Unlocking the Potential of Multiple BERT Models for Bangla Question Answering in NCTB Textbooks*

Abdullah Khondoker, Enam Ahmed Taufik, Md Iftekhar Islam Tashik, S M Ishtiak mahmud, Antara Firoz Parsa
Abid Hossain, Md Humaion Kabir Mehedi, and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{abdullah.khondoker, enam.ahmed.taufik, iftekhar.islam.tashik, sm.ishtiak.mahmud, antara.firuz.parsa
abid.hossain, humaion.kabir.mehedi}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—Evaluating text comprehension in educational settings is critical for understanding student performance and improving curricular effectiveness. This study investigates the capability of state-of-the-art language models—RoBERTa, Bangla-BERT, and BERT Base—in automatically assessing Bangla passage-based question-answering from the National Curriculum and Textbook Board (NCTB) textbooks for classes 6-10. We compiled a dataset of around 3,000 Bangla passage-based question-answering instances and evaluated the models using F1 Score and Exact Match (EM) metrics. Our findings indicate variable performance among the models with respect to the criteria examined. Bangla-BERT displayed superior performance in F1 Score, while RoBERTA underperformed among all the models. These outcomes suggest that machine learning models can be viable tools for evaluating text comprehension in educational textbooks, albeit with room for optimization. The study lays the groundwork for future research aimed at implementing automated evaluation systems in educational institutions and provides insights into the strengths and limitations of each model in the context of Bangla text comprehension.**

***Index Terms*—RoBERTa, Bangla-Bert, BERT Base, Text Comprehension, Bangla Language National Curriculum and Textbook Board (NCTB), Educational Evaluation, F1 Score, Exact Match Score, Natural Language Processing, Automated Assessment, Question Answering, Language Models**

I. Introduction

The continuous evolution of Natural Language Processing (NLP) techniques is transformative, especially in the domain of Question Answering (QA). This study embarks on a novel journey to harness the capabilities of hybrid NLP models, focusing particularly on the Bengali language—a realm previously underrepresented in modern QA systems.

Central to the paper's objective is the creation of a tailored dataset, finely curated to cater to specific goals. The dataset, meticulously composed of around 3,000 passage-question-and-answer pairs, draws from the educational foundation of the Bangla language, encompassing classes six to ten. These pairs, expertly selected by human annotators in consultation with authoritative NCTB textbooks, form the bedrock of a comprehensive QA resource. The dataset is meticulously diversified, accommodating distinct question types.

To process this expansive dataset, we adopted a systematic methodology employing the power of the three distinct models: Bert-Base-Multilingual-Uncased [1], Bangla-Bert [2], and Roberta [3]. Through rigorous tokenization, preprocessing, and training, we endeavoured to sculpt a QA system optimized for Bengali language comprehension.

In the subsequent sections, we delve deeper into the dataset's nuances, elucidate the methodological approach, present our evaluation outcomes, and wrap up with a comprehensive conclusion, highlighting the implications and future prospects of our findings. A Sample diagram for the system is given below:

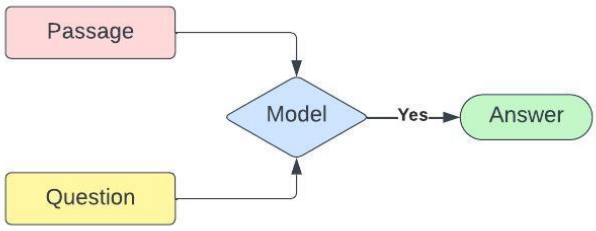

Fig. 1. The Task of Question Answering System

II. Literature review

Somnath Banerjee et al.'s [4] paper pioneers the development of a factoid question-answering system tailored for Bengali, a low-resource language. This research represents an initial step in addressing the complex challenges inherent in constructing such a system. The authors propose a system capable of answering natural language questions in a human-like manner. Questions are classified into five categories, with the system primarily focusing on factoid questions in three

