# Bangla Question Anwering System Using BERT Language Model.

**Sanjana S. Khan**
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
sanjana.sabah.khan@g.bracu.ac.bd

**Mehreen Khan**
Computer Science
Brac University
Dhaka, Bangladesh
mehreen.khan@g.bracu.ac.bd

**Dewan G. Mortoza**
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
dewan.golam.mortoza@g.bracu.ac.bd

**Annajiat Alim Rasel**
Senior Lecturer
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat@bracu.ac.bd

## Abstract

In recent years, there has been uncountable research on Question Answering (QA) systems in different languages. Still, QA is a critical NLP problem and a long-standing artificial intelligence milestone which is yet to reach perfection. Choosing the right language model and training it on a standardized dataset is the key to perfecting a QA system. Bidirectional Encoding Representations for Transformers (BERT) models are known to perform competently on complex information extraction tasks. For our work, we have prepared a dataset by combining selective data from two open source datasets. We have sliced the dataset into 3 splits - training, development and testing sets. In this paper, We are training and fine-tuning the BanglaBERT model using our dataset consisting of 3,798 tuples of question-passage pairs, and comparing our obtained accuracy score with other BERT based models - mBERT, RoBerta and DistilBert and SahajBert.

**Keywords** : Natural Processing Language, Question Answering system, Bangla, BERT

## 1 Introduction

Question Answering has been a very crucial part of any language study. The term Question Answering (QA) here comes from the reading comprehensive where the reader is given certain paragraphs to read and answer some questions related to those paragraphs. Using this concept, there are many models that are trained to do so using machine learning algorithms. They can understand the text context and answer the question related to that context. They can answer using the natural language that we are used to. It not only answers but also tries to correctly understand the sentiments of questions and look for the particular text or sentence it is looking for.

With the Internet making vast quantities of text available at our fingertips, QA systems have become critical for us to better understand and quickly extract key information from text. However, creating a training set for a Supervised Learning Model is difficult since each portion does not have a predetermined amount of sentences and answers can range from one word to many words. Computers need to understand the meaning of ambiguous language in text by using surrounding text to establish context. This problem was solved in 2018 when Google launched a language model - Bi-directional Encoder Representation from Transformer (BERT) - which can be used for Natural Language Processing (NLP) tasks (Devlin et al., 2019). BERT is a transfer learning model which means it is trained for one task but can be reused for another task. IT has a trained Transformer Encoder stack, with twelve in the Base version, and twenty-four in the Large version. There is many BERT provided by Hugginface for multiple tasks like, Summarization, Fill-Mask, Question-Answering etc.

Bangla is the official and national language of Bangladesh, with 98% of Bangladeshis using Bangla as their first language. Within India, Bangla is the official language of the states of West Bengal, Tripura and the Barak Valley region of the state of Assam. It is also a second official language of the Indian state of Jharkhand since September 2011. is the fifth most-spoken native language and the seventh most spoken language by total number of speakers in the world. Bangla is the fifth most spoken Indo-European language. Designing a QA system for Bangla is quite challenging as Bangla is a very different language from English in terms of alphabet, pronunciation, grammar and vocabulary. In this pager, we are designing a QA system in NLP for Bangla language that allows users to give questions and receive an automatic response using a given comprehension.

## 2 Related Work

For the past few years, there has been substantial research on English QA systems, but only a handful for Bangla. In 2021, Arnab Saha, Mirza I. Noor, Shahriar Fahim and three other authors from RUET, designed BERT-Bangla, a language model pre-trained on Bangla unlabelled text (Saha et al., 2021). They tested BERT-Bangla on different Bangla NLP classification tasks and attained better results than any other existing Bangla language models. They also developed the Bangla Question Answering Dataset (BQuAD) from scratch. Then, they fine-tuned the pre-trained language model on BQuAD for QA tasks and achieved an EM accuracy of 45% and F1 accuracy of 78.5%. Later in 2022, Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad and five other researchers from BUET, proposed two pretrained models in Bangla - BanglaBERT and BanglishBERT (Bhattacharjee et al., 2022). They created Bangla Natural Language Inference (NLI) and QA datasets.They also introduced the Bangla Language Understanding Benchmark (BLUB). In the supervised setting, BanglaBERT outperforms mBERT and XLM-R (base) by 6.8 and 4.3 BLUB scores, while in zero-shot crosslingual transfer, BanglishBERT outperforms them by 15.8 and 10.8 precisely.

In addition, there is some notable works for Arabic QA system . In 2014, Heba Abdelnasser, Maha Ragab, Reham Mohamed and four other researchers from Alexandria University,Egypt, proposed a new Arabic QA system - Al-Bayan, designed for the Holy Qur'an (Abdelnasser et al., 2014). This system can achieve upto 85% accuracy using the top-3 results. They used machine learning techniques to build a Semantic Information Retrieval module that maps fragments of natural language text into a weighted vector of Quranic concepts. In OSACT 2022 workshop, a shared task of Question Answering on the Holy Qur'an was given along with The Qur'anic Reading Comprehension Dataset (QRCD). One solution to this shared task, a proceeding paper was published on 3 June 2022 which proposed some post-processing operations to enhance the quality of answers aligning with the official measure (Elkomy and Sarhan, 2022). The authors fine-tuned a variety of BERT models optimized for the Arabic language and achieved a Partial Reciprocal Rank (pRR) score of 56.6% on the official test set.

## 3 Dataset

For our research, we have collected two datasets from open-sources and after thorough analysis we used a selective combination of both to form our dataset. The first dataset we have collected is created on by the researchers at Jahangirnagar University on December, 2022. They have developed a reading comprehension based question answering dataset in Bangla Language (Aurpa et al., 2022) for developing automatic reading comprehension systems. They have comprised a real-time and long passage (context) in the Bengali language which is collected from different Bangla articles, novels, biography, etc. Questions are then generated based on the given passages. The dataset consists of 3,798 tuples of question-passage pairs that are coupled with their extractive answers. For our purpose, we have split the question-passage pairs into training, development and testing sets. The dataset follows the same format as the commonly used reading comprehension dataset. The other collected dataset was created by the at BUET on July, 2022 (Bhattacharjee et al., 2022)

## 4 QA System Framework

## 5 Evaluation

## 6 Conclusion

## Acknowledgments

## References

Heba Abdelnasser, Maha Ragab, Reham Mohamed, Alaa Mohamed, Bassant Farouk, Nagwa El-Makky, and Marwan Torki. 2014. Al-bayan: An Arabic question answering system for the holy quran. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 57–64, Doha, Qatar. Association for Computational Linguistics.

Tanjim Taharat Aurpa, Richita Khandakar Rifat, Md Shoaib Ahmed, Md Musfique Anwar, and A. B.

M. Shawkat Ali. 2022. Uddipok: Reading comprehension based question answering dataset in bangla language.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohamemd Elkomy and Amany M. Sarhan. 2022. TCE at qur'an QA 2022: Arabic language question answering over holy qur'an using a post-processed ensemble of BERT-based models. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 154–161, Marseille, France. European Language Resources Association.

Arnab Saha, Mirza Ifat Noor, Shahriar Fahim, Subrata Sarker, Faisal Badal, and Sajal Das. 2021. An approach to extractive bangla question answering based on bert-bangla and bquad. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pages 1–6.