Executive Summary

This report provides a comprehensive examination of the role and impact of chatbots in medical consultations. As the healthcare industry undergoes digital transformation, chatbots are emerging as valuable tools for providing instant health information, facilitating patient communication, and augmenting medical services. This report explores the key benefits, challenges, and future trends associated with the adoption of chatbots in the medical field.

INTRODUCTION: This quote from Petter Bae Brandtzaeg, project leader of Social Health Bots project highlights one of the key reasons why building medical chatbots is very important! It gives you no restraint on asking as many medical questions as you like, whenever and wherever you like.There are significant advances that are made in the medical domain to help improve the interaction between a patient and a doctor. Earlier a patient has medical help only when the patient visits the hospital physically or when the doctor itself comes to the patient. But with the digitalization of the world some of the platforms offer patient-doctor interaction online. This interaction can be in many forms such as live phone calls, video calls, live chats, etc. Although live communication provides much more immersive interaction between the doctor and patient, in terms of scheduling it may be little easier if the patient can just ask the doctor for a medical query and the doctor can find his suitable time to help the patient. With widespread development of machine learning techniques, there is a place in this flow where machine learning can perfectly exploit to significantly reduce the workload or assist the doctor and can simultaneously help with the patient with faster query resolution.

Motivation : In a nation as linguistically diverse as Bangladesh, ensuring that individuals can access medical information in their preferred language is paramount. Language should never be a barrier to understanding symptoms, seeking timely medical advice, or making informed decisions about personal health. The Bangla Medical Chatbot sets out to break down this language barrier by offering a sophisticated platform that is capable of understanding and responding to a wide array of medical inquiries, from the common to the complex.

Key Findings:

1. Enhanced Patient Engagement

Chatbots play a pivotal role in engaging patients by offering immediate responses to health-related queries. Their ability to provide accurate information on symptoms, medications, and general health advice contributes to increased patient awareness and proactive healthcare management.

2. Time Efficiency and Accessibility

Automating routine inquiries and appointment scheduling through chatbots results in improved time efficiency for healthcare providers. Patients benefit from the accessibility of instant responses, reducing the need for lengthy waiting times and facilitating quicker access to medical information.

3. Challenges in Implementation

Despite the benefits, the implementation of medical chatbots poses challenges related to ensuring the accuracy of medical information, maintaining patient confidentiality, and addressing the ethical considerations of AI in healthcare. Striking a balance between automation and the human touch in medical interactions is crucial.

Problem Overview: In this case study we aim to automate some of these query resolutions with automation. This way when the patient asks a question, the chatbot itself that has been trained on such question-answer pairs can try to come up with an answer. If at all the patient is not satisfied with the answer generated by the chatbot, the chatbot itself will suggest some of the previously answered similar question-answer pairs. If the patient doesn’t get his or her answer then we may redirect him to an actual doctor.This way we can ensure that both the doctors ‘ and patients’ time is conserved. This is achieved by not overwhelming the doctor with a lot of patients’ questions

since they are filtered at each stage of the chatbot, the number of patients that still want to contact the doctor will reduce significantly. At the same time not making patients wait for the doctor to reply to the query significantly reduces the patient's waiting time and enables the patient to come up with a quick action in solving his/her problem.

