■ RESEARCH ARTICLE

Artificial Intelligence Based Chatbot in E-Health System

Kamil AKARSU [a†] **, Orhan Er** [b]

[a] Department of Computer Engineering, İzmir Bakırçay University, İzmir, Turkey
[†] kamilakarsu94@gmail.com, corresponding author

RECEIVED JUNE 8, 2023
ACCEPTED SEPTEMBER 26, 2023

CITATION Akarsu, K. & Er, O. (2023). Artificial intelligence based chatbot in e-health system. *Artificial Intelligence Theory and Applications, 3(*2), 113-122.

Abstract

The healthcare sector is undergoing a digital revolution due to the rapid growth of technology, and AI technologies are becoming more commonplace in the sector. Chatbots have become useful resources for people to get advice and information about their health issues. The creation and implementation of an AI-based chatbot, integrated with an e-health system, is the main topic of this article. This paper explains the development and creation of chatbots. The chatbot's language comprehension and response capabilities are enhanced through the use of AI techniques such as machine learning and natural language processing (NLP). In addition, the chatbot's user interaction procedure and data security precautions are covered. The paper also examines how the developed chatbot can be integrated into an e-health platform and provides the results of user testing. These evaluations focus on the chatbot's ability to provide accurate and insightful answers, understand user requirements, and provide useful advice. The test results show favourable user evaluations and indicate how well the AI-based chatbot performs in providing healthcare services.

Keywords: chatbot; chatbot in e-health system; sentence matching with deep learning.

1. Introduction

Developed by combining artificial intelligence and natural language processing technology, chatbots are robotic voice systems that can converse with humans in normal language. Using this technology, people can now access healthcare services more quickly and easily, revolutionizing the healthcare sector [1]. The development of chatbots has kept pace with developments in artificial intelligence and computer science. The original chatbots were primarily able to respond to users' simple requests for information, as they were built on a simple question-and-answer basis. However, with the help of deep learning methods and increasingly complex natural language processing algorithms, chatbots have become more sophisticated and intelligent over time.

The use of chatbots in healthcare has the potential to significantly improve patient access to care and increase service productivity. Many healthcare organizations and practitioners are using chatbots in areas such as advice, appointment scheduling, health information dissemination and symptom assessment [2].

There are several advantages to using chatbots in healthcare [3]. Firstly, because chatbots are available 24/7, patients can access healthcare in emergencies or out of

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2023 İzmir Bakırçay University

hours. Secondly, chatbots can protect users' confidential health information, alleviating privacy concerns. Chatbots can also answer common questions and help doctors focus on more complicated cases. However, there are significant challenges to using chatbots in healthcare. In order for chatbots to provide accurate and trustworthy answers, they must first be properly trained and provided with health data. This requires access to large and accurate datasets of high-quality training data. In addition, chatbots cannot always take on the role of human doctors and are not always accurate in identifying important medical issues. As a result, the decision to use chatbots should be carefully considered and implemented under the guidance of real doctors [4].

As a result, the use of chatbots in healthcare has the potential to improve patient access to care, increase healthcare productivity, and reduce the workload of healthcare workers. However, care must be taken to design, train and implement the technology correctly. Chatbots can help patients and healthcare professionals communicate and collaborate more effectively and provide patients with a more personalized and accessible experience with healthcare services [5].

2. Literature Review

By providing users with instant access to support, chatbots are essential in the health sector. They help users make informed decisions by providing personalized and reliable information on a range of medical conditions, symptoms and treatments. By freeing up healthcare professionals to focus on more important activities, chatbots can improve the effectiveness of healthcare systems. They enable people to access healthcare resources and assistance at any time, even in remote locations. By analyzing user input and providing relevant information, chatbots can also help in the early detection of health conditions, potentially leading to better health outcomes.

In a study conducted by M. Huang and colleagues [6] aimed to explain how to chatbot is a digital tool that uses text or voice to simulate human conversation and enable interactive communication. It can be accessed through smartphones or computers. Utilizing a chatbot for the follow-up of cancer patients during treatment can be a beneficial solution that also helps healthcare providers save time.

