Transactions on Automation Science and Engineering*. 14(2), 758-771. doi: 10.1109/TASE.2016.2597185.

Srba, I. & Bielikova, M. (2016) A comprehensive survey and classification of approaches for community question answering. *ACM Transactions on the Web*. 10(3), 1-63. doi: 10.1145/2934687.

Suligoj, F., Jerbic, B., Svaco, M. & Sekoranja, B. (2018) Fully Automated Point-Based Robotic Neurosurgical Patient Registration Procedure. *International Journal of Simulation Modelling*. 17(3), 458-471. doi: 10.2507/IJSIMM17(3)442.

Tavabi, N., Raza, M., Singh, M., Golchin, S., Singh, H., Hogue, G. D. & Kiapour, A. M. (2023) Disparities in cannabis use and documentation in electronic health records among children and young adults. *NPJ Digital Medicine*. 6(1), 138. doi: 10.1038/s41746-023-00885-w.

Tawfik, S. M., Elhosseiny, A. A., Galal, A. A., William, M. B., Qansuwa, E., Elbaz, R. M. & Salama, M. (2023) Health inequity in genomic personalized medicine in underrepresented populations: a look at the current evidence. *Functional & Integrative Genomics*. 23(1), 54. doi: 10.1007/s10142-023-00979-4.

Vatian, A., Dobrenko, N., Andreev, N., Nemerovskii, A., Nevochhikova, A. & Gusarova, N. (2019) Comparative analysis of approaches to building medical dialog systems in Russian. In: *20th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 2019, 14-16 November 2019*, *Manchester, UK*. Cham, Switzerland, Springer. Part I, pp. 175-183.

Veisi, H. & Shandi, H. F. (2020) A Persian medical question answering system. *International Journal on Artificial Intelligence Tools*. 29(6), 2050019. doi: 10.1142/S0218213020500190.

Wang, D., Liang, Y., Ma, H. & Xu, F. (2023) Refined Answer Selection Method with Attentive Bidirectional Long Short-Term Memory Network and Self-Attention Mechanism for Intelligent Medical Service Robot. *Applied Sciences*. 13(5), 3016. doi: 10.3390/app13053016.

Wasim, M., Asim, M. N., Khan, M. U. G. & Mahmood, W. (2019) Multi-label biomedical question classification for lexical answer type prediction. *Journal of Biomedical Informatics*. 93, 103143. doi: 10.1016/j.jbi.2019.103143.

Xue, X., Jeon, J. & Croft, W. B. (2008) Retrieval models for question and answer archives. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, 20-24 July 2008*, Singapore. pp. 475-482.

Yang, S., Yang, X., Zhang, R. & Liu, K. (2022) Hierarchical Progressive Network for Multimodal Medical Image Fusion in Healthcare Systems. *IEEE Transactions on Computational Social Systems.* 10(4), 1540-1558. doi: 10.1109/TCSS.2022.3165559.

Yilmaz, S. & Toklu, S. (2020) A deep learning analysis on question classification task using Word2vec representations. *Neural Computing and Applications*. 32, 2909-2928. doi: 10.1007/s00521-020-04725-w.

Yu, H., Liu, C., Zhang, L., Wu, C., Liang, G., Escorcia-Gutierrez, J. & Ghoneim, O. A. (2023) An intent classification method for questions in "Treatise on Febrile diseases" based on TinyBERT-CNN fusion model. *Computers in Biology and Medicine*. 162, 107075. doi: 10.1016/j.compbiomed.2023.107075.

Zhang, K., Wu, W., Wu, H., Li, Z. & Zhou, M. (2014) Question retrieval with high quality answers in community question answering. In: *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, CIKM 2014, 3-7 November 2014, Shanghai, China*. pp. 371-380.

Zhou, G., He, T., Zhao, J. & Hu, P. (2015) Learning continuous word embedding with metadata for question retrieval in community question answering. In: Zong, C. & Strube, M. (eds.) *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, *26-31 July 2015, Beijing, China*. Red Hook, NY, USA, The Association for Computational Linguistics. pp. 250-259.

Zhu, C. Y. (2023) Research on Emotion Recognition-Based Smart Assistant System: Emotional Intelligence and Personalized Services. *Journal of System and Management Sciences*. 13(5), 227-242.