Heliyon 8 (2022) e11052

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Reading comprehension based question answering system in Bangla language with transformer-based learning

Tanjim Taharat Aurpa [b,a,*], Richita Khandakar Rifat [a], Md Shoaib Ahmed [d], Md. Musfique Anwar [a], A. B. M. Shawkat Ali [c,e]

[a] Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh
[b] Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Bangladesh
[c] Central Queensland University, Melbourne, Australia
[d] Brian Station 23 Ltd, Dhaka, Bangladesh
[e] JU Data Mining Research Lab, Dhaka, Bangladesh

ARTICLE INFO

Dataset link: https://data.mendeley.com/datasets/s9pb3h2cjy/1

Keywords:
Bangla question answering
Transformer-based learning
Reading comprehension
Bangla language
Bangla reading comprehension

ABSTRACT

Question answering (QA) system in any language is an assortment of mechanisms for obtaining answers to user questions with various data compositions. Reading comprehension (RC) is one type of composition, and the popularity of this type is increasing day by day in Natural Language Processing (NLP) research area. Some works have been done in several languages, mainly in English. In the Bangla language, neither any dataset available for RC nor any work has been done in the past. In this research work, we develop a question-answering system from RC. For doing this, we construct a dataset containing 3636 reading comprehensions along with questions and answers. We apply a transformer-based deep neural network model to obtain convenient answers to questions based on reading comprehensions precisely and swiftly. We exploit some deep neural network architectures such as LSTM (Long Short-Term Memory), Bi-LSTM (Bidirectional LSTM) with attention, RNN (Recurrent Neural Network), ELECTRA, and BERT (Bidirectional Encoder Representations from Transformers) to our dataset for training. The transformer-based pre-training language architectures BERT and ELECTRA perform more prominently than others from those architectures. Finally, the trained model of BERT performs a satisfactory outcome with 87.78% of testing accuracy and 99% training accuracy, and ELECTRA provides training and testing accuracy of 82.5% and 93%, respectively.

1. Introduction

In an era where technological dependency and interaction are increasing, technology achieves the ability to understand human language and respond in a human-like manner. One of the essential concerns, education, also requires automated systems for various purposes. This urgency is resulting in numerous types of research regarding educational patterns like Reading Comprehension (RC). The preparedness defines RC to read any content, acknowledging it for integrating with previously acquired knowledge. RC indicates some question answers set (e.g., simple short questions, true-false, etc.) based on a given passage in education. RC models and systems are convenient in various sectors, such as question-answering systems. Implementing automated RC models not only eases the testing system for authorities but can also help students evaluate and prepare themselves.

Bengali is the world's sixth-largest language, and 228.7 million people, including India and Bangladesh, use it as their first language. At the same time, it bears great historical importance. UNESCO declared 21st February the International Mother Language Day, honoring the historical sacrifice of language martyrs who fought for the Bangla language to be their mother tongue. Since then, it has become the mother tongue for Bangladeshi people and the language for many children's primary education.

In March, due to COVID 19 Lockdown, almost 38 million Bangladeshi Students faced difficulties with their educational needs [34, 35]. For the prevention of the COVID-19 education crisis, there require more ad-

* Corresponding author at: Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Bangladesh.
E-mail address: taurpa.cse@iubat.edu (T.T. Aurpa).

https://doi.org/10.1016/j.heliyon.2022.e11052
Received 5 May 2022; Received in revised form 18 August 2022; Accepted 7 October 2022

vanced and automated systems in Bangla languages [38]. Contrary to all these glories, Bengali is one of the most minor studied languages in reading comprehension. It is still hard finding a fitting dataset for Bengali. Therefore demonstration of the RC-based Bangla question-answer system is essential using the latest technologies like transformer-based learning.

Transformers are the latest deep learning architecture that has proven adequate to deal with sequential data without recurrent networks like GRU or LSTM. It uses an attention mechanism with an encoder-decoder stack, and its escalated parallelization feature makes it possible to run large datasets more efficiently. Some impactful utilization of transformers can be seen in [21, 44]. One of the most significant transformer-based architectures is Bidirectional Encoder Representations from Transformers (BERT).

BERT has gained enormous popularity from its introduction to NLP. It is a transformer-based Pre train language model that uses the rebellious self-attention mechanism for classification and prediction tasks. This state of art language model outperforms almost all types of NLP equipment. In [51] BERT is used for text classification and brought forward different modified architectures of BERT for more satisfactory performance. This pre-trained state-of-art language model brings lofty accuracy in various NLP sectors as question answering (e.g. [29]), sentiment analysis (e.g. [27], [50], [42]), entity extraction as well as recognition ([49], [40], [6]). BERT found its way to rule other languages besides English with the introduction of mBERT. mBERT is a variation of the BERT model that is able to deal with different languages besides English. mBERT has been implemented for languages like Bangla, Greek, Danish, Turkish, etc. It contributes to multilingual text classification [25, 26, 33], offensive language detection [18], Word Sense Disambiguation [53], Translation Quality Estimation [16, 22], etc.

BERT conquered the NLP world once it was revealed. Nonetheless, in 2020, ELECTRA, a new pre-train language model (PLM), was introduced, which overcomes the constraints of mask language models (MLM). ELECTRA trains a model with a generator that works like MLM and a discriminator that is in charge of recognizing the tokens that the generator replaces. ELECTRA has been shown to be effective in a variety of NLP areas, including sentiment/emotion analysis (e.g., [48], [4]), fake news analysis (e.g., [19]), text mining (e.g., [31], and text mining (e.g., [31]. In [32] the domain we are concentrating on in this study, it also displays high performance in cyberbullying-related works. The use of ELECTRA for implementing automated systems in the Bangla language isn't that familiar yet.

Although English is getting a lot of studies on reading comprehension and question-answer-based systems, we notice Bengali is still far behind. The use of modern technologies such as BERT and ELECTRA in Bangla Reading comprehension is not highly available yet.

1.1. Research objectives

In this work, we have implemented a BERT-based framework for indicating the answers in Bangla RC, and this implementation provides the highest performance of other existing models. We also use another transformer architecture, ELECTRA, for this solution. To our best knowledge, no architecture has been implemented based on ELECTRA.

The main contribution of this work is to bring out a workable and automated RC solution using deep architectures without designing any inference unit or knowledge base. Some key points of this research have been summarized below:

- To bring out an extensive solution for the Bangla Reading Comprehension field.
- Implementation of Transformer-based (BERT and ELECTRA) framework that successfully predicts the answers for given passages and corresponding questions.
- To measure the outcome of our work with appropriate metrics like Accuracy and Loss.

1.2. Paper outline

We discuss various research works in the following Section 2 related to our work. We try to sum up their methodology and differentiate it from ours. Section 3 carries everything about preliminary concepts and our proposed framework. In Section 4, we have mentioned our Experimental Setup. Then the following Section 5 discloses our findings and results. Next, Section 6 is an explicit discussion of our research, and it summarizes our overall contribution. Lastly, Section 7 concludes the work and mentions our future research plans.

2. Related work

Reading comprehension(RC) is attracting a lot of interest from researchers at present. Many new works are being done in this field as well as existing issues are also being picked out to upgrade. In their paper, Zhou et al. [52] addressed the over-confidence and over-sensitivity issues in current RC models. Their experiment demonstrated that it improves the robustness of reading comprehension models. Bajgar et al. [9] proposed to move to a larger data set and, as a step toward it, proposed a new data set, the Book Test, which is similar to the Children's Book Test (CBT).

Lu Chen et al. [15] developed a model for answering inquiries about a website, as well as a data set called WebSRC. Their current work is restricted to a few standard sorts of inquiries and responses, and they still can not make use of the vast information available about the websites. Therefore we can see that many studies are being done on reading comprehensions, some of which are already fascinating and others are still improving. However, the issue persisted that the majority of the focus is still on English and the low-resource languages are still in their infancy.

