Enhancing Online Medical Consultations: A Novel Approach to Personalized Doctor Recommendation Using Semantic Features and Response Metrics

Yingbin Zheng, Yunping Cai, Yiwei Yan, Sai Chen, Kai Gong

Table of Contents

Enhancing Online Medical Consultations: A Novel Approach to Personalized Doctor Recommendation Using Semantic Features and Response Metrics

Yingbin Zheng[1] MS; Yunping Cai[2] ME; Yiwei Yan[1] MA; Sai Chen[2] ME; Kai Gong[1] MD

[1]The First Affiliated Hospital of Xiamen University Xiamen City CN
[2]Xiamen Meteorological Bureau Xiamen City CN

Corresponding Author:
Kai Gong MD
The First Affiliated Hospital of Xiamen University
10 Shanggu Road, Siming District
Xiamen City
CN

Abstract

Background: The rapid growth of online medical services has highlighted the significance of smart triage systems in helping patients find the most appropriate doctors. However, traditional triage methods often rely on department recommendations, and are insufficient to accurately match patients' textual questions with doctors' specialties. There is an urgent need to develop algorithms for recommending doctors.

Objective: To develop and validate a patient-doctor hybrid recommendation model with response metrics (PDHR model) for better triage performance.

Methods: A total of 646,383 online medical consultation records from the Internet Hospital of the First Affiliated Hospital of Xiamen University were collected. Semantic features representing patients and doctors were developed to identify the set of most similar questions and to semantically expand the pool of recommended doctor candidates, respectively. The doctors' response rate was designed to improve candidate rankings. These three characteristics combine to create the PDHR model. Five doctors participated to evaluate the efficiency of the PDHR model through multiple metrics and questionnaires, as well as the performance of SBERT and Doc2Vec in text embedding.

Results: The PDHR model reaches the best recommendation performance when the number of recommended doctors is 14. At this point, the model has an F1-score of 76.25%, a proportion of high-quality services of 41.05%, and a rating of 3.90. After removing doctors' characteristics and response rates from PDHR model, the F1-score decreased by 12.05%, the proportion of high-quality services fell by 10.87%, the average hit ratio dropped by 1.06%, and the rating declined by 11.43%. According to whether those five doctors were hit by PDHR model, SBERT achieved an average hit ratio of 88.60%, while Doc2Vec achieved an average hit ratio of 53.40%.

Conclusions: The PDHR model uses semantic features and response metrics to enable patients to accurately find the doctor that best suits their needs.

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
  Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
  Only make the preprint title and abstract visible.
  No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Running head: NOVEL APPROACH TO PERSONALIZED ONLINE DOCTOR RECS   1

Enhancing Online Medical Consultations: A Novel Approach to Personalized Doctor Recommendation Using Semantic Features and Response Metrics

Yingbin Zheng[1], MS; Yunping Cai[2], ME; Yiwei Yan[1], MA; Sai Chen[2], ME; Kai Gong[1*], MD

1 Biomedical Big Data Center, The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen 361003, China
2 Meteorological Disaster Prevention Technology Center, Xiamen Meteorological Bureau, Xiamen 361003, China
Corresponding Author:
Kai Gong freatink@xmu.edu.cn
The First Affiliated Hospital of Xiamen University
10 Shanggu Road, Siming District
Xiamen City, 361003
China

ABSTRACT

Background: The rapid growth of online medical services has highlighted the significance of smart triage systems in helping patients find the most appropriate doctors. However, traditional triage methods often rely on department recommendations, and are insufficient to accurately match patients' textual questions with doctors' specialties. There is an urgent need to develop algorithms for recommending doctors.

Objective: To develop and validate a patient-doctor hybrid recommendation model with response metrics (PDHR model) for better triage performance.

Methods: A total of 646,383 online medical consultation records from the Internet Hospital of the First Affiliated Hospital of Xiamen University were collected. Semantic features representing patients and doctors were developed to identify the set of most similar questions and to semantically expand the pool of recommended doctor candidates, respectively. The doctors' response rate was designed to improve candidate rankings. These three characteristics combine to create the PDHR model. Five doctors participated to evaluate the efficiency of the PDHR model through multiple metrics and questionnaires, as well as the performance of SBERT and Doc2Vec in text embedding.

Results: The PDHR model reaches the best recommendation performance when the number of recommended doctors is 14. At this point, the model has an F1-score of 76.25%, a proportion of high-quality services of 41.05%, and a rating of 3.90. After removing doctors' characteristics and response rates from PDHR model, the F1-score decreased by 12.05%, the proportion of high-quality services fell by 10.87%, the average hit ratio dropped by 1.06%, and the rating declined by 11.43%. According to whether those five doctors were hit by PDHR model, SBERT achieved an average hit

ratio of 88.60%, while Doc2Vec achieved an average hit ratio of 53.40%.

Conclusions: The PDHR model uses semantic features and response metrics to enable patients to accurately find the doctor that best suits their needs.

Keywords: Online medical service, Text analysis, SBERT, Smart triage systems, PDHR model

INTRODUCTION

Online medical consultation is increasingly popular as an alternative to traditional healthcare services because it is convenient, accessible, and affordable [1]. This type of doctor-patient interaction takes place electronically, connecting both parties through text, images, and videos. Its advantages include eliminating time and space constraints and accurately documenting the medication process [2], making it more attractive to many patients than in-person medical visits. As of the end of 2022, the number of users in China's internet medical and health market reached 363 million [3]. The rapid growth of online medical services and the vast amount of information available have created considerable difficulties for patients in finding the doctors best suited to their needs, leading to potentially mismatched consultations [4].

At present, most existing triage procedures rely on manual recommendation from schedulers to select departments for patients. As the number of consultations increases, manual provision of advice does not guarantee the professionalism and quality of medical services [5]. Additionally, schedulers are unable to provide 24-hour service, resulting in gaps in healthcare access and continuity of services. At this point, a common approach might be to develop an intelligent department recommendation model. Advancements in technology, particularly in the field of machine learning, present opportunities to improve the accuracy and efficiency of patient department assignment in healthcare systems. For example, Mullenbach et al. [6] integrated the attention mechanism and utilized LSTM to predict the patient's disease type for further triage. Li and Yu [7] used multi-filter residual convolutional neural networks to investigate the issue of department recommendation. Wang et al.[8] utilized the BERT model to study disease diagnosis and department recommendations. These approaches can potentially automate the process of assigning patients to appropriate departments, reducing the burden on schedulers and improving patient outcomes through more accurate and timely care.

However, due to the ongoing subdivision of departments, these department recommendation models still cannot accurately match medical needs with doctors' specialties. For example, obstetricians and gynecologists further specialize into subfields such as gynecology, obstetrics, reproductive endocrinology, infertility, prenatal diagnosis, and genetic counseling. This refined division not only improves the effectiveness of diagnosis and treatment, but also ensures that patients receive the most cutting-edge and professional care plans. In addition, even if the diseases treated are similar or the same, different medical institutions may have different department names. These problems have placed higher demands on hospital management, requiring more precise resource allocation to adapt to increasingly specialized services. Therefore, there is an urgent need to design personalized doctor recommendation models.

Personalized recommendation methods can help users manage massive amounts of information and knowledge [9] and are crucial for providing personalized medical services that meet the patient's needs [10]. For instance, Ju and Zhang [11] integrated geographical location and patients' questions to generate personalized recommendations. Liu et al. [12] proposed a recommended doctor model that considers characteristics of patients and doctors. Lu et al. [5] proposed a self-adaptive doctor recommendation system that considered doctor activity and patient feedback. These methods can be advantageous for both patients and online healthcare providers, as it minimizes the time and effort required to find a suitable match, thus ensuring efficient delivery of healthcare services [13].