distinct domains. Challenges in developing question-answering systems for low-resource languages, such as Bengali, include the prevalence and diverse positions of interrogatives and the scarcity of language processing tools. BFQA, their Factoid QA system for Bengali, comprises a pipeline architecture with three key components: question analysis, sentence extraction, and answer extraction. The authors introduce several question analysis processors, including QType and Expected Answer Type identification, named entity recognition, question topical target identification, and keyword identification. Evaluation employs the Mean Reciprocal Rank (MRR) metric, yielding an MRR of 0.32, albeit with varying performance across domains. Despite the achievement in advancing Bengali factoid question answering, this research acknowledges the system's lower accuracy compared to European languages, attributed to underperforming components like the shallow parser and named entity recognition system. In summary, this work contributes significantly to low-resource language question-answering while recognizing ongoing challenges.

In the paper by Zhou et al. [5], critical issues of overconfidence and over-sensitivity in existing Reading Comprehension (RC) models are effectively addressed. Through their experiments, the study showcases significant improvements in the robustness of RC models. The key innovation of the paper lies in the development of a method that incorporates external knowledge to enforce a range of linguistic constraints, including entity, lexical, and predicate constraints. This integration empowers the model to generate more precise predictions, not only for semantic differences but also for semantic equivalences in adversarial examples. Furthermore, Zhou et al. [5] introduce posterior regularization into RC models, a crucial addition that contributes to enhancing the resilience of the underlying RC models. By seamlessly incorporating linguistic constraints and posterior regularization into the learning process, the proposed method not only bolsters the robustness of base RC models but also successfully integrates these constraints, resulting in more accurate and adaptable RC systems.

Bechet et al. [6] embark on an intriguing exploration of question-answering within digitized archive collections, with a particular focus on its applicability to Social Science studies. Their research introduces an innovative approach that harnesses the power of a BART Transformer-based generative model, enhanced by semantic constraints, to tackle the task of question generation. This pioneering approach holds great promise in the field of natural language processing, especially within the niche of archive collection studies. The study meticulously conducts experiments on three distinct corpora: FQUAD, CALOR-QUEST, and ARCHIVAL, each presenting its own set of complexities and challenges. The findings from these experiments unveil a notable enhancement in the quality of questions generated when semantic annotations are incorporated, underscoring the effectiveness of this approach in improving question generation. However, an intriguing observation emerges during the evaluation of Machine Reading for Question Answering (MRQA) performance. When models trained on automatically generated questions are used for evaluation, the study notes no significant improvements, particularly on the CALOR-QUEST and ARCHIVAL datasets. This intriguing finding underscores the nuances associated with using automatically generated questions in the evaluation process, opening up avenues for further investigation in this domain. Nevertheless, the research shines in its performance on the formidable ARCHIVAL dataset, demonstrating the robustness and adaptability of the proposed approach, even in challenging contexts. The analysis thoughtfully highlights the distinctions between questions generated by experts and those sourced from crowdsourcing efforts, shedding light on the importance of question quality in this context. [6]

Rathod, Manav et al. [7], to address the lack of a Bangla Question/Answer Dataset, introduces BanglaRQA, a reading comprehension-based question-answering dataset for Bangla. It contains 3,000 context passages and 14,889 question-answer pairs, covering answerable and unanswerable questions across four question categories and three answer types. The paper [7] also evaluates four Transformer models on BanglaRQA, with the best model achieving 62.42 percent EM, and 78.11 percent F1 scores. However, further analysis reveals variation in performance across question-answer types, indicating room for improvement. The paper [7] demonstrates the effectiveness of BanglaRQA as a training resource by achieving strong results on the bn$_squaddataset$.