The range of research on question-answer models and domains is getting wider continuously. In their paper, Chen et al. [14] proposed using Wikipedia as the knowledge source to get to grips with open domain question answering. Their perspective combines search equipment based on bigram hashing and TF-IDF matching along with a multilayer Recurrent neural network (RNN) model. Roemmele et al. [36] proposed a system that generates automated questions from a given paragraph and uses State-of-art-model to find the answer to the question and present it in a human-like manner. Stroh et al. [41] implemented three deep learning models for QA tasks. Firstly, they baseline GRU model, which works well for one-word answers; secondly, they developed a better model developed by themself named Dynamic memory networks. Lastly, they implemented a simpler model of End-to-end memory networks. Annamoradnejad et al. [5] attempted to automate the moderating of QA websites. Its goal was to develop a model that could predict 20 subjective or quality elements of questions on QA websites. Even though their model attained a value of 0.046 after two epochs of training, it did not improve much in subsequent epochs, as measured by Mean-Squared-Error (MSE). This revealed our readiness to integrate QA systems into our daily lives. Education, business, and others will improve their game because of these practical and strong QA system concepts.

BERT is a commonly used model that is preferred for its efficiency and higher capability. Numerous works can be found that apply it [7]. Li et al. [27] in their paper on sentiment analysis, investigated contextualized embedding's modeling power from pre-trained models. They showed that their BERT-based architecture with a simplistic linear classification layer outperforms state-of-the-art results. Xue et al. [49] proposed a focused attention model that can be used for the relation extraction task and the joint entity. It combined a BERT language model into collaborative learning along with a dynamic range attention mechanism. And Liu et al. [29] utilized BERT to acquire contextualized

portrayal. It showed an average of 1.65% improvement than previously used BiLSTMs and CNNs. The utilization of this modern technology in the Bangla Natural Language does not become as familiar as in English. Utka et al. [46] show their efforts to improve Latvian SA for tweets. They employed a model of pretrained multilingual Bidirectional Encoder Representations from Transformers to improve the performance of SA for Latvian tweets (mBERT). They also explored further by pretraining the model using data from within the domain.

ELECTRA is used to pretrain transformer-based networks using a small amount of computing power than BERT. Butala et al. [12] provided a method for fine-tuning a pre-trained language model via parameter sharing to predict empathic concern and personal discomfort. They described how they used information from pre-trained language models for Track specific Tasks in their system entry. Pericherla et al. [32] tested the effectiveness of word embedding approaches on two classifier algorithms: Logistic regression and LightGBM, to see how well they could detect cyberbullying. These are all brilliant implementations of Electra, but it is still new to work in Bangla natural language processing.

Nowadays, the use of transformer-based architectures is in progress in the Bangla language. Singh et al. used transformer-based architectures for recognizing the Multilingual Complex Named Entity for Hindi and Bangla languages [39]. In [24], authors have introduced Bangla BERT a monolingual model. Another use of BERT can be seen in [8], where they use this architecture for classifying abusive comments.

As with other low-resource languages, research on the Bengali question-answer system still has not drawn enough attention. There is still no Bengali data set prepared. Mayeesha et al. [43] translated SQuAD 2.0 and utilized state-of-the-art transformer models like BERT, DistilBERT, and RoBERTa and trained a system on it. They focused more on the translation of the dataset rather than the classification. Uddin et al. [45] provided an end-to-end methodology for automating question answering that only answers the automated questions. They also did not employ BERT or other fine-tuned models, which may have improved their results. Banerjee et al. [10] attempted to build a factoid QA system for the Bengali language and named it BFQA. They accepted questions in natural language, then found answers from a certain document and proposed an answer ranking system for determining the best response. However, the system's accuracy was not comparable to that of European languages, as evidenced by the trials. Saha et al. [37] proposed BERT-Bangla, a language model pre-trained on a large amount of Bangla text. It was a context-aware QA system. They have a 73.9 percent accuracy rate. Two different papers evolved automated context-based Bengali Question Answering systems. One of them, Keya et al. [23] utilized the seq2seq Long short-term memory (LSTM) model to generate the answer. They used a context for creating questions, but the context was too simple, consisting of only one line. Another author Bhuiyan [11] used Bidirectional LSTM with an attention mechanism for the same task. Their context was straightforward, containing only a single line similar to the question. The scarcity of studies in this area for the Bangla Language is not covered yet. Still, there required more research attention here.

3. Proposed methodology

In this section, we will discuss the proposed methodology and the concepts relevant to it.

3.1. Preliminary concepts

Before explaining our proposed framework, we have discussed some preliminary concepts here.

3.2. Transformer-based learning

Transformer-based learning [47] is an astounding change for the popular and utilitarian artificial intelligence field NLP. This architecture

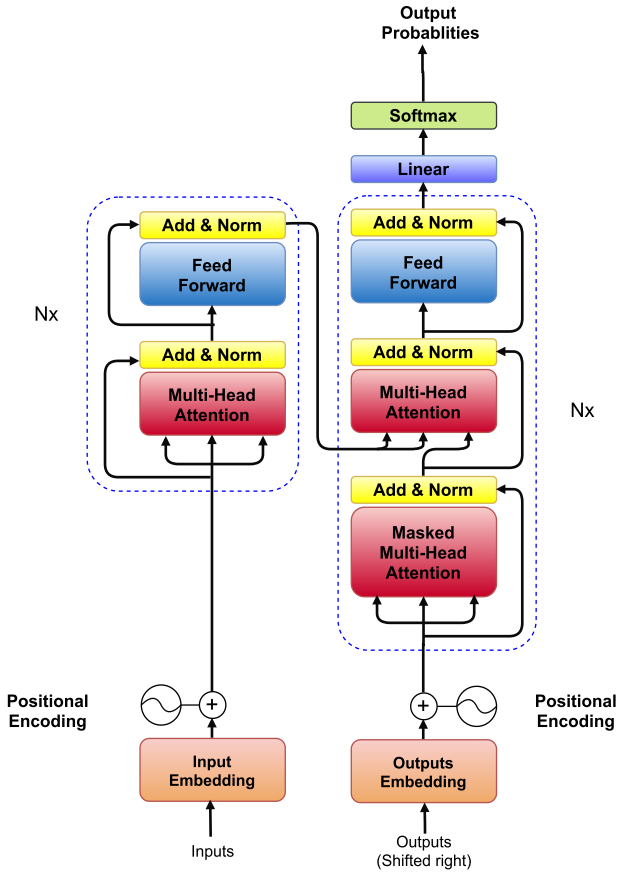

Fig. 1. The Transformer - model architecture. (This Figure's left and right halves sketch how the encoder and decoder of the Transformer respectively work using point-wise fully connected layers with stacked self-attention.)

uses the attention mechanism together with the encoder and decoder to handle sequential inputs. With auto-regressive steps the encoder in transformer maps input sequence $(x_1,..,x_n)$ to an uninterrupted representation z $(z_1,..,z_n)$, which follows to the output sequence $(y_1,..,y_m)$ through the decoder. Fig. 1 represents the architecture of the Transformer.

Two essential parts of transformers are:

Encoder and Decoder Stacks: Both Stacks consist of $N = 6$ layers with 2 sublayers. Two sublayers materialize the multi-head self-attention mechanism and a position-wise fully connected feed-forward network, respectively. The output of sublayers is *LayerNorm (x + Sublayer (x))* with the dimension $d_{model} = 512$, for the sublayer function, *Sublayer (x)*.