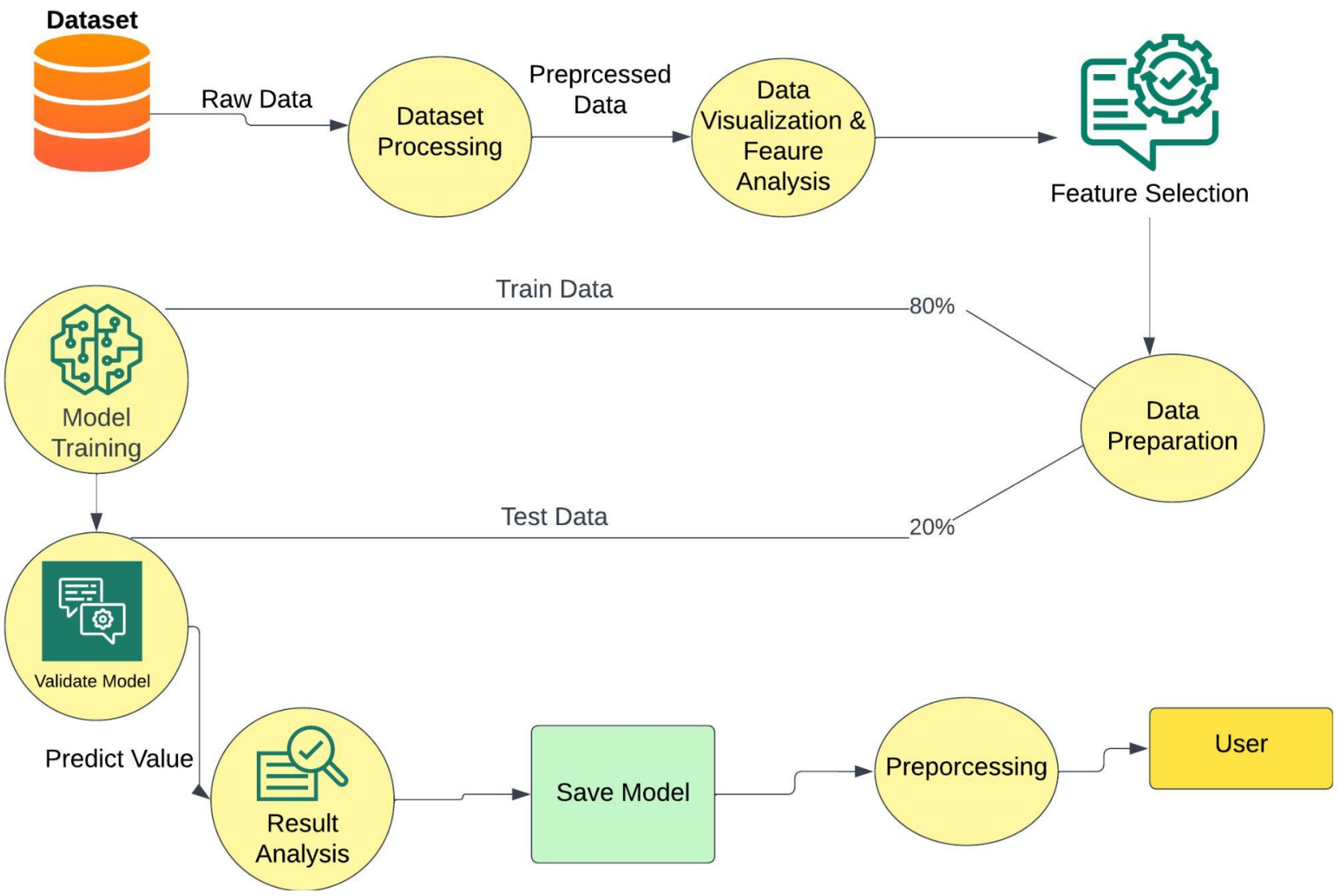
ML Formulation of the Problem: There are two main components to solve the problem which might include many sub-components. Since at first, we are trying to generate an answer to the patient's query we can pose this problem as a seq2seq task trained in supervised fashion from previously answered question-answer pairs.Secondly, when the patient is not satisfied with the answer provided, we have to show the patient similar questions with answers that have been previously answered by a doctor, this we do by doing a semantic search task in an unsupervised fashion. Although we will train neural networks to generate question and answer embeddings which we use to do this semantic search. These embeddings are generated using supervised learning binary classification approach which we will discuss as we progress through the blog.

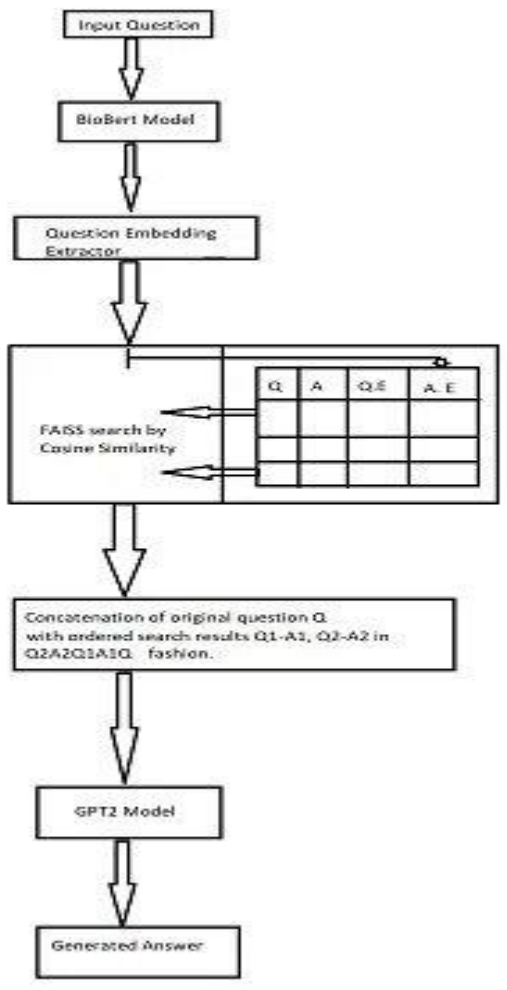
Dataset Analysis: We are using a dataset from kaggle which has question and answers of medical queries.

