SINA-BERT: A Pre-Trained Language Model for Analysis of Medical Texts in Persian

Anonymous ACL submission

Abstract

We have released SINA-BERT, a language model pre-trained on BERT (Devlin et al., 2018) to address the lack of a high-quality Persian language model in the medical domain. SINA-BERT utilizes pre-training on a large-scale corpus of medical contents including formal and informal texts collected from various online resources in order to improve the performance on health-care related tasks. We employ SINA-BERT to complete following representative tasks: categorization of medical questions, medical sentiment analysis, medical named entity recognition, and medical question retrieval. For each task, we have developed Persian annotated data sets for training and evaluation and learnt a representation for the data of each task especially complex and long medical questions. With the same architecture being used in each task, SINA-BERT outperforms BERT-based models that were previously made available in the Persian language.

1 Introduction

Patients, physicians and healthcare professionals are generating textual information every day using diverse formats that can be found in online resources. To improve the diagnosis and treatment of disease, text mining techniques are becoming increasingly important. Developing computational models of disease and applying these models to massive collections of textual information are significant challenges of computational medicine (Rakocevic et al., 2013).

Text mining methods have been considered in multiple research studies in medicine; the most important ones being Named Entity Recognition (NER), personal data anonymization, knowledge discovery (Bokharaeian et al., 2017), and terminology extraction (Luque et al., 2019). By employing text mining techniques, several healthcare systems can be developed such as Question Answering Systems (Ozyurt et al., 2020) and medical specialized search engines (Luo et al., 2008).

Recent progress in medical text mining is due to advancements in the deep learning techniques used for natural language processing (NLP). In particular, language models have shown remarkable advances in most NLP tasks and many current state-of-the-art methods often rely on Transformer-based pre-trained language models (Devlin et al., 2018; Radford et al., 2018).

While there are several BERT-based language models for the medical domain in English (Lee et al., 2020; Beltagy et al., 2019; Rasmy et al., 2020), Persian lacks such resources. In this paper, we present SINA-BERT, a pre-trained language representation model for the Persian biomedical domain. First, we initialize SINA-BERT with weights from ParsBERT (Farahani et al., 2020), which is a public domain Persian language model. Then, SINA-BERT is pre-trained on large Persian medical corpora, collected from medical and health related websites, journals, books, forums and news websites. These corpora contain 2.8M documents from both formal and informal texts.

To show our language model's effectiveness in medical text mining, SINA-BERT has been fine-tuned and evaluated on the following popular medical text mining tasks: question classification, sentiment analysis, NER, and question retrieval. We also provide an annotated data set for each task and compare the performance of SINA-BERT against the state-of-the-art models. Therefore, the contributions of this paper can be listed as follows:

- A large scale Persian medical corpus.
- A pre-trained language model for the Persian medical domain.
- A database containing 200k Persian medical questions answered by professional physicians for the task of question retrieval.
- Annotated data sets for tasks of medical sentiment analysis, medical NER, and categorizing medical questions in Persian.

- Three data sets for automatic evaluation of medical retrieval systems in Persian.
- Learning a representation for medical complex and long questions based on deep sentence representation and ranking.

This work opens up avenues for further investigation into Persian medical text analysis. The rest of this paper is organized as follows: Section 2 briefly reviews BERT-based language models in the medical domain. Then the procedure of pre-training SINA-BERT is presented in Section 3. After that, the evaluation results of SINA-BERT on downstream tasks are explained in Section 4. Finally, concluding remarks are given in Section 5.

2 Background

We have reviewed biomedical word embeddings and the BERT-based language models of medical-related domains below, as well as the Persian's pre-trained language models.

2.1 Language Models for Medical Domain

BioBERT (Lee et al., 2020) is a domain-specific language model which was initialized by BERT and pre-trained on a large-scale biomedical corpus containing PubMed abstracts and PubMed full-text articles. BioBERT is fine-tuned for three biomedical text mining tasks: NER, relation extraction, and question answering.