Multi-Head Attention: The multi-head self-attention mechanism in the transformer has three different uses. The decoder forwards queries to the next, and the encoder's output produces memory keys and values in attention layers. In encoder self-attention layers, all queries, keys, and values are generated from the output of the previous layer's encoder. Finally, after masking out softmax's input, the decoder maintains the auto-regressive property in the center of scaled dot-product attention. Here attention is computed for input consisting of queries and keys in dimension d_k and values in dimension d_v. For queries, keys, and values packed in matrix Q, K, and V, the attention is in Equation (1) and (2)

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_K}})V \tag{1}$$

$$\begin{aligned}\text{MultiHead}(Q,K,V) &= \text{Concat}(\text{head}_1,...,\text{head}_h)W^O \\ \text{where, head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned} \tag{2}$$

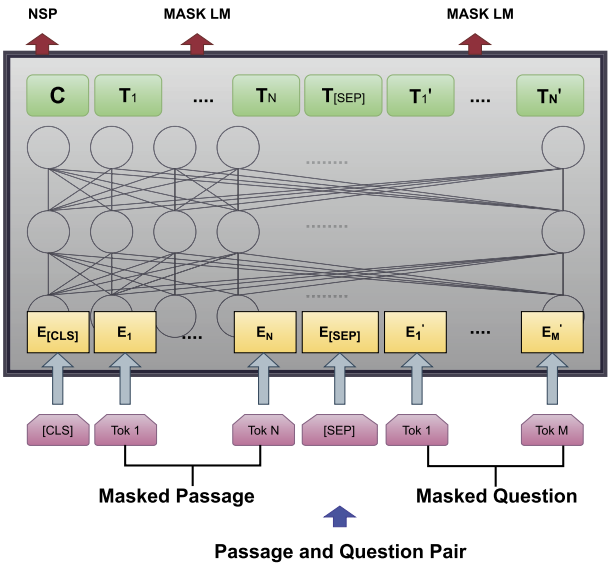

Fig. 2. BERT Architecture (BERT architecture has two steps Pretraining and Fine-tuning).

$W_i^Q \in \mathbb{R}^{d_{model} \times d_K}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_K}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_V}$ and $W^O \in \mathbb{R}^{hd_V \times d_{model}}$

3.3. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a multilayered bidirectional transformer encoder [20]. Input for BERT is an unambiguous token sequence that can contain one sentence or a couple of sentences. BERT has two steps, and they are:

- **Pre-training BERT:** BERT is pretrained on two unsupervised tasks called Masked LM (Language Model) and NSP (Next Sentence Prediction). Masked LM is about masking some random tokens and predicting them in order to get the pre-trained bidirectional model. NSP is about predicting the next sentence for a sentence pair. NSP is useful when two input sentences occur, and it is about understanding the relationship between the sentence pair. BERT has been pre-trained with BooksCorpus (800M words) [54] and English Wikipedia's text passages (not list, headers, or Tables) (2,500M words).
- **Fine-tuning BERT:** BERT For both single and coupled sentences, BERT is accredited for various downstream tasks, which enable choosing proper inputs. It initializes with pre-trained parameters, and all these can be fine-tuned with labeled data in downstream tasks.

Fig. 2 represents the BERT architecture, Where two input corpus are given to the classifier.

From Fig. 2 we can observe that for RC-based question answering tasks, input passage and questions are forwarded as a single packed sequence in the classifier. Two embeddings A and B are used by passage and questions, respectively. Here is inaugurated a start vector $S \in \mathbb{R}^H$ and also an end vector $E \in \mathbb{R}^H$ for Fine-tuning BERT. For a word i to be the starting of the answer is identified with the dot product of input tokens T_i and S. Before that, it followed the softmax over paragraph's all words by:

$$P_i = \frac{e^{S.T_i}}{\sum_j e^{S.T_j}} \tag{3}$$

Equation (3) is used for answer span's end and $S.T_i + E.T_j$ represents the candidate span's score from position i to position j. While $j \geq i$ showed prediction, the maximum scoring span is found. The veracious start and end positions' sum of log-likelihoods indicates the training objective.

3.4. Multilingual BERT (mBERT)

mBERT [28] is a BERT architecture that is pre-trained with 104 languages, including Bangla. For the classification of different languages, mBERT is trained with 10,000 sentences of each language. These sentences are collected from Wikipedia and contain at least 20 characters. From these data, 5000 are used for validation, and another 5000 are for testing purposes. It can distinguish between language-neutral components and language-specific components. Some probing tasks evaluated in mBERT are Language Identification, Language Similarity, Parallel Sentence Retrieval, Word Alignment, and Machine Translation.

3.5. Efficiently learning an encoder that classifies token replacements accurately (ELECTRA)

ELECTRA [17] has been proven as one of the most potent transformers architecture for its smaller architecture and higher performance. The working concept of ELECTRA is very close to Generative adversarial networks (GANs) and consists of a Discriminator and Generator. It is not a GAN-type architecture as, unlike GANs, the generator does not intend to escalate the discriminator loss and behave like Mask Language Model.

Generator G and Discriminator D both primarily have an encoder which that encodes the input $x = [x_1, \ldots\ldots, x_n]$ into a vector $h(x) = [h_1, \ldots\ldots, h_n]$ which is a contextual representation. For any t generator provide the probability of generating token x_t with softmax layer.

$$p_G(x_t \mid x) = \mathbf{exp}(e(x_t)^T h_G(x)_t) / \sum_{x'} \mathbf{exp}(e(x')^T h_G(x)_t) \tag{4}$$

In Equation (4), e is the token embeddings. The generator is trained like Mask Language Model. For input sequence $x = [x_1, \ldots\ldots, x_n]$, generator masked out tokens $m = [m_1, \ldots\ldots, m_k]^3$ at random position. These selected tokens are replaced with [MASK] using $x^{\text{masked}} = \text{REPLACE}(x, m, [\text{MASK}])$. The generator also can predict the original identity of these tokens.

Now the discriminator is ready to predict whether the generated token is real or not. It uses the sigmoid function as follows in Equation (5):

$$D(x,t) = \mathbf{sigmoid}(w^T h_D(x)_t) \tag{5}$$

The functionality of the discriminator is to differentiate replaced tokens by the generator. If generators $x_{corrupt}$ replace a masked-out-token, the discriminator tries to which $x_{corrupt}$ matches the input x constructed by Equation (6), (7):

$$\begin{aligned} m_i &\sim \text{unif}\{1, n\} \text{ for } i = 1 \text{ to } k \\ x^{\text{masked}} &= \text{REPLACE}(x, m, [\text{MASK}]) \end{aligned} \tag{6}$$

$$\begin{aligned} \hat{x}_i &\sim p_G(x_i \mid x^{\text{masked}}) \text{ for } i \in m \\ x^{\text{corrupt}} &= \text{REPLACE}(x, m, \hat{x}) \end{aligned} \tag{7}$$

The loss functions are in Equation (8) and (9)

$$\mathcal{L}_{\text{MLM}}(x, \theta_G) = \mathbb{E}\left(\sum_{i \in m} -\log p_G(x_i \mid x^{\text{masked}}) \right) \tag{8}$$

$$\begin{aligned} \mathcal{L}_{\text{Disc}}(x, \theta_D) = \mathbb{E}\Bigg(\sum_{t=1}^{n} -\mathbb{1}(x_t^{\text{corrupt}} = x_t) \log D(x^{\text{corrupt}}, t) - \mathbb{1} \\ (x_t^{\text{corrupt}} \neq x_t) \log(1 - D(x^{\text{corrupt}}, t)) \Bigg) \end{aligned} \tag{9}$$

Fig. 3 represents the token generation and discrimination process of ELECTRA.