In their study, R. Joshua and colleagues [7] aimed to evaluate whether a multilingual chatbot could effectively engage patients with limited English proficiency (LEP) and improve their outcomes after total joint arthroplasty (TJA). They recognized that language barriers can make the delivery of perioperative instructions challenging for LEP patients, and thus sought to assess the effectiveness of a multilingual chatbot in overcoming these barriers and enhancing patient outcomes in the context of TJA.

In other related study, J. Montenegro and colleagues [8] aimed to further explore the use of chatbots in assisting pregnant women during the prenatal and postnatal periods in Brazil. Similar to the previous study, a pilot study was conducted using a mixed-method design involving healthcare professionals and pregnant women. The participants interacted with the chatbot for seven days and then completed a survey. The results of this study aligned with the previous findings, indicating that pregnant women perceived the chatbot as educational and believed their physicians would approve of its use. The chatbot's performance expectation received positive ratings, while facilitating conditions had a relatively lower influence. Healthcare providers emphasized the benefits of clear language and comprehensive information for pregnant women. The study reaffirmed the viability and usefulness of the presented chatbot and highlighted the potential for future research to focus on enhancing conversation aspects through Natural Language Processing (NLP) techniques [9]

In a separate study, S. Pandey and colleagues [10] conducted research on the association between increased screen time and its potential health impacts, particularly the detrimental effects on mental health. They employed Deep Learning (DL) and Machine Learning (ML) techniques to examine the influence of technological obsessions on health outcomes. The deployment of chatbots in various industries has proven to be a transformative innovation. The study focused on the development of conversational Artificial Intelligence (AI) systems that enable operators to engage in conversations with machines, resembling human interactions. Two types of chatbots were designed and developed: retrieval-based and generative-based, with each type consisting of six different designs. The accuracy rates for the retrieval-based chatbots varied, with the highest accuracy achieved by the Bidirectional LSTM (Bi-LSTM) [11] at 91.57%. In comparison, the generative-based chatbots, with encoder-decoder designs, exhibited an accuracy rate of 94.45%. A notable distinction between the two types of chatbots is that generative-based chatbots[12] have the capability to generate new text, whereas retrieval-based chatbots are limited to responding based on existing knowledge and outputs.

In a study conducted by Y. Liu and colleagues [13] a comparative analysis was performed to investigate the factors influencing user satisfaction and usage intention in the context of chatbot adoption and development. The study focused on the contrasting experiences of mainland China, where chatbot services have become integrated into daily life, and Hong Kong, where challenges persist in enhancing and promoting chatbot offerings. Through qualitative exploration, critical factors such as perceived quality and privacy concerns were identified. The quantitative findings shed light on the distinct roles of these antecedents across the two regions. Notably, satisfaction was found to positively influence usage intention, with relevance, completeness, pleasure, and assurance emerging as significant factors in both regions.

3. Materials and Methods

3.1. Dataset

In this study, the Instructor Doctor-200k dataset was used to train the deep learning model. The Instructor Doctor-200k dataset, developed by Google AI and Stanford University, is a large collection of human-generated medical questions and answers. It contains more than 200,000 instruction, questions and answers developed by instructors and doctors in the field of medical education [14]. The purpose of this dataset is to develop and test natural language processing models for answering medical questions. The wide range of topics makes it suitable for building models that can answer a wide range of medical questions. Researchers and developers can use the dataset, which can be downloaded from the Google AI Research website, to improve medical question answering systems and increase access to medical information.

There are training and test sets in the Instructor Doctor 200k dataset. While the test set contains only 50,000 instruction, questions and answers, the training set contains 150,000. The questions and answers in the dataset are restricted to being between one and one hundred words long. The dataset also contains medical terms annotated using the Unified Medical Language System (UMLS). Medical software systems often use the UMLS, a complete lexicon of medical terminology. The dataset contains 3 features by structure. These; instruction, input and output [14].