Jacob Devlin et al. [8] explore the application of NLP techniques, specifically neural language models, for generating question/answer exercises from English texts. The aim is to support ESL teaching to children by generating beginner-level exercises. The proposed approach involves a four-stage pipeline: pre-processing, answer candidate selection, question generation using the T5 transformer-based model, and post-processing. Evaluation of benchmark datasets demonstrates comparable results to previous works. However, limitations are identified, such as imperfect co-reference resolution and errors in question generation. Future work involves refining the system by exploring other language models, expanding the evaluation corpus, and fine-tuning the model for specific English proficiency levels. The developed tool has the potential for integration into an educational platform for English language teaching. [8]

Yinhan Liu et al. [9] investigate the development of a deep learning-based question-answering system in Bengali, aiming to overcome the limitations and lack of progress in this field. By leveraging state-of-the-art transformer models, the research focuses on training a QA system using a synthetic reading comprehension dataset translated from SQuAD 2.0. Furthermore, a human-annotated QA dataset sourced from Bengali Wikipedia is utilized for evaluating the models. Comparative analysis with human children provides valuable insights and establishes a benchmark score. The research [9] emphasizes the importance of addressing the challenges in low-resource language settings, particularly in the context of reading comprehension-based question answering.

In this paper Bhattacharjee et al. [10] introduce the task of

Multi-Question Generation, aiming to generate diverse questions assessing the same concept. It addresses the limitation of existing systems that generate only one question per answer. The paper proposes an evaluation framework based on desirable question qualities and presents results comparing different question generation approaches. The authors highlight the issue of word overlap between generated questions and input passages and propose metrics to measure question answerability, semantic similarity, and distinct wordings. Future work includes exploring human evaluation metrics, reinforcement learning objectives, and advanced paraphrase systems. The paper suggests incorporating teacher evaluation to define desirable question properties and evaluate the educational impact of diverse question wordings. The publicly released pipeline holds the potential for enriching educational resources at scale. [10]

III. Dataset

In the context of our project's focus on the Bangla language, a pivotal component involves the creation of a dataset tailored to our objectives. Emphasizing the essence of customization, we undertook the task of dataset construction, marking a foundational stride in the development of an effective Bangla Question-Answering system. Our dataset is a result of a meticulous curation process involving around 3,000 question-and-answer pairs. These pairs, meticulously selected by human annotators consulting NCTB textbooks from classes six to ten, offer a contextual and informative foundation. Each passage in the dataset, averaging 387 words, provides substantial context for the subsequent question answering. Importantly, human annotators meticulously collected responses for each question type, ensuring the dataset's reliability and relevance. The primary objective driving this endeavour is the development of a proficient Bangla question-answering system. To this end, we painstakingly organized the dataset into training and validation subsets, with each subset elegantly encapsulated within CSV files. These files harmoniously interweave multiple passages with their corresponding questions and expertly annotated answers. By taking the reins in dataset creation, we lay the groundwork for a question-answering system deeply rooted in precision, contextual understanding, and linguistic nuance intrinsic to the Bangla language. Here Figure [02] is an example of our dataset:

Context: খাদ্যের উপাদান ছয়টি। সেগুলো হলো: শর্করা, আমিষ, স্নেহ, ভিটামিন, খনিজ লবণ এবং পানি। এগুলোর মধ্যে শর্করা, আমিষ ও স্নেহ পদার্থ (বা ফ্যাট) দেহ পরিপোষক খাদ্য। খাদ্যের স্নেহ এবং শর্করাকে বলা হয় শক্তি উৎপাদক খাদ্য এবং আমিষ যুক্ত খাদ্যকে বলা হয় দেহ গঠনের খাদ্য। ভিটামিন, খনিজ লবণ ও পানি দেহ সংরক্ষক খাদ্য উপাদান, যেগুলো দেহের রোগ প্রতিরোধে সাহায্য করে।
Question: কোনটি দেহ পরিপোষক খাদ্য?
Answer: আমিষ

Fig. 2. Sample of Dataset

IV. Methodology

This segment outlines the framework for an Automated Question Answering System in Bengali. Given the limited prior work in Bangla question answering, our objective is to develop a system that provides accurate responses to user queries, addressing the gap in this area. In this research endeavour, we meticulously processed the Bangla question-answering dataset to prepare it for subsequent modelling and evaluation. A step-by-step approach was undertaken to ensure the quality and integrity of the dataset. The model represented in Figure [03] of our workflow-