3.6. Proposed framework

Our proposed framework has been delineated in Fig. 4. This framework is consisted of three steps. They are discussed below:

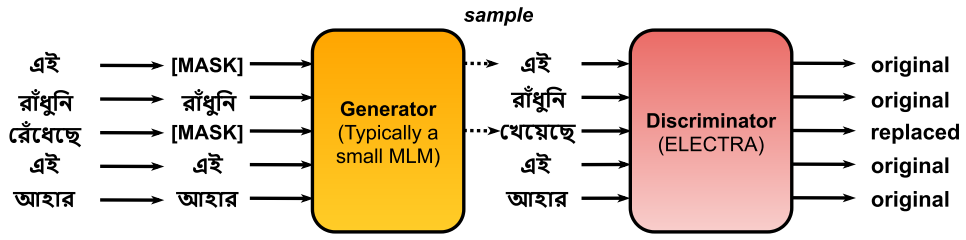

Fig. 3. ELECTRA: Replaced Token detection and Generation.

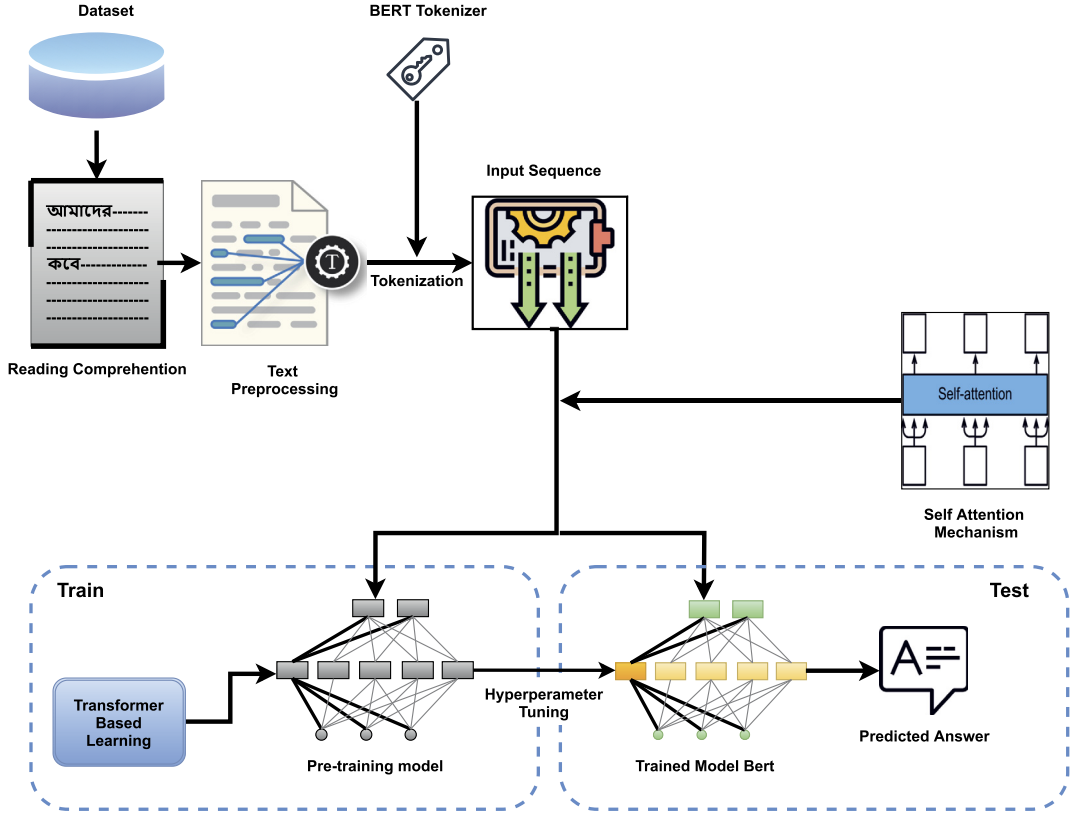

Fig. 4. Proposed Framework for Bangla RC. (Here sketch the working procedure of the proposed model, starting from data preprocessing to predicting answers of RC.)

3.6.1. Attention based input representation

At first, we process our reading comprehensions and tokenize the passages, questions, answers in order to create our input sequences. In the classifier, their two sequences are passes. The first one is the input sequence, and the second one is the attention mask.

Input Sequence: The first token in the input sequence is the default [CLS] token. Then the tokenized passage and question with default token [SEP] between them are added to it. Lastly, token [PAD] was added to pad the sequence.

[CLS] $[T_1^p]$ ..... $[T_n^p]$ [SEP] $[T_1^q]$ ..... $[T_m^q]$ [PAD]..

The previous sequence's embeddings are the final input sent to BERT. Here $[T_1^p]$.....$[T_n^p]$ and $[T_1^q]$.....$[T_n^q]$ are the tokens for passage and questions, respectively. The attention mechanism and positional encoding are applied to the input and passed to the classifiers.

3.6.2. Fine-tuning proposed BERT and ELECTRA models for downstream task

After the creation of the input sequence and attention mask, The BERT classifier is trained with the input sequence, attention mask, and the answers. The answer is also tokenized using the BERT and ELECTRA tokenizer. The proposed BERT model learns about input sequence with pretraining tasks MLM and NSP, and ELECTRA does the same by discriminating between original and replaced tokens by the generator. With the utilization of the pre-trained transformers, the models are fine-tuned on downstream task reading comprehension on our dataset. Here the proposed BERT model is an mBERT architecture utilized for Bangla Language. The proposed model has three input layers for inputting the attention mask, input ids, and token type ids. Then comes the transformer-based (BERT/ELECTRA) layers. Then the dense and flattened layers are added. Lastly, the output layer used the activation softmax function.

3.6.3. Rationalization of predicted answers

Now our classifiers are ready to predict answers for unlabeled Reading Comprehension. Finally, to justify our model's prediction quality, we use our unlabeled test dataset, which is also processed to generate input sequence and attention mask.

The algorithmic representation of our system framework is mentioned in Algorithm 1. The input of this algorithm are passages P_{test} and questions Q_{test} and the output is the corresponding answers of the given passages and questions A_{test}. In lines 1 to 8, the test process of the system is reflected using the process *PREDICTRC* which is described in lines 9 to 22. In this function *TrainModel* is the BERT/ELECTRA classifier. The *Preprocess* indicates the data preprocessing mentioned in section 4.4

Algorithm 1 Algorithmic representation.

Input: $RC_{test} = (P_{test}, Q_{test})$
Output: A_{test} = answers of P_{test} *and* Q_{test}

$i \leftarrow 0$
while $i < P_{test}.length$ **do**
  $P_m \leftarrow Preprocess(P_{test}[i])$
  $Q_m \leftarrow Preprocess(Q_{test}[i])$
  $X_{test} \leftarrow Tokenize(P_{test}, Q_{test}$
  $A_{test}[i] \leftarrow PredictRC(X_{test})$
  $i \leftarrow i + 1$
end while
procedure PREDICTRC(*Output*)
  $i \leftarrow 0$
  while $i < P_{train}.length$ **do**
    $P_{train} \leftarrow Preprocess(P_{train}[i])$
    $Q_{train} \leftarrow Preprocess(Q_{train}[i])$
    $A_{train} \leftarrow Preprocess(A[i])$
    $X_{train}[i] \leftarrow Tokenize(P_{train}, Q_{train})$
    $Y_{train}[i] \leftarrow Tokenize(A_{train})$
    $i \leftarrow i + 1$
  end while
  $TestModel \leftarrow TrainModel(X_{train}, Y_{train})$
  $Output \leftarrow TestModel(X)$
  Return Output
end procedure

Table 1. Model parameters and values.