However, there are still some shortcomings in previous researches. Most existing studies use satisfaction as a measure of doctor performance. But the authenticity of satisfaction ratings across different platforms is not always reliable, as many users tend to habitually provide positive feedback. In terms of the evaluating indicators for recommended doctors, most studies used accuracy as a single indicator and do not consider the service quality of recommended doctors. These limitations may result in consultation mismatches, higher patient waiting costs, and potentially reduced patient satisfaction. To the best of our knowledge, previous studies have not developed a triage system for recommending doctors that utilizes the transformer-based models, which are the cutting-edge models for nature language processing. The Bidirectional Encoder Representations from Transformers (BERT) [14] is a popular transformer-based model that has been pre-trained on common texts such as Wikipedia and the Brown Corpus. BERT is a state-of-the-art model that utilizes an attention-based mechanism [15, 16] to accurately understand the context of words, enabling unsupervised learning by linking text input and output through a decoder-encoder framework [17, 18]. However, the BERT model is not suitable for semantic similarity searches or clustering, which has led to the creation of a different sentence-embedding model called the Sentence-BERT (SBERT) model [19]. This modified version of BERT model was designed to be semantically meaningful and suitable for sentence similarity tasks. It works by integrating a Siamese network and a pre-trained BERT model, along with a pooling layer that generates a fixed-sized representation. The SBERT model can accurately identify if there is a significant match between two sentences, making it a useful tool for data mining, information retrieval, and text-matching [20].

The objective of this study is to develop a more precise algorithm that can better recommend professional and highly engaged doctors, thus improve the effective utilization of medical resources and the medical experience of online patients, through reducing the mismatches between medical needs and services. The practical benefits expected from our findings include the enhanced ability of online healthcare platforms to provide timely, relevant, and professional medical consultations that are closely tailored to each patient's unique needs. By implementing our advanced recommendation algorithm, we expect to not only identify the most appropriate specialists based on patient input, but also incorporate a comprehensive evaluation of physician performance metrics. This will ensure that patients are recommended physicians who are not only experts in their field, but also highly engaged and responsive, resulting in higher quality care.

We seek to answer these three questions:

1. How can we effectively construct features for patients and doctors to facilitate efficient doctor recommendations?

2. How can we incorporate the doctors' performance metrics into recommendation strategies to increase the chance of recommending highly active doctors?

3. How can the effectiveness of the recommendation strategy be verified considering both accuracy and service quality?

METHODS

Data Collection

This research collected a total of 646,383 online medical consultation records from the Internet Hospital of the First Affiliated Hospital of Xiamen University between 2016 and 2023. Each record contains the textual question, de-identified codes for the doctor and patient, the doctor's department, as well as the response status and time. The response status refers to whether the record has received a reply from the corresponding doctors. Response time is the duration between submitting a request and getting a response. Five examples of the questions generated during online medical consultations are displayed in Table 1.

Table 1 Examples of patients' consultation questions

Sample Number	Patient Code	Questions	Doctor Code	Department	Response Status[a]	Response Time
1	200321	You initially diagnosed me with left varicocele and ordered a color Doppler ultrasound examination. The results showed a moderate left varicocele with reflux. I would like to inquire whether, aside from surgery, this condition can be treated through medication, injections, or other noninvasive methods?	208	Urology	True	8 h 12 min
2	306878	My child has been experiencing discomfort in the throat and recurring fevers for four days before visiting a doctor, who considered it was pneumonia and started administering azithromycin. Today is the fifth day of treatment. After the first day of intravenous azithromycin, the fever subsided, but there is still occasional coughing with phlegm. I am worried about the potential for significant side effects. Should the child continue taking azithromycin?	372	Pediatrics	False	--
3	447138	I am currently on my period and have scheduled an ultrasound and mammography for this afternoon. Could you please tell me if there is a recommended waiting period before trying to conceive after a mammography?	133	Breast Surgery	True	20 min
4	591872	What does the glucose tolerance test report indicate? Could you please explain it to me?	423	Obstetrics	True	9 min
5	603826	My chin is red without feeling painful or itchy, and it has been like this for over a month. I've tried Clotrimazole but no obvious effects were achieved. Could you please tell me what condition this might be and what medication I should use?	418	Dermatology	True	2 h 20 min

True: The doctor has responded to the consultation; False: The doctor has not responded to the consultation.

These records were divided into two test datasets and one training dataset. For the first test dataset, the doctor with the highest number of consultations will be selected from each of the following departments with the most inquiries: gastroenterology, obstetrics, respiratory medicine, pediatrics, and dermatology. Their codes are 98, 141, 202, 512, and 601, respectively. 400 consultation records will be randomly selected from each of the aforementioned doctors. These doctors will then review these textual questions to determine if they are within their expertise. Any questions that a doctor is proficient in will be tagged, and eventually we randomly selected 200 records for each doctor from these tagged questions to compile a test dataset consisting of 1000 records. For the second test dataset, a sample of 10,000 consultations was randomly chosen from the total dataset, excluding the consultation samples from the first test dataset. The training dataset consists of the consultations remaining after the removal of the first and second test datasets. The random seed for this study was set to 2023.

Data Pre-processing

Data related to patients' consultation questions has been collected and presented in the form of natural language. Pre-processing of these unstructured data is crucial in machine learning framework [21] to remove unnecessary, duplicated, irrelevant, and noisy data [22]. This study involved several steps to process these consultation questions, including normalization, tokenization, part-of-speech tagging, and stop-word removal, thereby forming a reliable corpus.

The study calculated response rates and times for all doctors as shown in Equations (1) and (2):

$$response_{ratei} = \frac{N_R}{N} \qquad （1）$$

$$response_{timei} = \frac{S_T}{N_R} \qquad （2）$$

Where N_R denotes the count of consultations that Doctor D_i has responded to, with " responded " indicating that the response status is confirmed as true. Meanwhile, N indicates the total number of consultations that doctor D_i has received. Furthermore, S_T refers to the total response time for all the consultations that Doctor D_i has responded to.

Upper and lower bounds on response times were established to minimize the impact of extremely high and low values on the experiment. Response times above 95% were capped at 8 hours and 6 minutes, while those below the 5th percentile were raised to 9 minutes.

Feature Extraction

Feature extraction is the process of converting raw input data into a meaningful set of features [23] that can be understood by machine learning classifiers. In the feature extraction stage, two unique features for both patients and doctors were introduced.