Table 1. Sample data in dataset

Number	**input** (string)	**output** (string)	**instruction** (string)
1	"i had what feels like a muscle cramp about an hour ago under the left bottom rib. it lasted about a minute and then went away. i had no other pains or dificulties since. could this have been a simptom of a minor heart attack. do heart attack symptoms come one at a time or are there more than one symptom when they occur?"	"No this is not a symptom of great attack.... it is normal after some stressful activity. No need to worry. If same thing happen again let me know"	"If you are a doctor, please answer the medical questions based on the patient's description."
2	"My baby has been pooing 5-6 times a day for a week. In the last few days it has increased to 7 and they are very watery with green stringy bits in them. He does not seem unwell i.e no temperature and still eating. He now has a very bad nappy rash from the pooing ...help!"	"Hi... Thank you for consulting in Chat Doctor. It seems your kid is having viral diarrhea. Once it starts it will take 5-7 days to completely get better. Unless the kids having low urine output or very dull or excessively sleepy or blood in motion or green bilious vomiting...you need not worry. There is no need to use antibiotics unless there is blood in the motion. Antibiotics might worsen if unnecessarily used causing antibiotic associated diarrhea. I suggest you use zinc supplements (Z&D Chat Doctor)."	"If you are a doctor, please answer the medical questions based on the patient's description."
3	"my sone has left sided abd pain..lt pelvic and rt pelvic pain in the groin area.. can only stand for a short time and sitting impossible for the pain..ct showed inlarged lymph nodes..slight elevation in wt count. he has been 10 weeks and unable to get any relief other than rest and pain meds.. he is 32 6ft 3..approx 220 lb"	"Hi. If there is no relief, it is mandatory to get the biopsy of the node done ASAP to get a correct diagnosis. It may be a serious problem. As a rule - anything not getting OK within 1 to 3 weeks maximum should be removed and tested. Noe-a-days we can do this by laparoscopy too, one day care surgery...	"If you are a doctor, please answer the medical questions based on the patient's description."
4	"Hi, I have a bump on the top of my helix where usually a piercing would be, but i never had a piercing before, just informing you where this bump is. It looks like a blister, but it doesn t hurt. I don t know how long it s been there. If you could please tell me what it might be that would be great! Thank You."	"Hi, Dear,Welcome with your query to Chat Doctor. Studied your query in full depth of its details. Reviewed it in context of your health concerns. Based On the facts, Dear, you need to consult Surgeon for checking other causes of such a lump. But I would alert you to rule out TB with Erythema Nodes which may be missed as a cause of such a lump on the helix. Fibro-chon Chat Doctor. This reply would help you to plan further treatment soon with your treating doctors. Best of Luck and early recovery. Welcome any further query in this regard, which would be replied in next session. Good Day!!"	"If you are a doctor, please answer the medical questions based on the patient's description."

5	"My baby is 3 month old. She cannot sleep at night.she feels sleepy, so when she is on our lap she is sleeping ,,but when we are trying to put her down in bed , she wakes up.,please let me know the solution of this. We have also checked whether her nose are blocked or not . But she is not having blocked nose."	"DearWelcome to ChatDoctorWe understand your concerns went through your details. 3 year child only understands security and care. She clings to anything that provides security and care. Naturally, she is clinging to you. Let her do it. While putting her down, you also lie down with her. Nursing habit also can be imparted while laying down. Such a change of habit, will change her way of looking at security and care. These are related to training. You need to use alternate techniques of training. If you require more of my help in this aspect, please use this URL. http://goo.gl/aYW2pR. Make sure that you include every minute details possible. Hope this answers your query. Available for further clarifications. Good luck."	"If you are a doctor, please answer the medical questions based on the patient's description."