Hyperparameters	BERT	ELECTRA
learning rate (AdamW)	2e-04	5e-04
max_len	484	512
Batch_size	32	32
verbose	1	1
epoch	40	40

Our proposed methodology has produced a significant result for Bangla Reading Comprehension. It has the capability to handle very long and complex real-world passages and questions than the existing works. Besides, it provides significantly higher accuracy and reduced loss for unlabeled data samples.

4. Experimental setup

4.1. Experimental environment

Deep Learning models require high-end configurations for the purpose of parallel processing. Therefore we have maneuvered Google Colab [3, 13]. It is a Jupyter notebook platform based on cloud computing and provides necessary options for utilizing GPU and TPU. It is workable under Ubuntu OS with Tesla k-80 GPU of NVIDIA accompanied by 12 GB of GPU memory. It imparted python runtime and other required pre-configured libraries and packages to run deep learning tasks.

4.2. Hyperparameter tuning

Hyperparameters influence the weight initialization and data order. Thus finding the most significant values for hyperparameters benefits our model to predict accurately. Table 1 proclaims the most suitable values for our classifier. The most significant hyperparameters for a transformer-based model are its learning rate, batch_size, max_seq_length, epoch, etc. For a simple transformer, the values are learning rate = 4e-4, batch_size = 8, max_seq_length = 128 and epoch = 1. It by default uses the AdamW [30] optimizer.

In our proposed classifier we tune these hyperparameters as learning rate = 2e-4, batch_size = 32, max_seq_length = 484 and epoch = 40 for the BERT model. For ELECTRA these values are learning rate = 5e-4, batch_size = 32, max_seq_length = 512 and epoch = 40. We keep the optimizer as default. Choosing verbose = 1 helps us watch our output's progress easily. This hyperparameter tuned model is well performed than the baseline model.

4.3. Dataset composition and provision

There is no obtainable dataset for Bangla Reading Comprehension Tasks; this research field has gained less attention. To conduct noble research, we require a sufficient amount of data for the evolution of methodology. Hence we have collected our data for the experiment. We collect long 'Passages' from different significant Bangla writings on the internet and develop a Reading Comprehensions dataset with 3636 samples. In Table 2 we mentioned some samples of our data set. These data contain passages and questions as inputs and answers as outputs. Besides sample data, we also mention the English Translation of sample data. We translated the passages, questions, and answers manually. This Translation indicates the difference between Bangla and English text.

The maximum length for a passage, question, and answer are 359, 108, and 7, respectively. The largest passage and question pair length is 409. These values are counted after the tokenization of the corpus. We have manifested 20% of our data for testing and other data for training. 2908 Reading Comprehensions are used to train the BERT and ELECTRA classifier, and 728 are preserved for testing purposes.

4.4. Data prepossessing

Raw data comes with unnecessary characters and words. These may act as a barrier during classification. Therefore we don't put our data directly into the classifier. We preprocessed data and applied it [1, 2]. The following preprocessing steps are taken in order to bring out the most significant results:

- Besides words, the Raw data consists of many characters (e.g. ., $, %, #, *, -, etc.), which probably emanate a decreasing accuracy. Therefore we remove these characters from our corpus.
- Also, there are many Bangla stop words[1] in data that have no contribution to prediction tasks. Moreover, these words can be a barrier to higher accuracy. The abolition of these stop words is proven supportive of our accuracy.
- Bangla words exist in different forms. E.g. the word 'ধর' can be formed as 'ধরা', 'ধরে', 'ধরেন', 'ধরেছেন', 'ধরবেন', 'ধরছে', 'ধরবে', 'ধরব', etc. So we apply stemming and lemmatization and use the root word to process the corpus.

Below we have brought up the raw data and preprocessed data. This preprocessed data performed better than the raw data.

- **Raw Data**
  বাংলাদেশ একটি ক্ষুদ্র আয়তনের জনবহুল দেশ। এ দেশের আয়তন ১,৪৭,৫৭০ বর্গকিলোমিটার। এ দেশের মাটে
  আয়তনের ১৭ শতাংশ বনভূমি। বাংলাদেশের রাজধানীর নাম ঢাকা। বাংলাদেশে মাটে আটটি বিভাগীয় শহর রয়েছে। (Translation mentioned in Table 2)
- Preprocessed Data
  বাংলাদেশ, ক্ষুদ্র, আয়তন, জনবহুল, দেশ, দেশ, আয়তন, ১৪৭৫৭০, বর্গকিলোমিটার, দেশ, আয়তন, ১৭, শতাংশ,
  বনভূমি, বাংলাদেশ, রাজধানী, নাম, ঢাকা, বাংলাদেশ, আট, বিভাগীয়, শহর

5. Experimental evolution

5.1. Proposed model's performance

After the training process of our classifiers, we applied unlabeled test data and brought up some results based on the prediction of answers. We determine the accuracy and loss for evaluation. The predicted output from test input and the test outputs are compared for accuracy

[1] https://github.com/stopwords-iso/stopwords-bn.

Table 2. Dataset Sample and English Translation. (Each observation contains a passage, a question generated from the passage, and their corresponding Answer.)

	Passage	Question	Answer
Bangla Text	বাংলাদেশ একটি ক্ষুদ্র আয়তনের জনবহুল দেশ। এ দেশের আয়তন ১,৪৭,৫৭০ বর্গকিলোমিটার। এ দেশের মাটে আয়তনের ১৭ শতাংশ বনভূমি। বাংলাদেশের রাজধানীর নাম ঢাকা। বাংলাদেশে মাটে আটটি বিভাগীয় শহর রয়েছে।	বাংলাদেশের আয়তন কত বর্গকিলোমিটার?	১,৪৭,৫৭০
English Translation	Bangladesh is a small country with a large population. The area of the country is 1,46,570 sq km. Seventeen percent of the country's total area is forest land. The capital of Bangladesh is Dhaka. There are a total of eight divisional cities in Bangladesh.	What is the area of Bangladesh?	1,47,570
Bangla Text	আইনস্টাইনের স্পেশাল রিলেটিভিটি থিওরি প্রকাশের আগে ধরে নেওয়া হয়েছিল সময় এই মহাবিশ্বের একটি ধ্রুব বৈশিষ্ট্য। আইনস্টাইন দেখালেন, সময় ধ্রুব, ধারণাটি একটি বিভ্রম। অন্য মাত্রার মতো সময়ও আপেক্ষিক। কিন্তু এই আপেক্ষিকতা অনুভব করার মতো প্রযুক্তি বা সামর্থ্য কোনোটাই আমাদের নেই। এই কারণে সময়ের বেঁধে দেওয়া একটি নির্দিষ্ট গতিতে প্রতিনিয়ত ভবিষ্যতের দিকে হেঁটে চলেছি আমরা। আইনস্টাইন প্রমাণ করলেন, এই গতিকে বহুগুণে বৃদ্ধি করা যাবে যদি আমরা আলোর কাছাকাছি বেগ অর্জন করতে পারি। অর্থাৎ, তখন সময়ের বেঁধেদেওয়া গন্ডির মধ্যে আর মাথানিচু করে আমাদের চলতে হবে না।	কে প্রমাণ করলেন, গতিকে বহুগুণে বৃদ্ধি করা যাবে যদি আমরা আলোরে কাছাকাছি বেগ অর্জন করতে পারি?	আইনস্টাইন
English Translation	Before Einstein's special relativity theory, it was assumed that Time is a constant nature of the Universe. Einstein shows, 'Time being constant is an illusion.' Like other dimensions, Time is also relative. But to feel this relativity, we have neither such as technologies nor the ability. That's why every moment, we are walking towards the future with the velocity fixed by Time. Einstein proved that speed can be multiplied if we can achieve velocities close to the light. That means we need not follow the boundaries fixed by Time anymore. Within a few moments, we can travel to the future after several centuries.	Who proved that speed could be multiplied if we could achieve velocity closer to light?	Einstein
Bangla Text	পদ্মা বাংলাদেশের প্রধান বৃহত্তম নদী। এর দৈর্ঘ্য প্রায় ৩৬৬ কিলোমিটার। ইহা হিমালয়ের হিমবাহ থেকে উতপত্তি লাভ করে বাংলাদেশে পদ্মা নদী নামে। তারপর পদ্মা নদী গায়োলন্দে গিয়ে মিলিত হয়েছে যমুনার সাথে । মিলনের এই ধারা অব্যাহত রেখে মেঘনার সাথে আবার মিলিত হয় চাঁদপুরে এবং শেষমেষ মেঘনা নামেই বঙ্গোপসাগরে পতিত হয়েছে।	কাথোয় পদ্মা গিয়ে মিলিত হয়েছে যমুনার সাথে?	গায়োলন্দ
English Translation	The Padma is the largest river in Bangladesh. Its length is about 366 kilometers. Originating from the glaciers of the Himalayas, it has flowed over an the extended area inside of India and entered Bangladesh as the Padma River. Then the river Padma goes to Goalanda and joins the Jamuna. Continuing this trend of reunion, it reunited with Meghna at Chandpur and finally fell into the Bay of Bengal as Meghna.	Where did the Padma meet the Jamuna?	Goalanda