Patients' Features Modeling

This study employed the pre-trained Sentence Bidirectional Encoder Representations from Transformers (SBERT) model known as "distiluse-base-multilingual-cased", to convert all consultation questions into semantic representations and then calculates sentence embeddings for further analysis. As shown in figure 1 [19], the SBERT model processed Sentence A and Sentence B through BERT pooling to generate their respective embeddings, u and v. The similarity between these embeddings is then calculated using the cosine similarity method, which effectively measures how similar sentences are. The cosine similarity is expressed by Equation (3), where u and v represents two vectors:

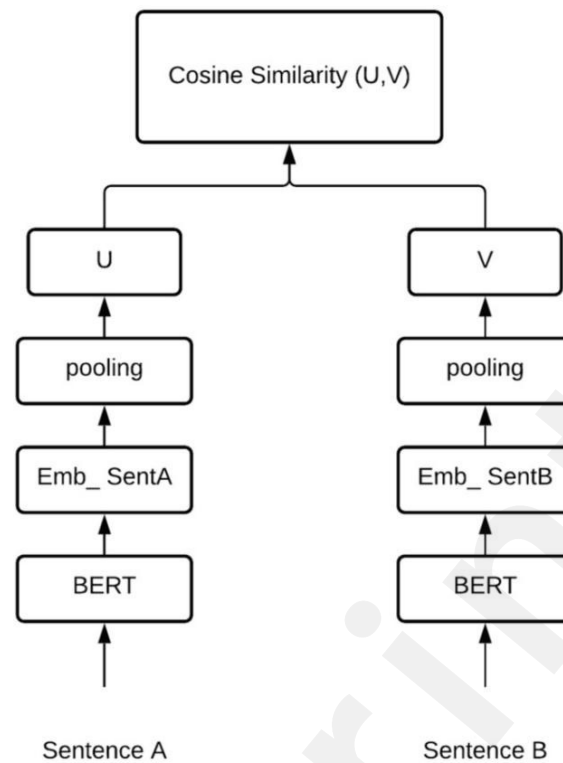$$¿(u,v) = \frac{u \cdot v}{\|u\| \cdot \|v\|} \qquad （3）$$

Figure 1 SBERT architecture

Doctors' Features Modeling

TF-IDF model [24, 25] is a commonly used method in text mining and information retrieval because it can capture the importance of words and has the potential to extract features from multiple texts. The formula of this algorithm is shown in equation (4):

$$TF-IDF(t,d)=TF(t,d)\cdot IDF(t) \qquad\qquad (4)$$

Where $TF(t,d)$ represents the frequency of a specific keyword t in document d, while $IDF(t)$ signifies the inverse document frequency. According to this formula, the higher the $TF-IDF(t,d)$ value, the more significant the feature is in the document.

This study utilized TF-IDF model to extract crucial information from a collection of patients' consultation questions aggregated by doctor codes, selecting the top 20 with the highest TF-IDF weights. This extracted information was then fed into a SBERT model to compute cosine similarity among doctors.

Recommendation

A patient-doctor hybrid recommendation model with response metrics (PDHR model) was developed by combining features of both patients and doctors. This model also takes into account the doctor's response rate to recommendation strategy. PDHR model, which is a type of top-k recommendation system, is designed to provide patients with a list of the top k doctors who are most likely to meet their medical needs, as illustrated in Figure 2.
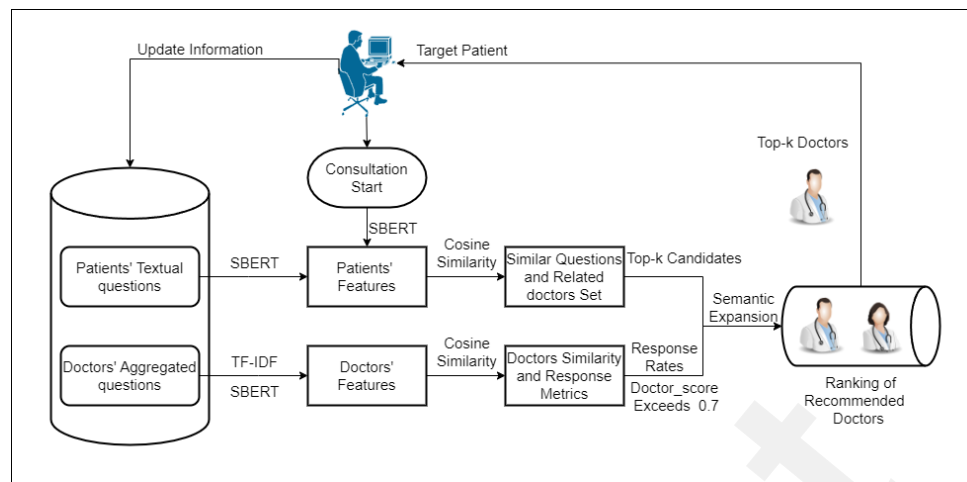
Figure 2 the architecture of PDHR model

Step1: Generate a candidate set using the patient feature-based (PFB) model

The patient feature-based (PFB) model was developed to identify sets of the most similar questions. When a new consultation starts, the consultation questions will be processed using SBERT to generate the corresponding embeddings. The patients' features were used to construct a similarity matrix among questions using cosine similarity. Then, consultation questions that are similar to the new consultation will be identified by comparing the patients' features. The doctors who are associated with these similar questions are considered potential candidates. The similarity score, known as the $init_{score}$, serves as the baseline for making recommendations. The top-k doctors are selected as candidates from the set of similar questions, where k is an adjustable hyperparameter in this model.

Step2: Expand the candidate set based on the patient-doctor hybrid (PDH) model

Since patients' textual questions are unprofessional, setting a similarity threshold based solely on the patient characteristics may limit the recommendation results. The patient-doctor hybrid (PDH) model ensures that all potential doctor recommendations are considered. This model is formed by combining the doctors' features with the PFB model to semantically expand the scope of candidates. It does this by creating an index called the $expand_{score}$, which is derived from the features of doctors. This index reflects the degree of similarity among doctors and helps determine which doctors have the necessary expertise and qualifications to provide the right care for a given patient. This approach can adjust biases in the system that may arise from recommending doctors based solely on similarities to patients' questions. The PDH model is shown in Equation (5):

$$doctor_{scorei} = init_{scorei} \cdot expand_{scorei} \qquad （5）$$

If the $doctor_{score}$ exceeds 0.7, it will be used for semantically expand the range of candidates. When the doctor is not derived from PDH model, the $expand_{score}$ is assumed to be 1.

Step3: Optimize the ranking of the candidate set by incorporating the response rate

The response rate can serve as an indicator to measure the efficacy of doctors' performance. An increase in the response rate suggests that doctors are more willing to treat patients. This can be viewed as a positive feedback loop, as higher response rates lead to more motivated doctors. Therefore, it's crucial to take into account the doctors' activity level along with the similarity index, as this can skew the recommendation results towards more active doctors, increasing the chance that inquiries will be answered and thus improving patient satisfaction. The final PDHR model is displayed in Equation (6). The top-k doctors were selected for recommendation based on the scores, where n represents the number of times doctor D_i is recommended:

$$score_i = response_{ratei} \cdot \sum_n doctor_{scorei} \qquad （6）$$

Evaluation

The proposed PDHR model's effectiveness will be evaluated using metrics including hit ratio, precision, recall, F1-score, and high-quality service proportion. In the first test dataset, a recommendation is considered correct when the selected doctor is among the top-k recommended doctors. The hit ratio refers to the proportion of correct recommendations to the total number of recommendations. In the second test dataset, a recommendation is regarded as accurate if the recommended doctor's maximum $doctor_{score}$ is greater than 0.7. Precision refers to the proportion of correctly recommended doctors to the total recommended doctors. On the other hand, recall is the proportion of correctly recommended doctors to the doctors that should have been retrieved in the sample. The F1-score is a valuable metric for assessing the recommendation algorithm's effectiveness, as it merges precision and recall to yield the best results. A higher F1-score signifies a more efficient algorithm. Precision, recall, and F1-score are calculated using the formulas described in (7)-(9):

$$Precision = \frac{TP}{TP+FP} \qquad (7)$$

$$Recall = \frac{TP}{TP+FN} \qquad (8)$$

$$F1-score = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \qquad (9)$$

Where TP is a true positive, FP is a false positive, FN is a false negative.

A quick response time is a critical element for high user satisfaction, allowing the system to promptly provide services that meet patient expectations. If a doctor responds quickly, the patient will perceive the quality of the doctor's service to be better than that of a doctor who takes longer time to respond. Therefore, the proportion of doctors who respond quickly among all recommended doctors, known as the high-quality services proportion, is a significant measure of evaluation. The calculation formula for high-quality services proportion is determined by the formula shown in (10):

$$hqos_i = \frac{N_f}{N} \qquad (10)$$

The term N_f represents the number of doctors whose response time is faster than the average response time, while N denotes the number of recommended doctors.