3.2. Model Fine Tuning

The dataset must be transformed into a format compatible with the BERT-Base [15] model before fine-tuning can begin. In this step, the texts are divided into tokens and each token is compared to an index in the dictionary used by BERT. Training on the dataset updates the BERT-Base model as part of the fine-tuning process [16]. Optimization approaches are used to reduce the loss of the model during training. These steps improve the BERT-Base model for the chatbot task. After the fine-tuning process, the BERT-Base model is adapted to better handle the work of building chatbots. The model takes user queries, embeds them in representations, and then generates the best responses.

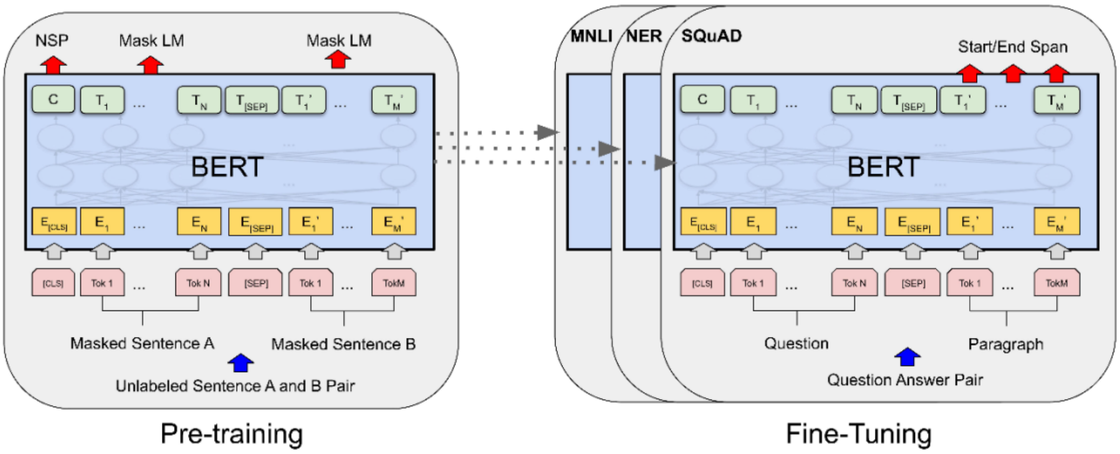


Figure 1. Base and fine-tuned BERT model architecture

We need to use the tokenizer provided by the library in order to use the pre-trained BERT. This is because the model has a specific, fixed vocabulary, and the BERT tokenizer has a special way of handling words that are not in the vocabulary.

In addition, we are required to add special tokens at the beginning and end of each sentence, pad & truncate all sentences to a single constant length, and explicitly specify which tokens are padded with the "attention mask".

3.2. Train The Model

BERT-base consists of 12 transformer layers. Each transformer layer takes a list of token embeddings and produces the same number of embeddings with the same hidden size (or dimensions) on the output. The output of the last transformer layer of the [CLS] token is used as sequence features to feed a classifier.

The Transformers library has a class called BertForSequenceClassification. This class is designed for classification tasks. However, we are going to create a new class so that we can specify our own chosen classifiers.

Below, we will create a BertClassifier class with a BERT model to extract the last hidden layer of the [CLS] token and a single hidden layer feedforward classifier.

3.3. Optimizer and Learning Rate Scheduler

To fine-tune our Bert classifier, we need to create an optimizer. The authors recommend the following hyperparameters:

- Batch size: 32
- Learning rate (Adam): 5e-5
- Number of epochs: 30

We will train our Bert classifier for 30 epochs. In each epoch, we will do training of our model and evaluation of its performance on the validation set.

3.4. Evaluation Criteria

Several performance measures [17] were compared to assess the performance of each model. Accuracy, Recall, Precision, F1 score rate are among the metrics used [18].