calculation. The loss function used here is SparseCategoricalCrossentropy.[2]

In Table 3 we summed up the accuracy for different values of hyperparameters. We ran the model for five different learning rates (1e-4, 2e-4, 3e-4, 4e-4 and 5e-4), three different batch sizes (12, 24 and 32) and for two different values of max_ sequence_length (128 and 484). The best accuracy was obtained for BERT, where the learning rate is 1e-4, the batch size is 32, and the max_ sequence_length is 484. In Table 3, we bolded out the best accuracy for easy visualization. We also show the different accuracy for ELECTRA with the change of hyperparameters in Table 4. Electra provides the best accuracy at 86.5 for the learning rate 5e-4 with batch size 32 and max_len 512.

To evaluate the performance of our classifier, we have broached two types of evolution metrics Accuracy and loss. Accuracies and Losses indicate the model's behavior over epochs. In Fig. 5, we plotted our training and testing accuracy of the BERT model. Here, the training and testing accuracy have been plotted using red and blue lines, respectively. It is noticeable that the training accuracy is 98.54% which means our classifier has been constructed properly. To evolute the classifier's prediction, we plotted testing accuracy, where we achieved the highest 87.78% accuracy for unlabeled testing data. We calculated both accuracies for 40 epochs, and these values are very significant for any Bangla Reading Comprehension System.

In Fig. 6, we plotted the training and validation loss with red and blue lines, respectively. Training and validation losses are reduced significantly here. We decrease validation loss by .12 over 40 epochs, and the training loss is almost near zero in some epochs.

In the same fashion, we plotted both training and testing accuracy for ELECTRA in Fig. 7. The training and testing accuracy for ELECTRA is determined as 93.0% and 82.52%. The training and testing losses for ELECTRA are shown in Fig. 8 where minimum training and testing loss are .24 and .49.

We determine the True Positives (TP), False Positives (FP), and False Negatives (FN) for the predicted answers.[3] True Positives are the number of tokens that are common in both predicted answers and the ground truth of answers. False Positives indicate the predicted tokens that are not in the test data's output, and False Negatives are the tokens from ground truth's answers that the classifier can't predict. Then we determine the Precision, Recall using the following Equation (10) and (11) with TP, FP, FN.

[2] https://keras.io/api/losses/probabilistic_losses/#sparse_categorical_crossentropy-function.

[3] https://kierszbaumsamuel.medium.com/f1-score-in-nlp-span-based-qa-task-5b115a5e7d41.

Table 3. Accuracy for the proposed BERT model for different hyperparameter values.

Learning Rate	Batch Size	max_len	Accuracy (%)
1e-4	12	128	83.0
		484	83.5
	24	128	80.7
		484	79.9
	32	128	84.2
		484	86.0
2e-4	12	128	86.5
		484	86.3
	24	128	86.1
		484	86.8
	32	128	85.4
		484	**87.78**
3e-4	12	128	76.7
		484	78.0
	24	128	80.9
		484	81.0
	32	128	83.0
		484	84.8
4e-4	12	128	84.1
		484	85.6
	24	128	83.0
		484	83.9
	32	128	86.1
		484	86.7
5e-4	12	128	82.1
		484	79.9
	24	128	80.3
		484	81.0
	32	128	81.4
		484	82.5

Table 4. Accuracy for the proposed ELECTRA model for different hyperparameter values.

Learning Rate	Batch Size	max_len	Accuracy (%)
1e-4	12	484	78.0
		512	80.5
	24	484	77.9
		512	81.6
	32	484	80.8
		512	81
2e-4	12	484	79.3
		512	81.0
	24	484	79.2
		512	80.8
	32	484	80.0
		512	80.0
3e-4	12	484	78.2
		512	75.7
	24	484	82.0
		512	79.7
	32	484	81.2
		512	81.1
4e-4	12	484	81.0
		512	81.9
	24	484	79.0
		512	80.0
	32	484	81.2
		512	80.3
5e-4	12	484	82.1
		512	79.8
	24	484	81.5
		512	79.8
	32	484	81.6
		512	**82.52**

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

Again determined precision and recall are used for identifying the F1 Score with Equation (12).

$$F1Score = \frac{2 * precision * recall}{precision + recall} \tag{12}$$

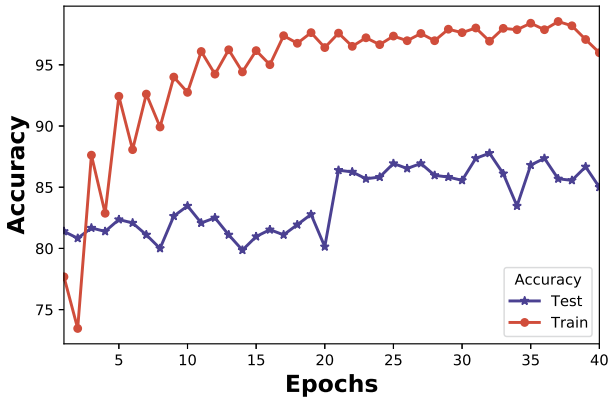

Fig. 5. Accuracy of our proposed BERT model. (Accuracy indicates the model's right predictions for unlabeled testing data.)

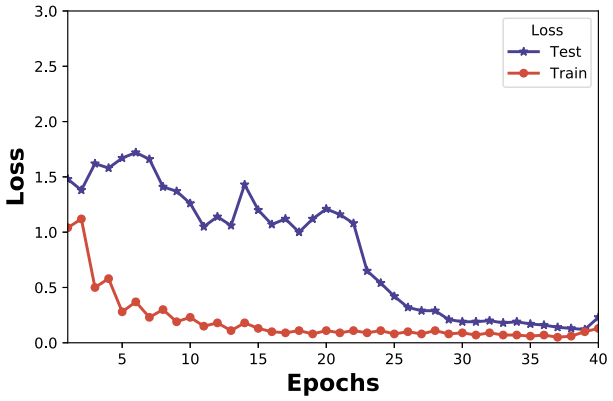

Fig. 6. Loss of our proposed BERT model. (Loss indicates the penalty for wrong predictions for the model.)