Baseline Experiments

For the PFB model, the purpose of this baseline was to determine whether excluding doctors' features from PDHR model would degrade performance. Three steps were taken to assess its performance compared to the PDHR model. Firstly, the hit ratio and ranking of the selected doctor in the recommendation set were calculated in the first test dataset. Secondly, the precision, recall, and F1-score of the recommendation results were computed in the second test dataset, and the consultation questions that each selected doctor was hit by these two models were collected. Finally, a questionnaire for assessed the rationality of the recommendations was conducted by randomly selecting 200 consultation questions (100 for each model) for each doctor from those consultation questions. The questionnaire included an evaluation of the relevance of each selected doctor with the consultation questions. The survey question is: "Based on your area of expertise, how would you rate the match between you and consultation question?" The questionnaire used a Likert five-point [26] scale for measurement, with scores ranging from 1 (very inappropriate) to 5 (very appropriate). The Mann-Whitney U test [27] was used to determine whether there was a statistical difference in the doctors' perceptions of how well the consultation questions from these two models matched their area of expertise.

For the PDH model, the proportion of high-quality services in the recommendation results were calculated to assess whether eliminating the response rate in the PDHR model will reduce service quality.

For Doc2Vec[28], we utilized it to create text embeddings for all patients' consultation questions to reconstruct PDHR model. The performances of Doc2Vec against SBERT were evaluated in the first test dataset to determine the effectiveness of transfer learning without contextual modeling. The model's performance is measured by the hit ratio and the ranking of the selected doctor within the set of recommendations.

RESULTS

Dataset Summary

Among the 646,383 consultation records, there were 193,675 patients and 858 doctors across 44 departments. According to Table 2, which provides a summary at the record level, 32.95% of the records were created by male patients, while female patients accounted for 67.05%. The predominant age group among patients was 20-39 years, representing 54.60% of the total. Patients most frequently consult senior doctors, who account for 62.65% of all consultations. The majority of consultations were initiated between 12-17 hours, accounting for 37.04% of the total, while the bulk of responses were received between 18-23 hours, accounting for 40.94%. The average response time of the doctors was three hours and forty minutes with an average response rate of 65.20%.

Table 2 Summary of the characteristics of the collected data records (N = 646,383)

Characteristic	Value, n (%)
Gender	
Male	212,983 (32.95)
Female	433,400 (67.05)
Age(years)	
<20	118,484 (18.33%)
20-39	352,907 (54.60%)
40-59	125,957 (19.49%)
>60	49,035 (7.58%)
Doctors' professional title	
Junior	10,766 (1.67%)
Intermediate	45,892 (7.10%)
Sub-senior	184,767 (28.58%)
Senior	404,958 (62.65%)
Consulted created moment (hours)	
0-5	15,686 (2.43%)
6-11	195,297 (30.21%)
12-17	239,401 (37.04%)
18-23	195,999 (30.32%)
Consultation responded moment (hours)	
0-5	13394 (2.61%)
6-11	115684 (22.73%)
12-17	171634 (33.72%)
18-23	208417 (40.94%)
Response status	
True	421,441 (65.20%)
False	224,942 (34.80%)

Case Analysis

This study recommends a question, as shown in Sample 4 in Table 1: "What does the glucose tolerance test report indicate? Could you please explain it to me?" This was done to verify the feasibility of PDHR model. Questions similar to the target patient's question and related candidate

doctors are displayed in Table 3.

Table 3 Top-10 questions and related doctors similar to the target patient's question

Question Code	Similarity	Doctor Code
394946	0.9766	178
559249	0.9765	423
317643	0.9618	141
409326	0.9419	164
238700	0.9214	456
2416	0.9004	707
173519	0.8990	304
551580	0.8976	330
466072	0.8906	632
93556	0.8861	391

The doctors who were similar to the candidate doctors were identified and included in the recommendation strategy along with response indicators. The recommended doctors of PDHR model are displayed in Table 4.

Table 4 Top-10 recommended doctors

Doctor Code	$score$	$doctor_{score}$
141	7.1877	0.9618
164	5.9507	0.9419
423	4.1309	0.9765
456	4.0309	0.9214
166	3.9408	0.8978
178	3.2916	0.9766
169	3.2082	0.7851
181	2.7208	0.7310
335	2.7096	0.7672
189	2.4849	0.7358

To evaluate the precision of the recommended results, we compared the consultation question within the recommended results. For example, Doctor 141's diagnosis includes "gestational diabetes" and " glucose tolerance", while Doctor 164's diagnosis includes "glucose tolerance" and "diabetes". This information matches the consultation question of sample number 4, suggesting that the recommended results are likely accurate.

Evaluating the Effectiveness and Service Quality

Figure 3 shows a comparison of PDHR model with PFB and PDH model in terms of various indexes. The x-axis represents the number of recommended doctors (K), with values ranging from 2 to 20 and increasing in increments of 2. PDHR and PFB model exhibit comparable performance when K is below 10. However, as K increases to 10 or more, PDHR model demonstrates a marked improvement over PFB model. PDHR model achieves its highest F1-score when K is 14, indicating optimal performance at this level. At this stage, the model has a precision of 71.26%, a recall of 82.02%, and an F1-score of 76.25%. Compared to PFB model, the precision has increased by 15.43%, the recall has increased by 9.10%, and the F1-score has increased by 12.05%. The results presented in table 4 indicate that although there are minor fluctuations in the rankings of selected doctors of PDHR model, its hit ratio has increased by 2.19% compared to PFB model. Those indicate that incorporating doctor features into the recommendation strategy can improve the effectiveness of the recommendation system.

Regardless of the value of K, PDHR model provides better $hqos_i$ than the PDH model. When K is set to 14, the high-quality service ratio of the PDHR model is 41.05%. This is an improvement of

10.87% over the PDH model. This suggests that incorporating response rate into the recommendation strategy can enhance service quality.
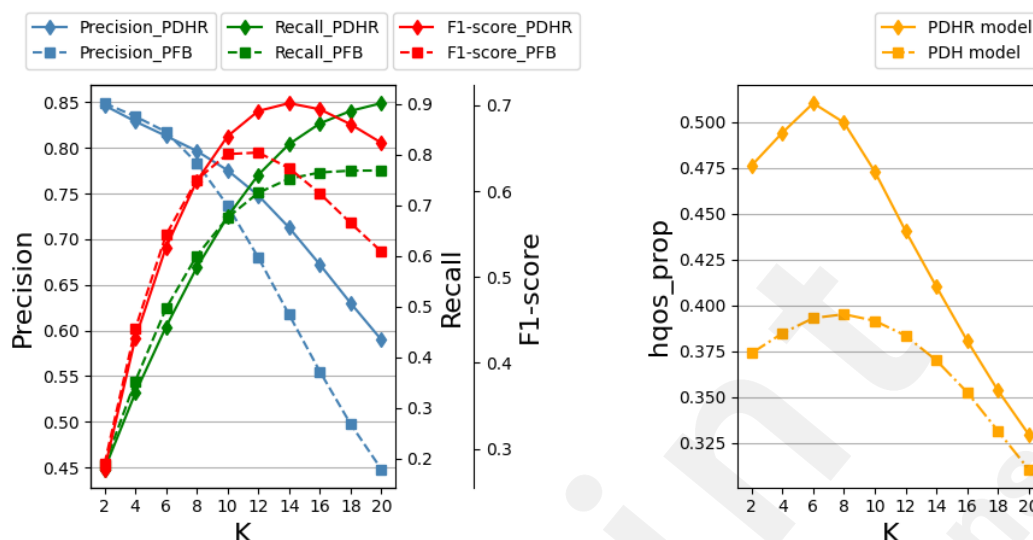

Figure 3 Comparison of our proposed PDHR model with PFB model and PDH model in terms of various indexes, including precision (blue), recall (green), F1-score (red) results, and $hqos_i$ (orange).