For binary classification problems, the discrimination score of the best (optimal) solution during classification training can be defined based on the confusion matrix as shown in Table 2. The row of the table represents the predicted class, and the column represents the actual class. From this confusion matrix, True Positive (TP) and True Negative (TN) indicate the number of positive and negative examples that were correctly classified. Meanwhile, False Negative (FN) and False Positive (FP) indicate the number of misclassified negative and positive examples, respectively. From Table 2, several commonly used metrics can be constructed to evaluate the performance of the classifier with different evaluation foci, as shown in Table 3. Due to multi-class problems, several of the metrics listed in Table 3 have been extended for multi-class classification evaluations.

Table 2. Confusion matrix for binary classification

	Real Positive Class	**Real Negative Class**
Predicted Positive Class	True Positive (TP)	False Negative (FN)
Predicted Negative Class	False Positive (FP)	True Negative (TN)

Accuracy is the most commonly used evaluation metric in practice for both binary and multi-class classification problems. Accuracy evaluates the quality of the generated solution based on the percentage of correct predictions over the total number of examples. The complementary accuracy metric is the error rate, which evaluates the generated solution based on the percentage of incorrect predictions. Both metrics have been widely used in practice by researchers to discriminate and select the optimal solution.

The advantages of accuracy or error rate are that this metric is easy to compute with less complexity, valid for multi-class and multi-label problems, easy to use scoring, and easy for humans to understand. As many studies point out, accuracy metrics have limitations in evaluation and discrimination processes. One of the main limitations of accuracy is that it produces less discriminative and less distinguishable values. As a result, it leads to less discriminative power for accuracy in selecting and determining the optimal classifier.

Table 3. Evaluation metrics

Metric	Formula	Evaluation
Accuracy (ACC)	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	In general, the accuracy metric measures the ratio of correct predictions to the total number of samples evaluated.
Error Rate (ERR)	$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN}$	Misclassification error measures the ratio of incorrect predictions to the total number of samples evaluated.
Recall	$Recall = \frac{TP}{TP + FN}$	This metric is used to measure the fraction of correctly classified positive patterns.
Precision	$Precision = \frac{TP}{TP + FP}$	Precision is used to measure accurately predicted positive patterns out of the total predicted patterns in a positive class.
F-Measure (FM)	$f_{\beta} - measure = \frac{(1 + \beta^2)x(recall * precision)}{\beta^2 + recall * precision}$	This metric is obtained by taking the harmonic mean of the precision and rating values.

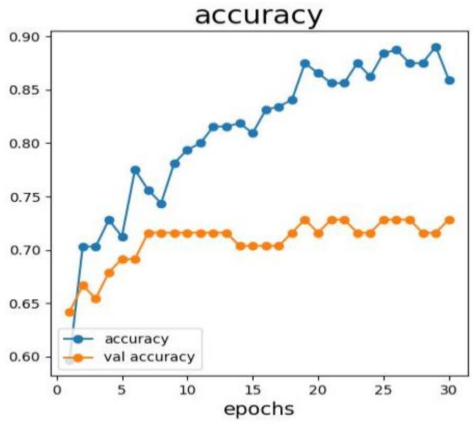
3.5. Experimental Results

Table 4. Performance values of model

Accuracy	Precision	Recall	F1 Score
0.88	0.89	0.66	0.77

The classifier results of our fine-tuned Bert model are shown in Table 4. The 88% Accuracy rate is actually at a debatable high level of performance for fine tunings made with data sets with such a high number of samples. We believe that we can increase the Accuracy and Precision rate by constantly changing the main elements that affect performance, i.e. the fine-tuning hyperparameters.

In Figure 2-3-4-5, we see the results of the derived classification performance metrics.