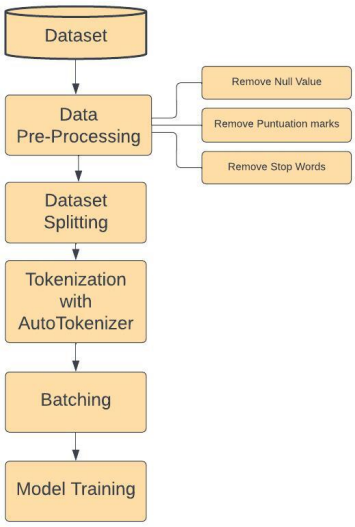

Fig. 3. Workflow

A. Data Preprocessing

The initial stage of data preprocessing involved the removal of null values from the dataset, ensuring that the foundational dataset was devoid of any missing or incomplete entries. Following this step, we employed a Basic Tokenizer to facilitate the creation of tokenized data. The Average number of tokens is shown in Figure [04]:

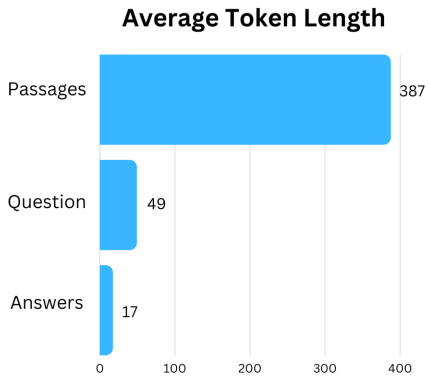

Fig. 4. Average number of tokens

This tokenization procedure was instrumental in the removal of punctuation marks (e.g. ., $, %, , *, -, etc.) and stop

words [11], streamlining the text data for subsequent analyses. Subsequently, we restored the dataset to its original format, preserving its coherence. To accurately identify the span of answers within the provided passages, we meticulously calculated the start and end indices of the answers. The answer column was then transformed into a structured dictionary, encompassing the answer itself along with its corresponding start and end indices.

B. Dataset Splitting

The dataset was divided into distinct training and testing subsets to facilitate the model development and evaluation process. Specifically, a 70%-30% train-test split was employed, ensuring a robust assessment of model generalization.

C. Tokenization with AutoTokenizer

Our empirical findings revealed that, among the considered tokenization techniques, AutoTokenizer exhibited superior performance during model training. This tokenizer, with its adaptive tokenization strategy, proved to be particularly adept at handling the intricacies of Bangla text. To equip the dataset for subsequent modelling endeavours, we harnessed the power of AutoTokenizer. The AutoTokenizer provided by the Hugging Face Transformers library was employed for tokenization. It tokenizes both the questions and contexts, breaking them down into subword units and mapping them to corresponding token IDs. Additionally, attention masks were generated to indicate which tokens should receive attention during training. After tokenization, the data was preprocessed to generate inputs that the model could process. These inputs included tokenized sequences, attention masks, and answer span positions. The resulting tokenized inputs were organized into a structured format, which included input IDs, attention masks, start positions, and end positions for answer spans. The tokenized inputs were then transformed into a list of dictionaries, where each dictionary contained input IDs, attention masks, start positions, and end positions. These dictionaries were structured to facilitate feeding the data into the model during training. This tokenization technique incorporated essential parameters, including truncation (enabled), padding (set to 'max-length'), maximum token length (512 tokens), and the return of attention masks and tokenized tensors (return-attention-mask=True, return-tensors='pt'). Notably, we refrained from adding any special tokens to preserve the native structure of the Bangla text.