Evaluating the Performance of Text Embedding

The results displayed in Table 5 compare the SBERT model with the Doc2Vec model in terms of text embedding, with K set to 14. The ranking indicates the position of the selected doctor within the recommendation set, while hit ratio represents the percentage of successful recommendations. The SBERT model surpasses the Doc2Vec model by 65.92% in hit ratio and the rankings of the selected doctors improved by 1.57. These findings suggest that using the SBERT model to create text embeddings could improve the performance of the recommendation system.

Table 5 Performance comparisons were conducted using different models for selected doctors

Doctor Code	Hit Ratio			Ranking		
	Doc2Vec-PDHR	SBERT-PDHR	SBERT-PFB	Doc2Vec-PDHR	SBERT-PDHR	SBERT-PFB
98	0.485	0.850	0.820	4.55	3.19	3.04
141	0.555	0.915	0.890	3.99	2.36	2.47
202	0.430	0.800	0.780	4.93	3.47	3.33
512	0.505	0.940	0.925	5.95	3.37	2.73
601	0.695	0.925	0.920	2.78	1.97	1.90
Average	0.534	0.886	0.867	4.44	2.87	2.69

Rationality Evaluation

As can be seen from Table 6, for each doctor, the PDHR model consistently receives higher ratings than the PFB model. The average rating of the PDHR model is 3.90, which is 11.43% higher than that of the PFB model. The p-values indicate that the differences in ratings between these two models are statistically significant, indicating that the PDHR model is capable of recommending better-performing doctors compared to the PFB model.

Table 6 Rationality evaluations of PDHR model and PFB model for selected doctors

Doctor Code	Rating		P-value
	PDHR	PFB	
98	3.89	3.48	0.0106
141	4.00	3.57	0.0150
202	3.79	3.33	0.0226
512	3.83	3.54	0.0491
601	3.98	3.56	0.0259
Average	3.90	3.50	-

DISCUSSION

Principal Findings

In this study, we developed an innovative doctor triage algorithm named the PDHR model. This model improves the accuracy of matching patients' textual questions with the doctors' specialties and optimizes the ranking of candidates according to the doctors' service performance. Consequently, PDHR model may help increase both the efficiency and the quality of online medical services by recommending active doctors with the most appropriate specialties.

Challenges and Solutions for Online Triage Systems

Triage service is a preliminary service in medical diagnosis [29], serving as the first point of contact for patients in healthcare. They are crucial for improving efficiency and precision of medical services. In offline outpatient clinics, patients' choices are limited due to the doctors' fixed schedules, especially if the appointment times cannot be changed. Therefore, triage is usually performed at the department level [30]. In contrast, online consultation services typically do not adhere to a fixed schedule, and all doctors can provide services online, so patients have a wider range of choices. However, the current triage systems have inherited the offline departmental recommendation form, which provides limited assistance to patients. Additionally, due to the ongoing division of departments, the naming conventions of these departments have become confusing, resulting in possible overlap in disease areas that doctors specialize in across various departments. Furthermore, with the development of regional medical platforms [31], doctors from different regions and multiple hospitals may share the same online consultation platform, which complicates the supply of medical services. Therefore, it is imperative to develop a new type of doctor recommendation system.

The construction of a triage system must first consider the matching of the doctors' specialties and the patients' medical needs, which is a prerequisite for the effective operation of online medical services [32]. The triage system needs to accurately create user profiles for doctors and patients, analyzing their characteristics, and achieve precise matching. Traditional triage systems typically use the profiles of professional-level doctors. These profiles mainly include disease names and other professional terminology that doctors specialize in, which can be difficult to match with patients' textual questions. Patients often ask their doctors questions using non-professional, colloquial descriptions of symptoms rather than precise disease names. Therefore, using professional descriptions to create doctor profiles does not semantically match well with patient's questions. Some studies have attempted to extract doctor features using textual questions from patients [5, 11, 12]. Our study draws on this approach, using natural language processing technology to build doctor characteristics based on a large corpus of patient inquiries, thus constructing the profiles from the patient's perspective and aligning more closely with patient needs.

In addition to expertise, the quality of service provided by doctors is equally important in ensuring effective online diagnosis and treatment. In online services, the quality of service is particularly reflected in the response time and rate, as well as the thoroughness of the content provided. Formally,

response time and rate are obvious and accessible indicators. Previous researches did not take these into account when developing triage systems. Therefore, we have included consideration of the response rate in the scoring calculation of our model's ranking. Since there is a correlation between response rate and time, our results showed that this approach also significantly improved the recommended doctors' response time.

Feasibility and Potential Extensions of the Proposed Model

The most significant difference between online and offline medical consultations is the limited availability of data. When doctors cannot physically examine patients, the dialogue generated during the consultation becomes the primary source of usable information. Due to potential incompatibilities and lack of data sharing between online and offline systems [26], patients' medical histories are often missing on most online consultation platforms, making it more challenging to extract patient characteristics. In terms of the quality of medical services, patient satisfaction with doctors is an indicator that can be referenced. However, the authenticity of satisfaction ratings on different platforms is not always reliable, as many users tend to habitually give positive feedback. Therefore, whether satisfaction ratings should be included in the model remains to be studied and verified. The PDHR model was designed to use the minimal information to match doctors and patients. Despite the limited number of variables included, the advantage is that the algorithm is portable across different platforms, offering greater versatility and suitability for widespread adoption. Subsequently, different platforms can modify the model to suit their specific circumstances, including incorporating past medical histories and satisfaction ratings. They can also adjust various hyperparameters within this framework, such as adjusting the weight of the response rates or setting a different number of recommendations to meet the needs of different platforms. Additionally, the user interface can display information that the model utilized or disregarded, providing additional support for patient decision-making. Considering that some doctors may not be familiar with online platforms, it is also feasible to show indicators of their offline services. For instance, it is important to consider if a doctor has a sufficient number of in-person appointments and the level of satisfaction expressed by patients regarding those services. If a doctor's expertise is a good match for a specific type of consultation and they have outstanding offline reviews, despite not being highly active online, patients could consider switching to offline consultations during clinic hours.

Real-World Application Challenges

Implementing the PDHR model in a real-world scenario presents a number of challenges. First and foremost, ethical issues related to privacy, consent, and confidentiality are major concerns. It is of the utmost importance that patients' medical histories and personal information are handled with the utmost care, in accordance with the patients' consent and in compliance with local and international laws. The PDHR model training process includes pseudonymization or anonymization of data by removing or replacing personally identifiable information, as well as the use of SBERT for text vectorization, which transforms textual data into numerical vectors that represent the semantic meaning of the text but do not contain explicitly identifiable information, preventing the extraction of personal information directly from the vectors to protect individual privacy. It is also important to consider data confidentiality, encrypting data at rest and in transit protects against unauthorized access. Using strong encryption standards, such as Advanced Encryption Standard (AES) for data at rest and Transport Layer Security (TLS) for data in transit, can ensure that even if data is intercepted or accessed, it remains unreadable and secure. Implement strong authentication mechanisms to verify the identity of users accessing the system. Use multi-factor authentication (MFA) to add an extra

layer of security. Implement role-based access control (RBAC) to ensure that users can access only the data relevant to their role, maintaining the principle of least privilege.