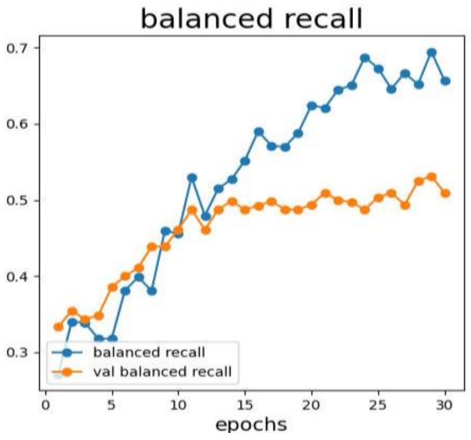
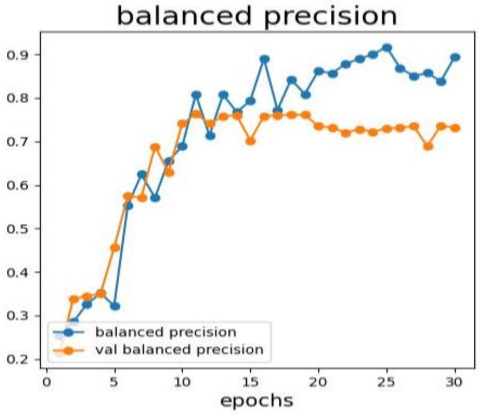


Figure 2. Accuracy graph

Figure 3. Balanced recall graph



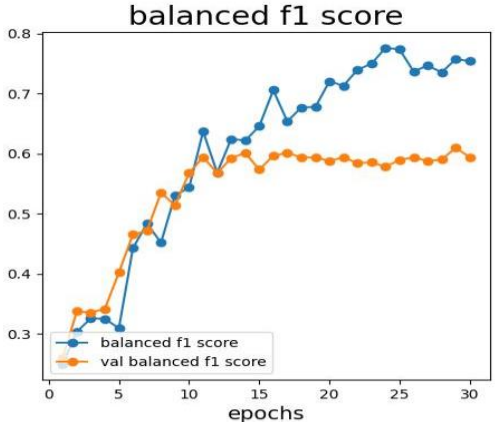


Figure 4. Balanced F1 score graph

Figure 5. Balanced precision graph

4. Conclusion and Future Works

In conclusion, the integration of artificial intelligence (AI) in the form of a chatbot into the e-health system has the potential to completely transform the way healthcare is delivered. The AI-based chatbot has proven to be an effective tool in providing personalized and timely healthcare information, assisting in triage, and improving patient engagement and satisfaction. The chatbot's performance in understanding user queries and providing accurate responses has been enhanced using machine learning algorithms, including deep learning models such as BERT. The use of chatbots in the e-health system has shown promising results in reducing healthcare costs, increasing accessibility, and improving overall healthcare outcomes.

Despite the progress and success of AI-based chatbots for e-health, there are several areas that warrant further exploration and improvement. First, the integration of more advanced natural language processing techniques, such as sentiment analysis and emotion detection, can enable the chatbot to better understand and respond to the user's emotional state, thereby providing more empathetic and personalized interactions. In addition, the integration of multimodal capabilities, such as speech and image recognition, can further enhance the functionality of the chatbot and enable it to provide more comprehensive support in the diagnosis and monitoring of health conditions. In addition, continuous refinement and expansion of the chatbot's knowledge base through regular updates and integration with reliable medical databases and resources will ensure that it remains up to date with the latest medical information. Integration with electronic health record (EHR) systems and interoperability with other healthcare platforms can also facilitate seamless information sharing and enable the chatbot to provide more personalized recommendations based on an individual's medical history. Ethical considerations and privacy are also critical areas that require attention. Ensuring that the chatbot complies with privacy regulations and maintains data security is paramount to building user trust. Ongoing research and development should focus on implementing robust privacy measures and conducting regular audits to address potential vulnerabilities. In conclusion, the future of AI-based chatbots in e-health is promising. By addressing the aforementioned areas for improvement and exploring new avenues, we can unlock the full potential of chatbots to revolutionize healthcare delivery, improve the patient experience and ultimately improve overall healthcare outcomes.

References

[1] C. Zielinski *et al.*, “Chatbots, ChatGPT, and Scholarly Manuscripts - WAME Recommendations on ChatGPT and Chatbots in Relation to Scholarly Publications,” *Afro-Egyptian Journal of Infectious and Endemic Diseases*, vol. 13, no. 1, pp. 75–79, Mar. 2023, doi: 10.21608/AEJI.2023.282936.