SCIBERT (Beltagy et al., 2019) is a language model which was pre-trained on a sizeable multi-domain corpus of scientific publications. This corpus contains 1.14M papers randomly selected from Semantic Scholar. SCIBERT was evaluated on sequence tagging, sentence classification, and dependency parsing; all with data sets from various scientific domains.

Clinical BERT (Alsentzer et al., 2019) was pre-trained on a corpus of approximately 2 million clinical notes. It improved the performance of clinical NLP tasks such as extracting Protected Health Information (PHI) during the process of anonymising medical records for de-identification.

BEHR (Li et al., 2020) is a Transformer-based deep neural sequence transduction model for electronic health records (EHR). The aim of training BEHR is to use a given patient's past EHR to predict his/her future diagnoses (if any). This model was trained and evaluated on the data from nearly 1.6 million individuals.

MedBERT (Rasmy et al., 2020) is another language model which was pre-trained on large-scale structured EHRs to benefit downstream disease-prediction tasks. This model was fine-tuned for the prediction of heart failure in patients with diabetes and the prediction of pancreatic cancer.

HQADeepHelper (Luo et al., 2020) is a deep learning system that includes a wide range of healthcare question answering models; most of which are based on the pre-trained BERT or SCIBERT.

BioWordVec (Zhang et al., 2019) is an open set of biomedical word embeddings that combines subword information from unlabeled biomedical text with a widely-used biomedical vocabulary.

2.2 Language Models in Persian

Two multi-lingual language models support Persian: multi-lingual BERT (Devlin et al., 2018) and XLM-RoBERTa (Conneau et al., 2020) to the best of our knowledge. However, the size and the domain of the corpora used by them are not apparent.

ParsBERT (Farahani et al., 2020) is a monolingual BERT for the Persian language, which was pre-trained on a general domain corpus of 2.8M documents. ParsBERT was evaluated on NER and sentiment analysis tasks. The domains of the data sets used in these evaluations are news and online shopping respectively.

3 Approach

Persian is among the under-resourced languages. Although there are language models that support Persian, none of them were pre-trained on a large Persian medical corpus. Understanding medical texts and solving medical tasks like question answering attract many researchers. However, the lack of a high-performance language model in this domain is a severe obstacle for them. In this section, we describe our Persian medical corpus and the details of pre-training SINA-BERT.

3.1 Data Collection

Although there are plenty of online Persian texts related to health and medicine, no large corpus is available. So, to train a medical language model in Persian, we had to gather together a large collection of texts from several online sources. The topic of these texts includes health, medicine, nursing, pharmacy, medical ethics and law, folk medicine, Persian medicine, lifestyle, nutrition, etc. This corpus contains 2.8M documents which were collected

from the following sources:

- health and medical news websites
- web sites publishing scientific materials about health, nutrition, lifestyle, etc.
- journals (abstract and full papers) and conference proceedings
- academic written materials
- medical reference books and theses
- online health-related forums
- medical and health-related pages of Instagram
- medical channels and groups of Telegram

The collected documents are then normalized and cleaned so they are free of HTML tags, hyperlinks, CSS, javascript, etc.

Normalization is an essential pre-processing task in Persian because, unlike English, some Persian letters can be written in different forms with different ASCII codes. We have developed a new normalizer module in which mapping into a standard character is provided for all of the characters that appear in the corpus. Wired characters are mapped into empty characters, which means they are removed.

3.2 Pre-Training SINA-BERT

SINA-BERT is based on the $\text{BERT}_{\text{BASE}}$ model architecture (Devlin et al., 2018) which includes 12 hidden layers, 12 attention heads, and 768 hidden sizes. The total number of parameters of this configuration is 110M. The initialization of parameters is taken from ParsBERT (Farahani et al., 2020) which is a public domain BERT-base model in Persian. The tokenizer of ParsBERT is also borrowed. As per the original BERT and ParsBERT, the pre-training objective is the Masked Language Model (MLM), in which 15% of tokens are randomly masked. The training batch size is 6, the learning rate is 5e-7, and each sequence contains 512 tokens at most.