Integration with existing healthcare IT systems is another challenge. Many healthcare providers use legacy systems that may not be immediately compatible with newer models such as PDHR. This requires the development of interfaces or middleware that can seamlessly connect the model to various healthcare IT infrastructures without disrupting existing workflows.

In addition, patient and physician acceptance is an integral part of the implementation process. Many users may be skeptical or resistant to changing traditional consultation methods. Educating both physicians and patients about the benefits of the PDHR model, such as increased efficiency, better physician-patient matching, and ultimately improved healthcare outcomes, is critical to facilitating adoption.

Evolving Text Feature Extraction

Text embedding is a fundamental method for text feature extraction, where Doc2Vec is an effective means of implementing text embedding [33]. However, with the advent of transformer-based models, previous text embedding methods are gradually being replaced by SBERT in the industrial service sector. First, Doc2Vec provides a static embedding for each word, best used for tasks that can benefit from representations without the need for understanding word-context relationships [34]. SBERT provides dynamic contextual embeddings that allow for a deeper understanding of the meaning of words in context. It also has the ability to transfer knowledge and analyze subwords [35], which are essential for more complex language comprehension tasks.

Second, the computational efficiency of SBERT compared to traditional methods such as Doc2Vec is primarily influenced by its transformer architecture. Transformers take advantage of parallel processing, which significantly speeds up the training phase. However, they also tend to be resource-intensive, primarily due to the need for larger memory footprints to handle the contextual embeddings and underlying mechanisms. For the PDHR model, this means that there could be increased computational requirements, especially when processing a large corpus of patient queries or generating physician profiles.

Third, when considering the scalability of SBERT within the PDHR model for widespread use, several strategies can help mitigate potential challenges. Horizontal scaling, or adding more machines to spread the load, is a straightforward approach but can increase costs. More efficient strategies include the use of cloud-based services that offer dynamic scaling options to accommodate fluctuating demand without the need for constant, high-capacity infrastructure. Another key consideration is optimizing input sequences. By limiting the length of textual input without losing critical information, the PDHR model can reduce the processing required for each query, making the system more responsive. In addition, caching frequently accessed embeddings and using batch processing for embedding generation can significantly reduce the overall computational load.

It also indicates that as technology progresses, the underlying technical components of models must be regularly updated and refined to enhance the system's overall efficiency. This is a real-world challenge that any online medical triage system in operation will encounter.

Limitations and Future Directions

There are some limitations and further solutions. First, it is important to collect data from multiple sources. The current study was limited to one hospital, which casts doubt on whether the findings are relevant in different contexts. To overcome this limitation, subsequent researches should aim to collect information from various sources to evaluate the efficiency of the proposed algorithm. Second, there were some irrelevant contents in our datasets. For example, questions like "Doctor, will you be available tomorrow? Where can I find you?" These business process related questions are

often mixed with medical questions describing symptoms and represent noises in the dataset [36]. Even though preprocessing methods may reduce these noises, manual involvement might still be necessary to enhance the data quality. Third, a common limitation of Deep Neural Networks (DNN) [37] is the lack of a natural method to explain their predictive results, which makes it difficult to understand why specific samples are predicted to be similar. Models based on transformers make it very challenging to identify when unfair biases or spurious correlations might drive predictions. Therefore, we have introduced the involvement of doctors. If doctors could provide more information based on order details, such as scoring from the perspectives of professional suitability and willingness to accept orders, it might effectively improve the final performance of the model. Fourth, it is critical to regularly update the doctors' professional information because this information changes over time. Relying on outdated data can result in less than ideal recommendations. To ensure that doctors' profiles are up to date, a time range feature can be implemented. This feature automatically deletes data beyond the specified time range and periodically updates the model with only the latest data. This approach can improve the chances of making accurate recommendations for active doctors and reduce the chances of those still in training or changing areas of expertise. Fifth, obtaining valid feedback from patients and doctors is essential to validating the model's benefits to patients in real-world settings on a larger scale. For instance, surveys can be conducted on patients' usage of the system, whether they have adopted the system's recommendations, and their feedback on whether the system has been helpful. Observing changes in metrics such as the number of consultations for the same condition before and after using the system, comparing their outcomes, and examining the health economic effects of the system are also important. Finally, we were unfortunately only able to obtain textual data to develop the PDHR model. However, the triage system framework proposed in this paper has the potential to incorporate various types of data beyond text. It is possible to integrate multimodal information, such as text, images, audio, and video, using vector embedding techniques to create new vector features. Based on this, calculating similarities could potentially achieve more precise matching.

CONCLUSIONS

This paper present a patient-doctor hybrid recommendation model with response metrics that uses natural language processing techniques to tackle online medical triage tasks. The system filters out relevant doctors, aiding patients in finding the doctors that best suit their actual medical requirements. This approach has significant practical value and can be incorporated into various health website systems to enhance the quality of doctor recommendations.

ACKNOWLEDGEMENT

DATA AVAILABILITY

The datasets involving patients' individual information during the study are not publicly available due to the privacy protection and data security regulations.

AUTHORS' CONTRIBUTIONS

KG contributed to the study concept and provided overall guidance. YY contributed to the acquisition of data. YZ contributed to model development and manuscript drafting. YC and SC contributed to data annotation. KG and YZ critically revised the manuscript. All authors contributed to the interpretation of the results and gave final approval of the manuscript.

CONFLICTS OF INTEREST

The authors have no conflicts of interest.

References:

1.   Haleem A, Javaid M, Singh RP, et al. Telemedicine for healthcare: Capabilities, features, barriers, and applications. *Sens Int*. 2021;2:100117. Epub 2021/11/23. doi: 10.1016/j.sintl.2021.100117. PubMed PMID: 34806053; PubMed Central PMCID: PMCPMC8590973.

2.   Hajli MN. Developing online health communities through digital media. *International Journal of Information Management*. 2014;34(2):311-4. doi: https://doi.org/10.1016/j.ijinfomgt.2014.01.006.

3.   (CNNIC) CINIC. The 52th China Statistical Report on Internet Development. 2023.

4.   Mosadeghrad AM. Factors Affecting Medical Service Quality. *Iran J Public Health*. 2014;43(2):210-20. Epub 2014/02/01. PubMed PMID: 26060745; PubMed Central PMCID: PMCPMC4450689.

5.   Lu W, Zhai Y. Self-Adaptive Telemedicine Specialist Recommendation Considering Specialist Activity and Patient Feedback. *Int J Environ Res Public Health*. 2022;19(9). Epub 2022/05/15. doi: 10.3390/ijerph19095594. PubMed PMID: 35564988; PubMed Central PMCID: PMCPMC9101090.

6.   Mullenbach J, Wiegreffe S, Duke J, et al., editors. Explainable Prediction of Medical Codes from Clinical Text2018 June; New Orleans, Louisiana: Association for Computational Linguistics.

7.   Li F, Yu H. ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. *Proc AAAI Conf Artif Intell*. 2020;34(5):8180-7. Epub 2020/02/01. doi: 10.1609/aaai.v34i05.6331. PubMed PMID: 34322282; PubMed Central PMCID: PMCPMC8315310.

8.   Wang J, Zhang G, Wang W, et al. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. *Journal of Cloud Computing*. 2021;10(1):4. doi: 10.1186/s13677-020-00218-2.

9.   Liu H, He J, Wang T, et al. Combining user preferences and user opinions for accurate recommendation. *Electronic Commerce Research and Applications*. 2013;12(1):14-23. doi: https://doi.org/10.1016/j.elerap.2012.05.002.