Performance Metrics: There are different components in this task which require different performance metrics. For the task of seq2seq generation, instead of relying on bleu score for measuring the efficiency of our model, we rather evaluate it by the manual interpretation of the quality of answers generated. The reason for not using bleu score is, even though it is generally used for evaluating seq2seq tasks, it is still not highly reliable for us to judge our model on this bleu score, at least for our biomedical task, as for some sentences it may give a very good bleu score and for some others it might give very low bleu scores. To better judge the model, the answers generated by the model should be open sourced to be judged by medical experts, only then we can come up with a correct evaluation of the model.For the task of semantic search although we do not have a labelled dataset indicating which questions are similar to calculate the overall metric, we use cosine similarity to evaluate the embeddings of question and answer returned by the model.

Overall Architecture:Our overall architecture is inspired from DocProduct, from which we use the idea of using pre-trained BioBert model to extract embeddings for previously answered question and answer pairs and use these embeddings to do a semantic search on the existing question-answer pairs. we have introduced a key component that makes sure while fine-tuning the BioBert model we do not directly pass the randomly generate negative samples but instead we make sure that the tags for these randomly generated negative sample answer do not contain any of the tags that are present for the original question. This way we make sure that our quality of embeddings is maintained by not mistakenly treating similar question and answer as negative samples.We are able to successfully integrate hugging face transformer models library to achieve

this task. Overall, in comparison to our reference, not only we have added an additional filtering system but also produced a clean reusable code that exploits the open source hugging face library.The overall architecture in the inference stage looks like this.