4 Validation on Medical Tasks

We validated SINA-BERT on five tasks. Since the lack of data sets for these tasks in Persian, we prepared annotated data for each task. These resources have been used in the evaluation of SINA-BERT and could be employed in further studies on Persian medical IR and QA tasks. In each task, SINA-BERT's performance is compared with the below state-of-the-are language models already available in Persian:

- BERT-Base, Multi-lingual Cased (mBERT) (Devlin et al., 2018) which is a multi-lingual language model that supports 102 languages including Persian.
- XLM-RoBERTa (Conneau et al., 2020) which is pre-trained for one hundred languages including Persian.
- ParsBERT (Farahani et al., 2020) which is the base model of SINA-BERT.

In contrast to SINA-BERT, the above language models were pre-trained on general domain data.

Table 1: Accuracy of the Persian language models applied to the fill-in-the-blank task.

Model	Accuracy
XLM-RoBERTa	12.83
mBERT	13.88
ParsBERT	39.44
SINA-BERT	**50.71**

4.1 Fill-in-the-Blank

The first task was fill-in-the-blank. We searched through a famous Persian website, Niniban[1], which is an online magazine. There are several forums on this site in which people discuss all medical and health-related matters, ask their questions and answer other people's questions. While the tone of the magazine's writing is completely formal, forums are mostly informal. Among all the materials on this website, 10,000 random sentences were selected. 15% of the tokens in each sentence were then masked randomly. The Persian language model was used to predict the masked tokens. This data set was excluded from the corpus we used to pre-train SINA-BERT, so we considered the exact matching of the masked token with a predicted word to be true. Therefore, we consider the number of true cases divided by the total number of masked tokens to be an indication of the model's accuracy. Table 1 shows that SINA-BERT significantly outperforms other models. Also, ParsBERT is the second-best model because it was pre-trained with a larger Persian corpus in comparison with mBERT and XLM-RoBERTa.

4.2 Medical Question Classification

Medical Question Answering (MQA) systems have gained considerable attention. Question Classification (QC) is a major task within these systems because MQA systems may be designed to answer only some specific kinds of medical questions.

[1] http://niniban.com/

Table 2: Accuracy of the language models evaluated on the question classification task.

Model	Prec.	Rec.	Macro F_1	Accu.
mBERT	88.41	87.41	87.89	90.80
XLM-RoBERTa	90.61	88.78	89.65	92.50
fastText + CNN	90.90	91.30	91.10	93.52
ParsBERT	93.01	93.13	93.07	94.66
SINA-BERT	**94.91**	**94.63**	**94.77**	**96.14**

Questions can be classified with consideration of different aspects such as anatomy, disease causes, treatment type, etc. (Roberts et al., 2014, 2016). A common classification is based on the type of doctor that should respond to that question.

To prepare a data set to validate SINA-BERT on this task, we used the QA data set we collected for the task of question retrieval, which will be explained in Section 4.5. Each QA has a meta-data that denotes the category of the question and the specialty of the doctors who answered that question. We selected “pediatric gastroenterology" as one of the most frequent categories in the database. Among the QA from this category, 1000 random samples were selected and labeled “1". Also, 3400 random samples were selected from other categories and labeled “0". To ensure that the automatic labels were correct, all samples were manually checked by two annotators. As a result, a data set containing 4400 QA was prepared.

Using our data set, we ran a binary classifier. The [CLS] token of the last layer was fed into a linear classification layer. A dropout of 0.1 was applied and cross-entropy loss was optimized using Adam (Kingma and Ba, 2014). The model was fine-tuned for 10 epochs using a batch size of 8 and a learning rate of 2e-5.

Table 2 shows the results of applying BERT-based models to the task of identifying questions related to pediatric gastroenterology. In addition to the BERT-based models, a Convolution Neural Network (CNN) was implemented which uses fastText (Bojanowski et al., 2017) word embedding as the initialization of the Embedding layer. This embedding was trained on the same corpus that was used for the pre-training of SINA-BERT. According to the macro F_1 and accuracy measures, SINA-BERT outperforms other language models. The results obtained by this experiment confirm that SINA-BERT surpasses other Persian language models in understanding the content of medical questions.