10.   Sharma D, Singh Aujla G, Bajaj R. Evolution from ancient medication to human-centered Healthcare 4.0: A review on health care recommender systems. *International Journal of Communication Systems*. 2019;n/a(n/a):e4058. doi: https://doi.org/10.1002/dac.4058.

11.   Ju C, Zhang S. Doctor Recommendation Model Based on Ontology Characteristics and Disease Text Mining Perspective. *Biomed Res Int*. 2021;2021:7431199. Epub 2021/08/25. doi: 10.1155/2021/7431199. PubMed PMID: 34426788; PubMed Central PMCID: PMCPMC8379386.

12.   Liu J, Li C, Huang Y, et al. An intelligent medical guidance and recommendation model driven by patient-physician communication data. *Front Public Health*. 2023;11:1098206. Epub 2023/02/14. doi: 10.3389/fpubh.2023.1098206. PubMed PMID: 36778565; PubMed Central PMCID: PMCPMC9909411.

13.   Bayeh Y, Tsehay CT, Negash WD. Health system responsiveness and associated factors for delivery care in public health facilities, Dessie City Administration, South Wollo zone, Ethiopia: Cross-sectional study design. *BMJ Open*. 2023;13(7):e069655. Epub 2023/07/22. doi: 10.1136/bmjopen-2022-069655. PubMed PMID: 37479512; PubMed Central PMCID: PMCPMC10364148.

14.   Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019. p. 4171-86.

15.   Liu X, Esser D, Wagstaff B, et al. Capsule robot pose and mechanism state detection in ultrasound using attention-based hierarchical deep learning. *Sci Rep*. 2022;12(1):21130. Epub 2022/12/09. doi: 10.1038/s41598-022-25572-w. PubMed PMID: 36476715; PubMed Central PMCID: PMCPMC9729303.

16.   Mitchell JR, Szepietowski P, Howard R, et al. A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study. *J Med Internet Res*. 2022;24(3):e27210. Epub 2022/03/24. doi: 10.2196/27210. PubMed PMID: 35319481; PubMed Central PMCID: PMCPMC8987958.

17.   Chen W, Zhong C, Peng J, et al. DxFormer: a decoupled automatic diagnostic system based on decoder-encoder transformer with dense symptom representations. *Bioinformatics*. 2023;39(1). Epub 2022/11/22. doi: 10.1093/bioinformatics/btac744. PubMed PMID: 36409016; PubMed Central PMCID: PMCPMC9825744.

18.   Arnaud E, Elbattah M, Gignon M, et al. Learning Embeddings from Free-text Triage Notes using Pretrained Transformer Models. HEALTHINF: PROCEEDINGS OF THE 15TH INTERNATIONAL JOINT CONFERENCE ON BIOMEDICAL ENGINEERING SYSTEMS AND TECHNOLOGIES - VOL 5: HEALTHINF; 2022-02-09; Lisbonne, Portugal2022. p. 835-41.

19.   Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks2019. 3973-83 p.

20.   Alksher MA, Azman A, Yaakob R, et al. A review of methods for mining idea from text. 2016 THIRD INTERNATIONAL CONFERENCE ON INFORMATION RETRIEVAL AND KNOWLEDGE MANAGEMENT (CAMP)2016. p. 88-93.

21.   Abduljabbar DA, Omar N. Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination. *Journal of theoretical and applied information technology*. 2015;78:447-55.

22.   Neugebauer J, Tothova V, Dolezalova J. Use of Standardized and Non-Standardized Tools for Measuring the Risk of Falls and Independence in Clinical Practice. *Int J Environ Res Public Health*. 2021;18(6). Epub 2021/04/04. doi: 10.3390/ijerph18063226. PubMed PMID: 33804715; PubMed Central PMCID: PMCPMC8004039.

23.   Scherer R, Rao R. Non-Manual Control Devices: Direct Brain-Computer Interaction. In: Pereira J, editor. Handbook of Research on Personal Autonomy Technologies and Disability Informatics. Hershey, PA, USA: IGI Global; 2011. p. 233-50.

24.   Xu S, Leng Y, Feng G, et al. A gene pathway enrichment method based on improved TF-IDF algorithm. *Biochem Biophys Rep*. 2023;34:101421. Epub 2023/03/17. doi: 10.1016/j.bbrep.2023.101421. PubMed PMID: 36923007; PubMed Central PMCID: PMCPMC10009669.

25.   Jones KS. A statistical interpretation of term specificity and its application in retrieval. Document retrieval systems: Taylor Graham Publishing; 1988. p. 132–42.

26.   Li C, Liu J, Qian G, et al. Double chain system for online and offline medical data sharing via private and consortium blockchain: A system design study. *Frontiers in public health*. 2022;10:1012202-. doi: 10.3389/fpubh.2022.1012202. PubMed PMID: 36304235.

27.   Fay MP, Malinovsky Y. Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. *Stat Med*. 2018;37(27):3991-4006. Epub 2018/07/10. doi: 10.1002/sim.7890. PubMed PMID: 29984411; PubMed Central PMCID: PMCPMC6377068.

28.   Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: Eric PX, Tony J, editors. Proceedings of the 31st International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR; 2014. p. 1188--96.

29.   Sexton V, Dale J, Bryce C, et al. Service use, clinical outcomes and user experience associated with urgent care services that use telephone-based digital triage: a systematic review. *BMJ open*. 2022;12(1):e051569-e. doi: 10.1136/bmjopen-2021-051569. PubMed PMID: 34980613.

30.   Liu T-T, Cheng C-T, Hsu C-P, et al. Validation of a five-level triage system in pediatric trauma and the effectiveness of triage nurse modification: A multi-center cohort analysis. *Front Med (Lausanne)*. 2022;9:947501-. doi: 10.3389/fmed.2022.947501. PubMed PMID: 36388924.

31.   Zhang L, Wang X, Xiao H, et al. Governance mechanisms for chronic disease diagnosis and treatment systems in the post-pandemic era. *Frontiers in public health*. 2022;10:1023022-. doi: 10.3389/fpubh.2022.1023022. PubMed PMID: 36582374.

32.   Fekonja U, Strnad M, Fekonja Z. Association between triage nurses' job satisfaction and professional capability: Results of a mixed-method study. *J Nurs Manag*. 2022;30(8):4364-77. Epub 2022/10/19. doi: 10.1111/jonm.13860. PubMed PMID: 36206481.

33.   Asudani DS, Nagwani NK, Singh P. Impact of word embedding models on text analytics in deep learning environment: a review. *Artif Intell Rev*. 2023:1-81. doi: 10.1007/s10462-023-10419-1. PubMed PMID: 36844886.

34.   Kinariwala S, Deshmukh S. Short text topic modelling using local and global word-context semantic correlation. *Multimed Tools Appl*. 2023:1-23. doi: 10.1007/s11042-023-14352-x. PubMed PMID: 36747894.

35.   Argote L. Knowledge Transfer Within Organizations: Mechanisms, Motivation, and Consideration. *Annu Rev Psychol*. 2023. Epub 2023/10/04. doi: 10.1146/annurev-psych-022123-105424. PubMed PMID: 37788572.

36.   Wang YM, Sun Y, Wang B, et al. Transfer learning for clustering single-cell RNA-seq data crossing-species and batch, case on uterine fibroids. *Brief Bioinform*. 2023;25(1). Epub 2023/11/22. doi: 10.1093/bib/bbad426. PubMed PMID: 37991248; PubMed Central PMCID: PMCPMC10664408.