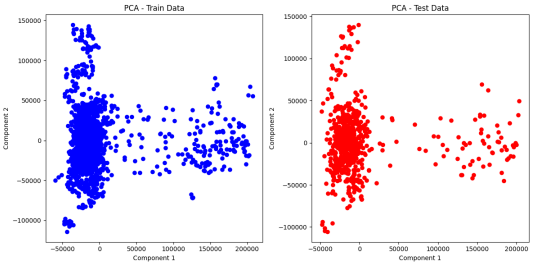

Fig. 5. PCA - Train & Test for RoBERTa Base

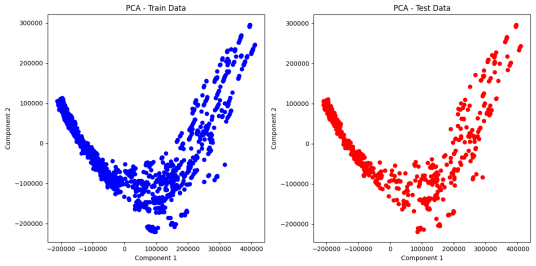

Fig. 6. PCA - Train & Test for BERT Base

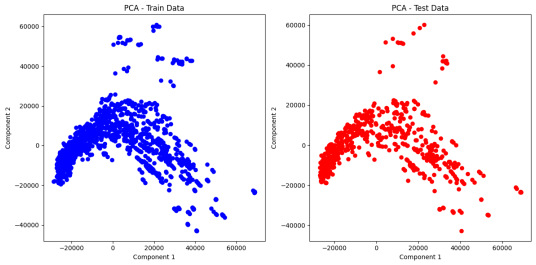

Fig. 7. PCA - Train & Test for Bangla-BERT

D. Null Value Handling

A rigorous check for null values was conducted post-tokenization. In instances where null values were encountered, we judiciously replaced them by adopting a strategy based on the maximum allowable token length.

E. Batching

For efficient model training, two distinct batch sizes were utilized. The first configuration involved a train batch size of 16 and a test batch size of 8, while the second configuration featured a train batch size of 32 and a test batch size of 16. This bifurcation aimed to explore the impact of batch size on model performance.

F. Model Training

Three distinct models were chosen for the training phase: Bert-Base-Multilingual-Uncased [1], Bangla-Bert [2], and Roberta [3]. To attain optimal performance, these models underwent extensive training with varying epochs and learning rates. The long training process was used to guarantee an accurate evaluation of the models' capabilities and to deliver the highest level of accuracy possible. In each iteration of the training loop, batches of tokenized question-context pairings were processed, producing projected answer spans. By comparing predicted answer locations with actual response placements, the loss was computed. Backpropagation gradients were used to adjust the model's parameters and reduce loss. At regular intervals, the system assessed its performance using validation data, and logging metrics like accuracy. Throughout training, we logged vital details including progress, losses, and metrics, offering insights into the learning process.

REFERENCES

[1] K. L. K. T. Jacob Devlin, Ming-Wei Chang, "Bert: Pre-training of deep bidirectional transformers for language understanding." in *Proceedings of NAACL-HLT 2019*, vol. 1, no. 1, Jun. 2019, pp. 4171–4186.

[2] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98

[3] N. G. J. D. M. J. D. C. O. L. M. L. L. Z. V. S. Yinhan Liu, Myle Ott, "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692.

[4] S. K. N. Somnath Banerjee and S. Bandyopadhyay., "Reading comprehension based question answering system in bangla language with transformer-based learning," *International Conference on Text, Speech and Dialogue*, 2014.

[5] M. Zhou, M. Huang, and X. Zhu, "Robust reading comprehension with linguistic constraints via posterior regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2500–2510, 2020.

[6] F. Bechet, E. Antoine, J. Auguste, and G. Damnati, "Question generation and answering for exploring digital humanities collections," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 4561–4568.

[7] S. M. S. Ekram, A. A. Rahman, M. S. Altaf, M. S. Islam, M. M. Rahman, M. M. Rahman, M. A. Hossain, and A. R. M. Kamal, "BanglaRQA: A benchmark dataset for under-resourced Bangla language reading comprehension-based question answering with diverse question-answer types," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2518–2532.

[8] G. Berger, T. Rischewski, L. Chiruzzo, and A. Rosá, "Generation of english question answer exercises from texts using transformers based models." IEEE, 2022.

[9] T. T. Mayeesha, A. M. Sarwar, and R. M. Rahman, "Deep learning based question answering system in bengali," *Journal of Information and Telecommunication*, vol. 5, no. 2, pp. 145–178, 2021.

[10] M. Rathod, T. Tu, and K. Stasaski, "Educational multi-question generation for reading comprehension," in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 216–223.

[11] Genediazjr, "Stopwords bengali (bn)." [Online]. Available: https://github.com/stopwords-iso/stopwords-bn