4.3 Medical Sentiment Analysis

People often interact with other users with similar health conditions on social networks and health forums and share their experiences about doctors, drugs, treatments, or diagnosis. Therefore, sentiment analysis in medical setting (Yadav et al., 2018; Denecke and Deng, 2015) has attracted much attention in recent years.

To assess patients’ satisfaction with their physician’s performance, a data set containing 5,000 comments was collected from Persian online medical counseling websites. This data is mostly comments from people on the quality of the counsel they received from online doctors. They were manually labeled with Satisfaction (1500 comments), Un-satisfaction (1202 comments), and No-idea (2298 comments), so we defined a 3-classes classification task for this data set. From this set of comments, 5% were used for testing, 10% for validation, and the rest for the training.

To perform the evaluation, the embedding vectors of the comments generated by SINA-BERT and other base models were given to a CNN classifier. This classifier, which consists of 100 filters of different sizes [2, 3, 4, 5, 6] along with the max-pooling layer, predicts the label of each comment based on the given embedding vectors. These hyper-parameters are tuned using the validation set and the model with the best accuracy was selected. Moreover, Adam optimizer with a learning rate of 2e-5, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used with batch size 16. The training was performed for 3 epochs.

The results of the sentiment analysis based on SINA-BERT and other basic models are shown in Table 3. Due to the randomness of the initial weights, each model was executed 5 times, and the average of them was reported. As can be seen, SINA-BERT has a higher performance compared to multi-lingual models such as mBERT and XLM-RoBERTa. In the case of ParsBERT, its performance was close to SINA-BERT due to the fact that medical terminologies are normally less commonly used in the comments of users. For example, many people just said “that was good", “this doctor is not so good", “last prescription didn’t work for me at all", etc., which means most of the comments were short and simple and lacked professional vocabulary.

Table 3: Accuracy of the language models evaluated on the medical sentiment analysis.

Model	Prec.	Rec.	Macro F_1	Accu.
mBERT	0.91	0.90	90.06	0.90
fastText + CNN	0.91	0.91	90.06	0.91
XLM-RoBERTa	0.92	0.92	91.62	0.92
ParsBERT	0.93	0.93	92.82	0.93
SINA-BERT	**0.95**	**0.94**	**94.49**	**0.94**

Table 4: Description of Persian medical NER data set.

Data Set	Tokens	Entity Counts			
		disease	drug	symptom	treatment
Training	56,860	712	1946	1448	199
Test	14,899	188	531	375	52
Validation	6,213	89	221	141	14
Total	77,972	989	2698	1964	265

Table 5: F_1 scores of the language models evaluated on the medical NER.

Model	Named Entities					
	disease	drug	symptom	treatment	micro F_1	accu.
fastText	65.90	74.26	60.37	30.77	65.52	75.22
XLM-RoBERTa	**86.00**	88.39	67.63	25.64	82.91	86.61
mBERT	84.34	92.80	73.74	**81.18**	**86.62**	88.68
ParsBERT	80.31	**92.98**	**74.93**	71.79	86.21	89.97
SINA-BERT	83.46	92.21	74.29	77.33	86.27	**90.87**

4.4 Named Entity Recognition

Medical NER systems are developed to extract information such as drugs, diseases, and pathogens from a text. There are some annotated corpora with NER tags in Persian (Taghizadeh et al., 2020; Poostchi et al., 2016); however, none of them includes medical entities to the best of our knowledge.

To create a NER data set in the medical domain, we randomly selected 500 questions related to pediatric gastroenterology from the data set which was prepared in Section 4.2. These questions were annotated with four entities: *disease*, *symptom*, *treatment*, and *drug*. Treatment refers to all kinds of actions caring for patients to combat disease except for the use of prescription drugs and medical tests. Each question was annotated by two annotators and the agreement between them was about 92% based on the tagged phrases.

The annotated data set was randomized and split into 70% for training, 20% for testing, and 10% for validation. Table 4 presents the statistics of the annotated data. It tabulates the counts of tokens as well entity-wise counts.