37.   Kuperwajs I, Schutt HH, Ma WJ. Using deep neural networks as a guide for modeling human planning. *Sci Rep*. 2023;13(1):20269. Epub 2023/11/21. doi: 10.1038/s41598-023-46850-1. PubMed PMID: 37985896; PubMed Central PMCID: PMCPMC10662256.
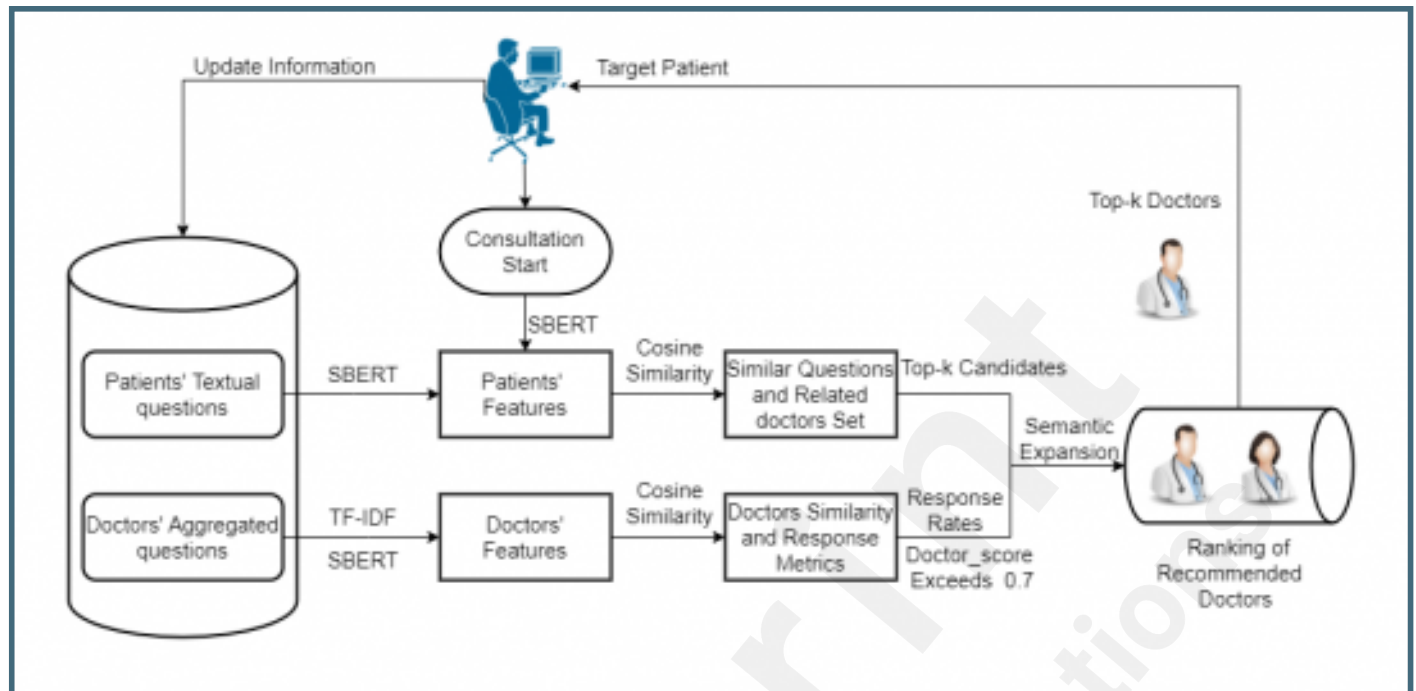
Supplementary Files

Figures

SBERT architecture.
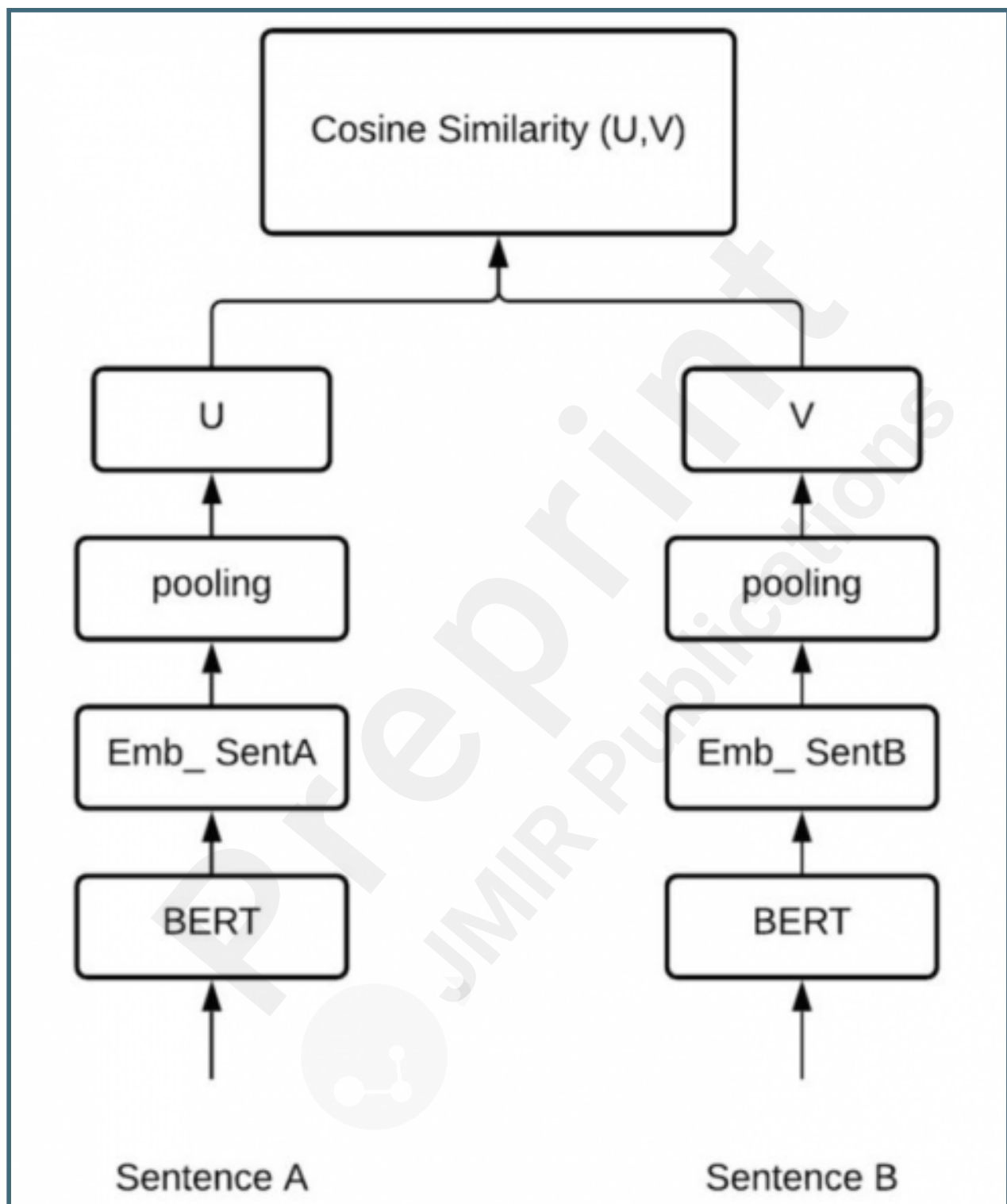
the architecture of PDHR model.

Comparison of our proposed PDHR model with PFB model and PDH model in terms of various indexes, including precision (blue), recall (green), F1-score (red) results, and hqos_prop (orange).