[2] N. Bhirud, S. Tatale, S. Randive, S. Tataale, and S. Nahar, “A Literature Review On Chatbots In Healthcare Domain Computational Feasibility of Paninian Grammar for Indian Languages' Analyses View project Machine Learning View project A Literature Review On Chatbots In Healthcare Domain,” *International Journal Of Scientific & Technology Research*, vol. 8, p. 7, 2019, Accessed: Jun. 06, 2023. [Online]. Available: www.ijstr.org

[3] S. Laumer, C. Maier, F. Tobias Gubler, and F. Tobias, “Chatbot Acceptance In Healthcare: Explaining User Adoption Of Conversational Agents For Disease Diagnosis,” 2019, Accessed: Jun. 06, 2023. [Online]. Available: https://aisel.aisnet.org/ecis2019_rp/88

[4] “View of Doctor Recommendation Chatbot: A research study.” https://sabapub.com/index.php/jaai/article/view/310/240 (accessed Jun. 06, 2023).

[5] Y. Windiatmoko, R. Rahmadi, A. F. Hidayatullah, R. Pradhan, J. Shukla, and M. Bansal, “K-Bot' Knowledge Enabled Personalized Healthcare Chatbot”, doi: 10.1088/1757-899X/1116/1/012185.

[6] M.-Y. Huang, C.-S. Weng, H.-L. Kuo, and Y.-C. Su, “Using a chatbot to reduce emergency department visits and unscheduled hospitalizations among patients with gynecologic malignancies during chemotherapy: A retrospective cohort study,” 2023, doi: 10.1016/j.heliyon.2023.e15798.

[7] J. P. Rainey *et al.*, “A Multilingual Chatbot Can Effectively Engage Arthroplasty Patients With Limited English Proficiency,” 2023, doi: 10.1016/j.arth.2023.04.014.

[8] J. Luis, Z. Montenegro, C. André Da Costa, and L. P. Janssen, “Evaluating the use of chatbot during pregnancy: A usability study,” *Healthcare Analytics*, vol. 2, p. 100072, 2022, doi: 10.1016/j.health.2022.100072.

[9] E. D. Liddy, “Natural Language Processing Natural Language Processing Natural Language Processing 1,” 2001, Accessed: Jun. 06, 2023. [Online]. Available: https://surface.syr.edu/istpub

[10] S. Pandey and S. Sharma, “A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning,” *Healthcare Analytics*, vol. 3, p. 100198, 2023, doi: 10.1016/j.health.2023.100198.

[11] A. Graves, N. Jaitly, and A.-R. Mohamed, “Hybrıd Speech Recognition With Deep Bidirectional LSTM”.

[12] J. Kapoči and ¯ Ut˙ E-Dzikien˙, “A Domain-Specific Generative Chatbot Trained from Little Data”, doi: 10.3390/app10072221.

[13] Y.-L. Liu, B. Hu, W. Yan, and Z. Lin, “Can chatbots satisfy me? A mixed-method comparative study of satisfaction with task-oriented chatbots in mainland China and Hong Kong,” 2023, doi: 10.1016/j.chb.2023.107716.

[14] “zl111/ChatDoctor · Hugging Face.” https://huggingface.co/zl111/ChatDoctor (accessed Jun. 06, 2023).

[15] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Accessed: Jun. 06, 2023. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[16] H. Wang, C. Focke, R. Sylvester, N. Mishra, and W. Wang, “Fine-tune Bert for DocRED with Two-step Process”, Accessed: Jun. 06, 2023. [Online]. Available: https://github.com/

[17] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, “An improved method to construct basic probability assignment based on the confusion matrix for classification problem,” *Inf Sci (N Y)*, pp. 250–261, 2016, doi: 10.1016/j.ins.2016.01.033.

[18] “Classification Model Evaluation Metrics”, doi: 10.14569/IJACSA.2021.0120670.