To create a Persian medical NER model, we adopted the network architecture of Beheshti-NER (Taher et al., 2020). In this model, the data is tokenized and given to SINA-BERT. Each sequence had 512 tokens at most. Then the representation of a sentence which was obtained from SINA-BERT was given to a fully-connected layer followed by a Conditional Random Field (CRF) layer. We added a dropout layer having a probability of 0.1 before the fully connected layer.

The parameter setting for the NER model was as follows: batch size was 8; AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 3e-6 was employed; the loss function is negative log-likelihood; the number of epochs was 15, and size of the fully connected layer was 786.

Table 5 presents the detailed results for all named entities. The evaluation was performed at the level of words. Among four classes, *drug* and *disease* obtained the highest scores. Because the name of drugs and diseases are somewhat independent of the context and can be specified by using the gazetteers. In contrast, symptoms and treatments are highly dependent on the context and so, the performance of models on these classes was lower than *drug* and *disease*. Table 5 shows that SINA-BERT outperformed the other models in terms of accuracy and reached to the state-of-the-art results based on micro F_1 score.

4.5 Medical Question Retrieval

A growing number of people including patients, doctors and healthcare professionals utilize Information Retrieval (IR) systems to seek answers to their questions. These questions vary from definitional questions, i.e., “What is X?", to complex questions pertinent to a patient’s illness such as how to assess symptoms in order to seek medical help and diagnosis (Cao et al., 2011).

In the task of question retrieval, a list of Question-Answer (QA) pairs are retrieved from a database of QA which are the most similar to the user’s question. This retrieval system supports decision-making for diagnosis and treatment. We collected a set of 200K medical QA pairs. They were gathered from 20 Persian websites that provide online services for medical consultancy. Each question of this database has been already answered by at least one physician. These QA pairs are cleaned and normalized. Analysis of these medical questions shows that they vary in length from a short sentence to one or more paragraphs, as well as vary in tone from professional to personal and

emotional. The average and standard deviation of question length are 69.0 and 78.3 tokens respectively.

Most of the retrieval models take pre-trained representations and either 1) obtain a document representation from individual word representations that is subsequently used for ranking, or 2) combine representation similarities in some way to rank documents (Gysel et al., 2018). A common method of generating question representation from word representation is to average the word representations. However, this basic representation can be improved. Therefore, we propose the following representations:

- SINA-BERT_all: The average of all embeddings in the last layer of the network.
- SINA-BERT_rsw: It is similar to SINA-BERT_all; but the stop-words are removed from the average pooling.
- SINA-BERT_kw: Instead of giving the complete question to the network, n most important key phrases are selected together with two words before and after them as the context for key phrases. The enhanced key phrases are separated with [SEP] tokens, and this sequence is then given to the model. Key phrases are selected based on TF-IDF score.
- SINA-BERT_kw_rcnt: It is similar to SINA-BERT_kw; but two words after and before the key phrases are ignored just before the average pooling and only the embedding of key phrases are considered.

Therefore, we adopted an unsupervised approach toward ranking documents as follows: Given a user's query, the representation of this query is obtained and the similarity of it to all the questions of the database is calculated using their cosine similarity. The topmost similar ones are retrieved and presented to the user. In the next sections, we compare SINA-BERT with the current state-of-the-art models and report their scores.

4.5.1 User-Oriented Evaluation

In the first evaluation, 70 QA pairs of the database were selected randomly and separated from other QA pairs. These QA pair were supposed to be the user's queries that were given to the retrieval system. In response to each user's query, the most similar QA pair of the database (top one) was judged by a human. There is a multiple-choice format for the judgment:

- Similar questions: two patients had similar conditions and their request was the same or very similar.
- Similar topics: two questions had similar topics; however they were not the same.
- Different topics: two questions had different topics.

These three options received scores of 1, 0.5, and 0 respectively. However, in a rigid evaluation, only the first case got a score of 1 and the others got 0. The accuracy was then calculated based on these scores. These human judgments were double checked to be fair across different language models.

Table 6: Accuracy of the language models on the task of medical question retrieval evaluated by human judgment.