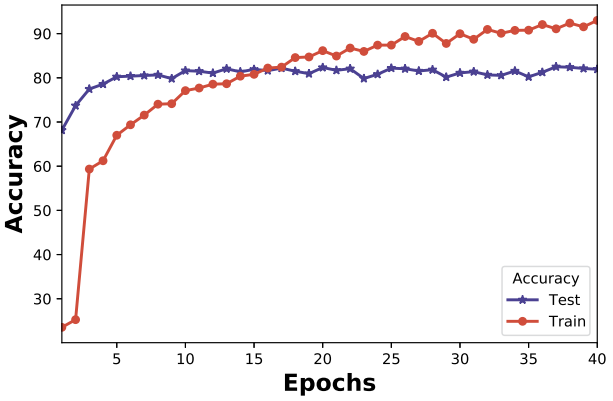

Fig. 7. Accuracy of our proposed ELECTRA model. (Accuracy indicates the model's right predictions for unlabeled testing data.)

Finally, all these values of TP, FP, FN, Precision, Recall, and F1 score are mentioned in Table 5.

5.2. Comparison with existing models

For the justification of our proposed methods' state-of-the-art performance, we applied some other techniques to our data. First of all, we applied our data to the simple RNN model. For that, we combined the passage and question in a GRU (Gated recurrent units) cell before projecting it to the dense layer and forwarded it to the softmax layer to generate output. This model worked well only for one-word answers and provided us the lowest accuracy comparatively.

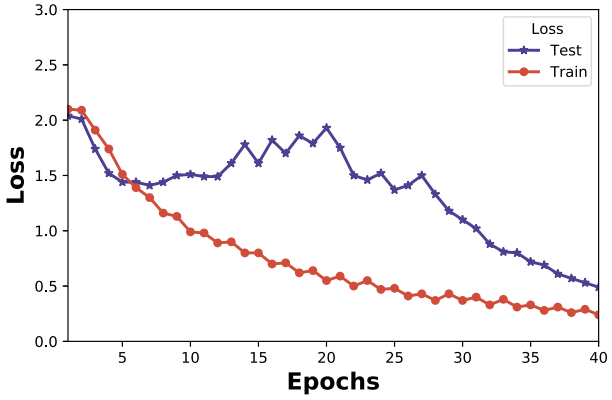

Fig. 8. Loss of our proposed ELECTRA model. (Loss indicates the penalty for wrong predictions for the model.)

Table 5. True Positive, False Positive, False Negative, Precision, Recall, and F1 Score of Proposed Bert and ELECTRA models.

Model ⟶ Metrics ↓	BERT	ELECTRA
True Positive	9575	9001
False Positive	690	1015
False Negative	643	892
Precision	93.27%	89.87%
Recall	93.71%	90.98%
F1 Score	93.49%	90.42%

Long Short Term Memory (LSTM) outperformed simple RNN. We used the softmax activation function for our Seq2Seq LSTM model. We merged the passage and question to the embedded layer and applied it to the LSTM.

Then we used the attention mechanism with bidirectional LSTM consisting of both backward and forward LSTM in our data. We calculated an attention weight from the encoder's output and determined the attention vector using the softmax. This model outperformed both LSTM and Simple RNN.

Fig. 9 shows the highest accuracy and F1 score for Simple RNN, LSTM, and bidirectional LSTM with attention and our proposed methods over 40 epochs. We can visualize that BERT outperforms other models, and ELECTRA also provides higher accuracy than these existed methods on our data.

6. Discussion

While Systems like RC-based Question Answering are becoming essential in modern autonomous Education systems, low-resource languages like Bangla are facing a deficiency of sufficient NLP research and dataset. Therefore in this paper, we try to illuminate the implementation of such an automatic system using the latest NLP technologies, BERT and ELECTRA. Among these two architectures, BERT outperformed ELECTRA, as mBERT is pretrained on the Bangla language. To solve the lacking data issue, we bring up a noble dataset for Bangla reading comprehension consisting of real-world Bangla Passages, Questions, and Answers. Then we trained our models with sufficient training data to compute the necessary weight for prediction and applied unlabeled testing data to determine how correctly the model performed the prediction. Besides, we perceived the noteworthy values of all hyperparameters of our proposed model. We identify evolution metrics accuracy and loss based on the performance of our model. The BERT model outperformed with significant 99% training accuracy and 87.78% validation accuracy. ELECTRA's performance is also remarkable in this research. Both training and testing losses of notably removed for the model. Finally, we compare our model's performance with existing methods in Section 5.2 and visualize that our model's performance is higher than others.

One significant limitation of that work is that the work is language specific. Therefore it may not provide satisfactory performance in other low-resource languages. We have used 2908 observations in our work. Increasing the amount of data can be helpful for the improvement of the performance. Another limitation of this work is that it is not tested with the synonyms of answers.

7. Conclusion and future work

The main motive of this work is to bring out an efficient and automated technique for Bangla Reading Comprehension, which can enact a progressive role in the Bangla Education system. We have developed a model that predicts the answer for any passage and reading comprehension question. To implement the model, we have utilized the latest NLP technique, the transformer-based learning BERT, which uses the self-attention mechanism and the pre-training language model for prediction.

We applied the proposed methodology in a real-world benchmark dataset entirely new to Bangla Language. This dataset can be a noble contribution to Bangla NLP. We determined a significant evaluation matrix for the vindication of the method. Our proposed framework outperforms other deep learning models for predicting answers and dispenses 87% testing accuracy and 99% training accuracy, which is remarkable for Bangla RC. It is visible that our training accuracy is higher than the testing. We intend to add more samples to our experiments to ameliorate our testing accuracy. Moreover, we want to come up with an algorithmic solution for a better accuracy performance for test data.

In the future, we want to implement this research work as an embedded system in order to develop a more efficient real-world Bangla Reading Comprehension system. We want to apply this methodology for other Reading comprehension questions like true-false, Fill in the blank, and multiple-choice questions. Moreover, we plan to expand our dataset so that we will be able to introduce an impactful data source for Bangla Reading Comprehension.

Declarations

Author contribution statement

Tanjim Taharat Aurpa: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Richita Khandakar Rifat: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Md Shoaib Ahmed; Md. Musfique Anwar: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

A. B. M. Shawkat Ali: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data associated with this study has been deposited at https://data.mendeley.com/datasets/s9pb3h2cjy/1.

Declaration of interests statement

The authors declare no conflict of interest.