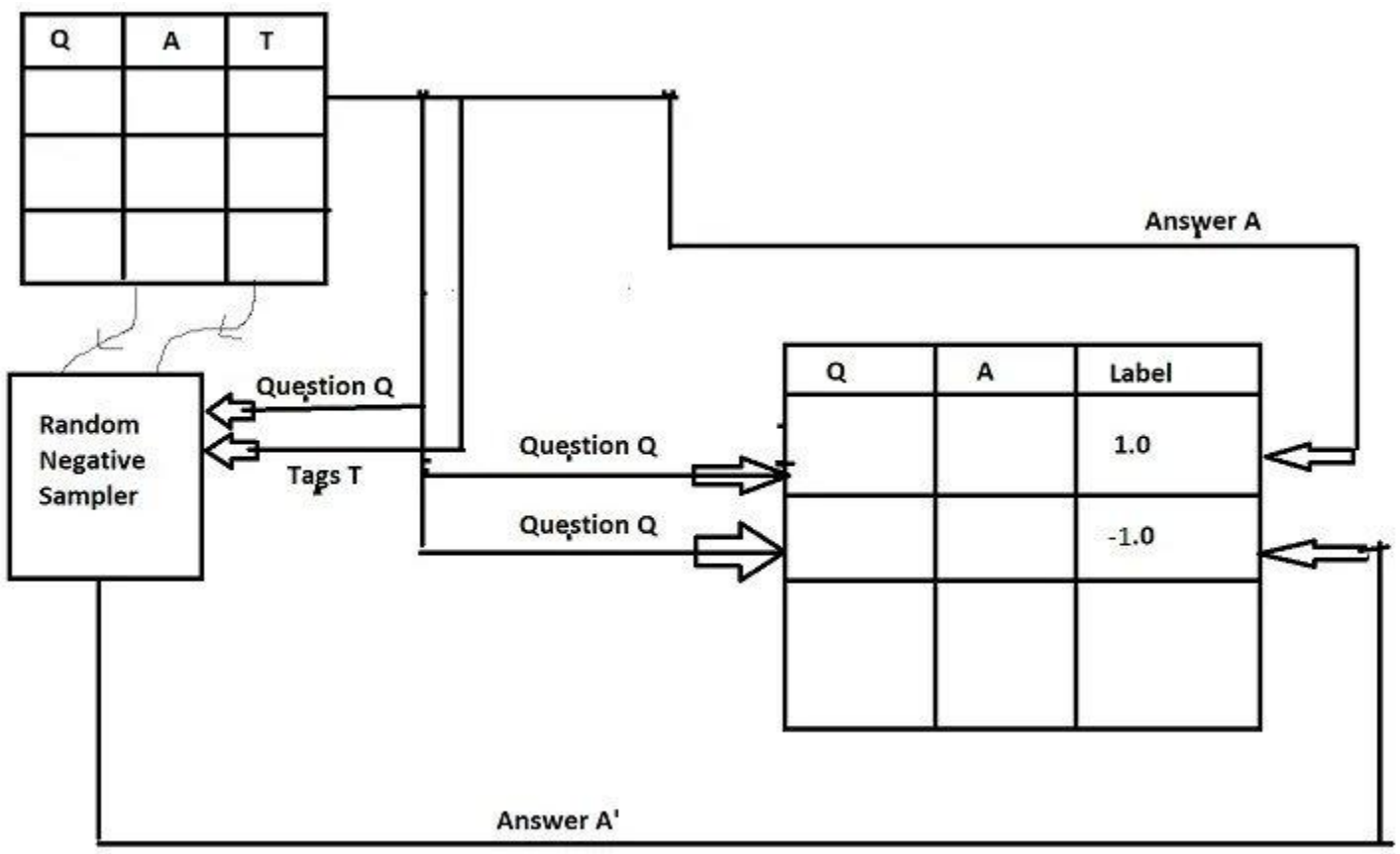




First, when a patient asks a question, using the fine-tuned BioBert model and the trained Question Embedding extractor network, the question embedding are extracted. Further using these embeddings, FAISS searches answers and the corresponding questions that are “cosinely” similar to the given question in terms of embeddings and return these similar question-answer pairs in sorted order of their cosine similarity with the given question.

These ranked question and answer pairs Q1-A1, Q2-A2, Q3-A3 are then concatenated with the original Question Q as: Q3A3Q2A2Q1A1Q. The reason for such an order is to make sure that most similar question-answer pair Q1A1 is as close as possible to the original question asked. This concatenated string in then passed as context to the fine-tuned GPT2 model so that it generates the output sentence in the Q3A3Q2A2Q1A1QA' where A' is the generated answer.

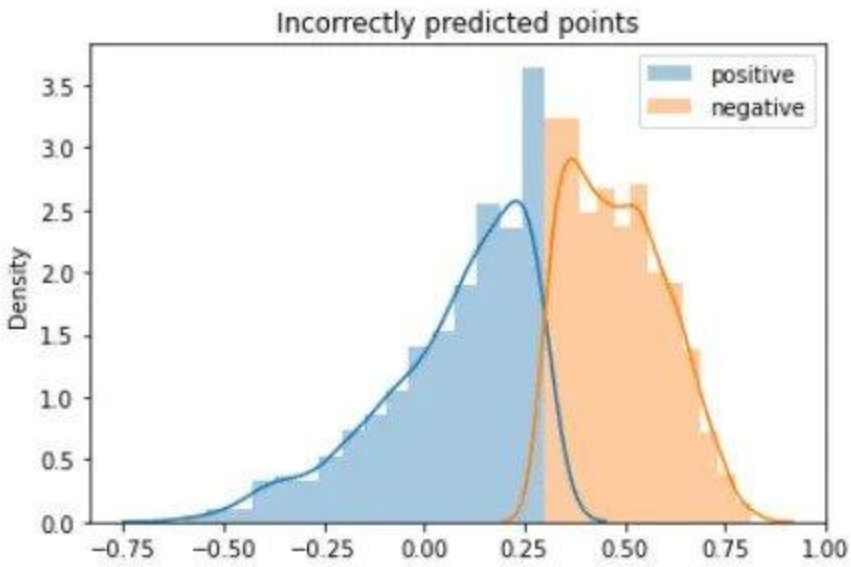
Dataset Preparation:

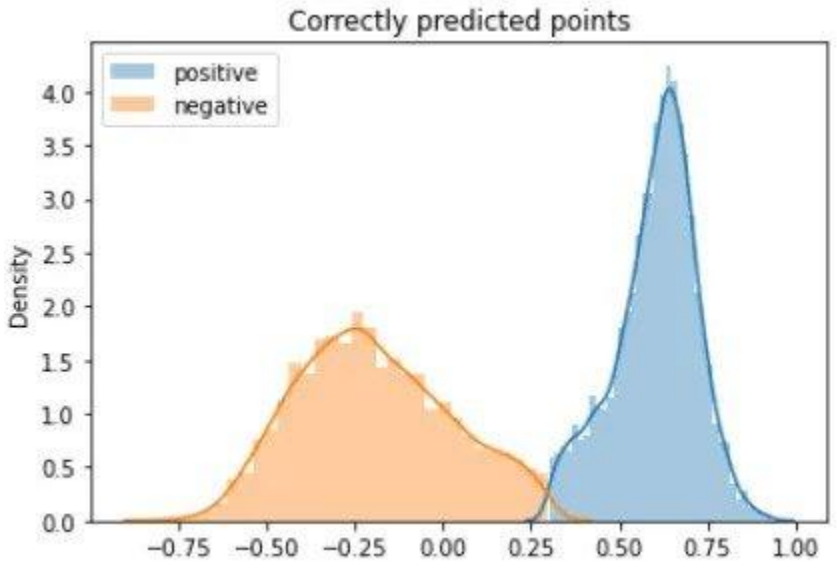


In our dataset, along with question and answers, we were also given tags which indicate the category to which the question and answer pair belongs. So for every Question Q and the Corresponding Answer A and tags T, we take the question-answer pair as the positive label(1.0) and for generating negative samples, we use both the original Question Q and the corresponding Tags T and generate negative sample Q',A', T'. We generate this negative sample by randomly picking up a tuple and checking whether the intersection between T and T' is null. If the mentioned intersection is null then we take this Answer A' and pair it with the original Question Q and consider this as the negative sample. If the intersection is not null, then we have to repeat the process of random sampling and checking for intersection.

Data Pre-processing: There are a few other minor details that we need to discuss about how we have preprocessed the question and answer text data. Preprocessing of the given question and answer data consisted of decontracting some of the words such as "won't" to "will not", remove unnecessary symbols such as !,etc. We also do not want to pad most of the question and answer with the padding constant. Therefore, when we calculated different percentiles of both question and answer data, we found that about 100% of all questions data and 99% of all answers data lie within 500 words. So post tokenizing, we truncated our sequences to a max sequence length of 512, this way we can use both the default pre-trained weights of BioBert and we do not lose any information. We also needed to pass attention mask to the BioBert model so that it understands that it has to focus on the actual content rather than on the padding token.

Model Evaluation on Validation Data: We compared the cosine similarities returned by the model for the validation dataset for both correctly classified negative and positive points, as well as the cosine similarities returned by the model for misclassified points that are originally negative and positive points. Please have a look at the below plots for the same.





From the above plots, we can observe that the correctly classified negative and positive points are well separated from each other while the incorrectly classified positive and negative points are very close to the threshold of 0.3 for which the model gave best train and validation accuracy. These two plots indicate that our model when differentiating correctly makes sure that we are well able to cosinely separate positive and negative points and when it fails to classify, it only does by a small amount.

Future Trends: The future of medical chatbots involves advancements in Natural Language Processing (NLP) and machine learning, enabling chatbots to understand and respond to medical queries with increased accuracy. Integration with Electronic Health Records (EHRs) and telemedicine platforms will further enhance the role of chatbots in the evolving landscape of healthcare services.

Future Work: In the future, we can experiment on how a transformer model can work on solving this problem directly as a seq2seq problem. Even if transformer model can directly generate answers to the given question, we still need BioBert to extract similar patient question-answer pairs. We can also work on gathering more biomedical data so that the model doesn't overfit to the given data and can generate more appropriate answers.

Conclusion:In conclusion, medical chatbots are proving to be valuable assets in the healthcare ecosystem, contributing to improved patient engagement, streamlined communication, and efficient healthcare delivery. Despite challenges, the continued development of AI in healthcare positions chatbots as valuable tools for both patients and healthcare professionals.

References:

https://arxiv.org/abs/1706.03762
https://arxiv.org/abs/1810.04805
https://arxiv.org/ftp/arxiv/papers/1901/1901.08746.pdf
https://d4mucfpksywv.cloudfront.net/better-language-models/languagemodels.pdf%C2%A0
https://github.com/ash3n/DocProduct#start-of-content
https://appliedaicourse.com