Model	Accuracy (01)	Accuracy
XLM-RoBERTa	18.57	30.00
mBERT	18.57	31.42
ParsBERT	25.71	32.86
SINA-BERT_all	30.00	41.43
SINA-BERT_kw	35.71	42.14
SINA-BERT_rsw	35.71	45.00
SINA-BERT_kw_rcnt	**35.71**	**47.85**

Table 6 presents the scores obtained by SINA-BERT and other language models. All versions of SINA-BERT significantly outperform other BERT-based models. Comparing four methods for producing document representation from word embeddings reveals that extracting keywords and removing stop words improves the accuracy of SINA-BERT_all by 1.7% and 8.6%, respectively. Adding the contexts into the keywords before feeding them into the model and removing the context word embeddings during the average pooling result the most improvement over SINA-BERT_all, i.e. about 15.5%, and result in the highest scores. This means that context words are necessary to produce the meaningful embedding of keywords; however, average of keyword's embedding is sufficient to build sentence representation.

4.5.2 Paraphrased Test Data

In the second evaluation, 200 QA pairs were selected from the database at random. These QAs were divided into four parts and each part was given to a human native in Persian with an academic degree to read the question carefully and produce a paraphrase for it. The guideline was to "*rewrite the question by changing the writing style, words, tone*

Table 7: Performance of different methods of question representation on single-stage retrieval task using paraphrased test data.

Model	R@1	R@5	R@10
XLM-RoBERTa	28.19	35.10	37.76
mBERT	27.65	34.57	40.42
ParsBERT	31.38	38.29	40.95
SINA-BERT_all	36.17	43.08	47.34
SINA-BERT_kw	40.42	45.74	48.93
SINA-BERT_rsw	42.02	50.00	54.78
SINA-BERT_kw_rcnt	**44.14**	**53.19**	**55.31**

Table 8: R@1 of the retrieval models applied to the noisy queries.

Model	Noise Percentage				
	0.1	0.2	0.3	0.4	0.5
XLM-RoBERTa	86.1	25.0	4.9	0.9	0.3
mBERT	97.8	83.2	34.6	9.3	1.3
ParsBERT	98.4	79.8	28.8	6.9	1.3
SINA-BERT_rsw	48.1	32.1	13.5	4.3	0.8
SINA-BERT_kw	93.7	48.1	9.8	1.7	0.3
SINA-BERT_all	**99.2**	86.8	28.2	5.3	0.5
SINA-BERT_kw_rcnt	99.1	**97.1**	**85.1**	**57.9**	**23.5**

of the text, etc. at most at possible until no change in the meaning".

Each paraphrased question was a query given to the retrieval system, and therefore the prime question is expected to be retrieved. To measure the performance of a retrieval system, we used the R@k metric, so we retrieved top k questions (k= 1, 5, and 10), and checked if the prime question was among the retrieved questions.

Table 7 presents the comparison of different language models. The overall scores are similar to Table 6, and SINA-BERT_all outperforms all the state-of-the-art language models. Among the proposed methods for filtering tokens from the average pooling, SINA-BERT_kw_rcnt shows the most improvement of R@10; obtaining 16.8% higher than the SINA-BERT_all.

4.5.3 Noisy Queries

In the third evaluation, 1000 QAs from the database were selected randomly. In each question, m percent of tokens were replaced with random tokens from the vocabulary. m varies from 0.1, 0.2, to 0.5. This noisy data set is given to all methods. It is expected that the prime question is retrieved when the noisy question is the query. So, we evaluated the retrieval methods by using the R@1 metric. As Table 8 demonstrates, the highest scores are obtained by SINA-BERT_kw_rcnt. This method outperforms all systems by a substantial margin; especially for higher noise percentages.

Table 9: Comparison of different retrieval methods on the paraphrased data set.

Model	R@1	R@5	R@10	MRR
UKP-DistilBERT	36.36	49.73	54.54	43.73
TF-IDF	50.00	62.23	66.47	56.66
UKP-XLMR-paraph	50.26	63.10	69.51	57.00
SINA-BERT	**68.87**	**75.51**	**76.53**	**69.95**

4.5.4 Comparing with Text Mining Methods

In the last evaluation, different methods of document presentation were compared by using an unsupervised re-ranking approach: Firstly, an initial list of documents is retrieved by a simple and fast unsupervised bag-of-words method, e.g. BM25 (Robertson et al., 1995), which are then re-ranked by the BERT-based models that produce the document representation from the word representation.

In addition to SINA-BERT_kw_rcnt, we employed UKP-DistilBERT (Reimers and Gurevych, 2020) and UKP-XLMR-paraph (Reimers and Gurevych, 2020) which are two multi-lingual sentence embeddings. The training of these models is based on the idea that a translated sentence should be mapped to the same point in the vector space as the original sentence. Therefore a mono-lingual model, e.g. mBERT, is used to generate sentence embeddings for the source language and then train a new system on translated sentences to mimic the original model. These models are available in more than 50 languages including Persian. The similarity of two questions is computed based on the cosine similarity of sentence embedding of the two questions.

The data set used in this experiment is the paraphrased data as was described in Section 4.5.2. Table 9 represents the recall scores obtained by different sentence representation methods. For a better comparison, we report the scores of the bag-of-word model of TF-IDF. The re-ranking method, which is based on SINA-BERT_kw_rcnt, significantly outperforms UKP-DistilBERT and UKP-XLMR-paraph. TF-IDF obtains a higher recall in comparison with the sentence representation method of UKP-DistilBERT. Although this is contrary to our expectations, the main reason for this result is that two paraphrased medical questions have many common keywords such as the names of drugs, names of diseases, names of medical treat-

Table 10: An example query with the best retrieved question. The last row shows the manual judgment.

Query: پسر چهار ماهه ام که شیر خشک آبتامیل مصرف میکند یک ماه هست خوب شیر نمیخورد .شیشه شیر را هی پس میزند در صورتی که قبلا خیلی بیش از حد نیز شیر میخورد. چیکار کنم که دوباره خوب شیر بخورد؟ My 4 month old son, who was drinking Aptamil powdered milk, refuses the bottle of milk, while he was already drinking too much. What can I do?		
SINA-BERT	**ParsBERT**	**UKP-Distil-BERT**
دختر 3ماه من قبل از واکسن دوماهگی خوب شیر می خورد اما بعد از ان در ساعات بیداری شیر نمیخوره، مجبور شدم بهش شیر خشک بدم ولی شیر خشکم نمیخوره، فقط وقتی که خوابه می تونم بهش شیر بدم لطفا راهنمایی کنید	دخترم هشت و نیم ماهشه. شیر بیومیل ای آر میخوره. از دیشب چندین بار استفراغ داشته و هر چی میخوره سریع بالا میاره. اسهال هم داره ولی فعلا کم. غذا نخوره چند روز فقط شیر بخوره بهتر هست؟	سلام دخترم 1سال و 8 ماه شه یک ماهی میشه به جز شیر هیچ چیزه دیگه ای نمی خوره قبلا هم که روزی سه الی 4 شیشه شیر می خورد الان یک الی دو شیشه شیر بیشتر نمی خوره شربت اشتها اور هم بهش دادم (زینگ سولفات) اما تاثیری نداشته لطفا راهنمایی کنید اگه داروی هم هست بهم بگید
My 3-month-old daughter was breastfeeding well before the two-month vaccination, but after that, she does not breastfeed during waking hours. I had to give her powdered milk but she does not drink powdered milk, I can only breastfeed her when she is asleep, please help	My daughter is eight and a half months old. She Eats Biomil ER milk. She has vomited several times since last night and everything she eats she throwing up quickly. She also has diarrhea, but not much at the moment. Not eating food for some days, just drinking milk, is this better?	Hi. My daughter who is 1 year and 8 months, has not eaten anything except milk for a month. Already, she used to drink three to 4 glasses of milk a day. Now she does not drink more than one or two bottles of milk. I also gave her an appetizing syrup (zinc sulfate) but it did not work. Please help me if there is any medicine.
Similar question (1)	Different-Topic (0)	Similar-topic (0.5)

ment, etc. The medical domain is a named entity-rich area that changes the scores in the favor of the bag-of-word models when evaluated on the paraphrased test data.

Table 10 shows an example of a Persian medical query and the best-retrieved question by SINA-BERT in comparison with ParsBERT and UKP-Distill-BERT. This query is related to the field of pediatric gastroenterology, in which a mother asks for some advice for her baby who refuses to drink milk. All the retrieved questions also refer to the baby's nutrition. However, deeply considering the retrieved questions reveals that although there are many common words between them, the issue that araises only in the question on the left which was retrieved by SINA-BERT, is similar to the query and the others are relatively different. This example therefore shows that SINA-BERT can be employed in understanding medical documents such as field of pediatric gastroenterology.

5 Conclusion and Future Work

This paper was the first work on developing a medical language model in Persian. A BERT-based language model was pre-trained by collecting a large corpus of both formal and informal Persian texts from online resources. SINA-BERT was validated on five tasks and we have prepared a data set for each one. SINA-BERT outperformed the state-of-the-art Persian language models in all tasks. The margin between it and the other models in the task of question retrieval is much more than in the classification tasks of question classification and sentiment analysis; mainly because the supervision that exists in the classification tasks somewhat closes the gap between SINA-BERT and the other language models. However, for the unsupervised task of question retrieval, the significant differences reveal that training a language model across a large medical data set greatly benefits its understanding of related texts.

As for future works, there is a wide range of tasks in the area of Persian medical text analysis such as information extraction from clinical notes, medical NER, biological relation extraction, medical entity linking, disease prediction, etc., which can be solved using the SINA-BERT; subject to provision of the annotated data sets. Finally, the achievements of this research provide the foundation for further studies of Persian health and medical related tasks.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Behrouz Bokharaeian, Alberto Diaz, Nasrin Taghizadeh, Hamidreza Chitsaz, and Ramyar Chavoshinejad. 2017. SNPPhenA: a corpus for extracting ranked associations of single-nucleotide polymorphisms and phenotypes from literature. *Journal of biomedical semantics*, 8(1):14.

YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. 2011. AskHERMES: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–288.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. ParsBERT: Transformer-based model for Persian language understanding. *arXiv preprint arXiv:2005.12515*.

Christophe Van Gysel, Maarten De Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems (TOIS)*, 36(4):1–25.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: transformer for electronic health records. *Scientific reports*, 10(1):1–12.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Feng Luo, Xiaoli Wang, Qingfeng Wu, Jiaying Liang, Xueliang Qiu, and Zhifeng Bao. 2020. HQADeepHelper: A deep learning system for healthcare question answering. In *Companion Proceedings of the Web Conference 2020*, pages 194–197.

Gang Luo, Chunqiang Tang, Hao Yang, and Xing Wei. 2008. Medsearch: a specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 143–152.

Carmen Luque, José M Luna, Maria Luque, and Sebastian Ventura. 2019. An advanced review on text mining in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1302.

Ibrahim Burak Ozyurt, Anita Bandrowski, and Jeffrey S Grethe. 2020. Bio-answerfinder: a system to find answers to questions from biomedical texts. *Database*, 2020.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. PersoNER: Persian Named Entity Recognition. In *COLING 2016-26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Goran Rakocevic, Tijana Djukic, Nenad Filipovic, and Veljko Milutinović. 2013. *Computational medicine in data mining and modeling*. Springer.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *arXiv preprint arXiv:2005.12833*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014. Automatically classifying question types for consumer health questions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1018. American Medical Informatics Association.

Kirk Roberts, Laritza Rodriguez, Sonya E Shooshan, and Dina Demner-Fushman. 2016. Resource classification for medical questions. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1040. American Medical Informatics Association.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Nasrin Taghizadeh, Zeinab Borhanifard, Melika GolestaniPour, and Heshaam Faili. 2020. NSURL-2019 task 7: Named entity recognition (NER) in farsi. *CoRR*, abs/2003.09029.

Ehsan Taher, Seyed Abbas Hoseini, and Mehrnoush Shamsfard. 2020. Beheshti-ner: Persian named entity recognition using bert. *arXiv preprint arXiv:2003.08875*.

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018. Medical sentiment analysis using social media: towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.