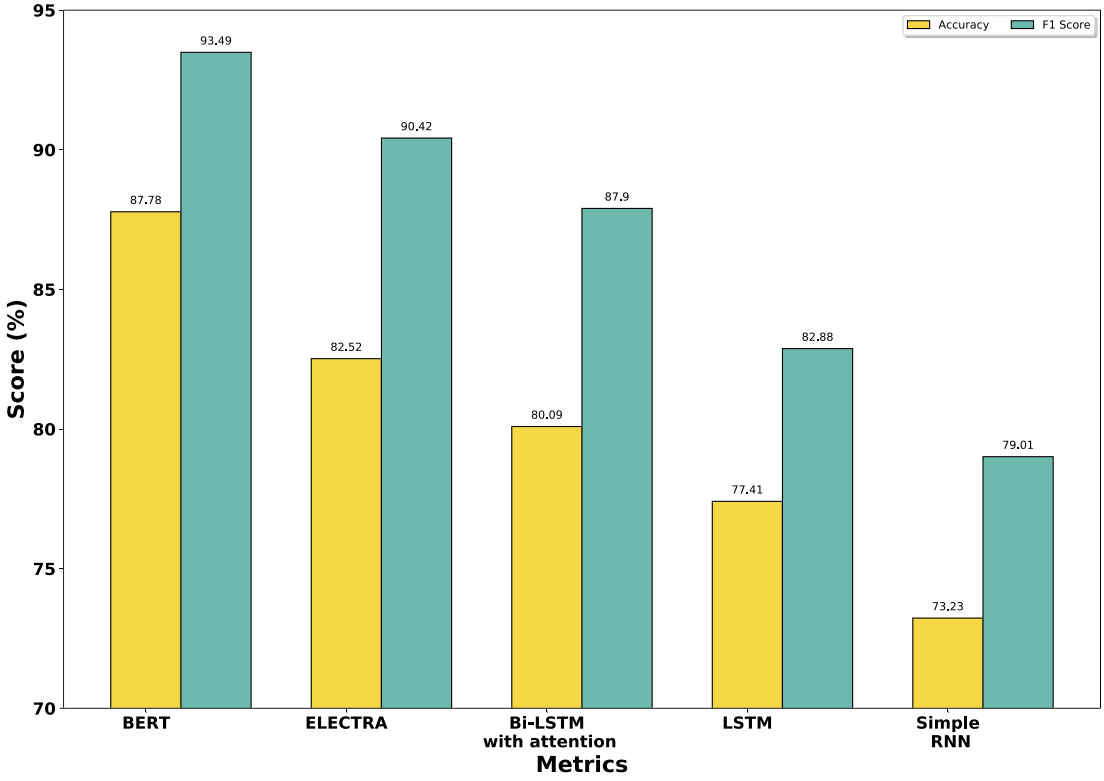

Fig. 9. Comparing Classifiers with their Accuracy and F1 scores. (A comparison of the Accuracy and F1 scores of several deep neural network architectures have been sketched here.)

Additional information

No additional information is available for this paper.

References

[1] M.S. Ahmed, T.T. Aurpa, M.M. Anwar, Online topical clusters detection for top-k trending topics in Twitter, in: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2020, pp. 573–577.

[2] M.S. Ahmed, T.T. Aurpa, M.M. Anwar, Detecting sentiment dynamics and clusters of Twitter users for trending topics in COVID-19 pandemic, PLoS ONE 16 (2021) e0253300.

[3] M.S. Ahmed, T.T. Aurpa, M.A.K. Azad, Fish disease detection using image based machine learning technique in aquaculture, J. King Saud Univ, Comput. Inf. Sci. 34 (2022) 5170–5182.

[4] N. Al-Twairesh, The evolution of language models applied to emotion analysis of Arabic tweets, Information 12 (2021) 84.

[5] I. Annamoradnejad, M. Fazli, J. Habibi, Predicting subjective features from questions on QA websites using BERT, in: 2020 6th International Conference on Web Research (ICWR), IEEE, 2020, pp. 240–244.

[6] I. Ashrafi, M. Mohammad, A.S. Mauree, G.M.A. Nijhum, R. Karim, N. Mohammed, S. Momen, Banner: a cost-sensitive contextualized model for Bangla named entity recognition, IEEE Access 8 (2020) 58206–58226.

[7] T.T. Aurpa, M.S. Ahmed, R. Sadik, S. Anwar, M.A.M. Adnan, M. Anwar, et al., Progressive guidance categorization using transformer-based deep neural network architecture, in: International Conference on Hybrid Intelligent Systems, Springer, 2021, pp. 344–353.

[8] T.T. Aurpa, R. Sadik, M.S. Ahmed, Abusive Bangla comments detection on Facebook using transformer-based deep learning models, Soc. Netw. Anal. Min. 12 (2022) 1–14.

[9] O. Bajgar, R. Kadlec, J. Kleindienst, Embracing data abundance: booktest dataset for reading comprehension, arXiv preprint, arXiv:1610.00956, 2016.

[10] S. Banerjee, S.K. Naskar, S. Bandyopadhyay, Bfqa: a Bengali factoid question answering system, in: International Conference on Text, Speech, and Dialogue, Springer, 2014, pp. 217–224.

[11] M.R. Bhuiyan, A.K.M. Masum, M. Abdullahil-Oaphy, S.A. Hossain, S. Abujar, An approach for Bengali automatic question answering system using attention mechanism, in: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2020, pp. 1–5.

[12] Y. Butala, K. Singh, A. Kumar, S. Shrivastava, Team Phoenix at WASSA 2021: emotion analysis on news stories with pre-trained language models, arXiv preprint, arXiv:2103.06057, 2021.

[13] T. Carneiro, R.V.M. Da Nóbrega, T. Nepomuceno, G.B. Bian, V.H.C. De Albuquerque, P.P. Reboucas Filho, Performance analysis of google colaboratory as a tool for accelerating deep learning applications, IEEE Access 6 (2018) 61677–61685.

[14] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading Wikipedia to answer open-domain questions, arXiv preprint, arXiv:1704.00051, 2017.

[15] L. Chen, X. Chen, Z. Zhao, D. Zhang, J. Ji, A. Luo, Y. Xiong, K. Yu, Websrc: a dataset for web-based structural reading comprehension, arXiv preprint, arXiv:2101.09465, 2021.

[16] S. Chowdhury, N. Baili, B. Vannah, Ensemble fine-tuned mBERT for translation quality estimation, arXiv preprint, arXiv:2109.03914, 2021.

[17] K. Clark, M.T. Luong, Q.V. Le, C.D. Manning, Electra: pre-training text encoders as discriminators rather than generators, arXiv preprint, arXiv:2003.10555, 2020.

[18] D. Colla, T. Caselli, V. Basile, J. Mitrović, M. Granitzer, GruPaTo at SemEval-2020 task 12: retraining mBERT on social media and fine-tuned offensive language models, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1546–1554.

[19] K.A. Das, A. Baruah, F.A. Barbhuiya, K. Dey, Ensemble of ELECTRA for profiling fake news spreaders, in: CLEF (Working Notes), 2020.

[20] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.

[21] S.E. Friedman, I.H. Magnusson, S.M. Schmer-Galunder, Extracting qualitative causal structure with transformer-based NLP, arXiv preprint, arXiv:2108.13304, 2021.

[22] H. Gonen, S. Ravfogel, Y. Elazar, Y. Goldberg, It's not Greek to mBERT: inducing word-level translations from multilingual BERT, arXiv preprint, arXiv:2010.08275, 2020.

[23] M. Keya, A.K.M. Masum, B. Majumdar, S.A. Hossain, S. Abujar, Bengali question answering system using seq2seq learning based on general knowledge dataset, in: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2020, pp. 1–6.

[24] M. Kowsher, A.A. Sami, N.J. Prottasha, M.S. Arefin, P.K. Dhar, T. Koshiba, Bangla-BERT: transformer-based efficient model for transfer learning and language understanding, IEEE Access 10 (2022) 91855–91870.

[25] J. Krishnan, A. Anastasopoulos, H. Purohit, H. Rangwala, Cross-lingual text classification of transliterated Hindi and Malayalam, arXiv preprint, arXiv:2108.13620, 2021.

[26] A. Kulkarni, M. Mandhane, M. Likhitkar, G. Kshirsagar, J. Jagdale, R. Joshi, Experimental evaluation of deep learning models for Marathi text classification, arXiv preprint, arXiv:2101.04899, 2021.

[27] X. Li, L. Bing, W. Zhang, W. Lam, Exploiting BERT for end-to-end aspect-based sentiment analysis, arXiv preprint, arXiv:1910.00883, 2019.
[28] J. Libovický, R. Rosa, A. Fraser, How language-neutral is multilingual BERT?, arXiv preprint, arXiv:1911.03310, 2019.
[29] A. Liu, Z. Huang, H. Lu, X. Wang, C. Yuan, BB-KBQA: BERT-based knowledge base question answering, in: China National Conference on Chinese Computational Linguistics, Springer, 2019, pp. 81–92.
[30] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint, arXiv:1711.05101, 2017.
[31] I.B. Ozyurt, On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining, bioRxiv, 2020.
[32] S. Pericherla, E. Ilavarasan, Performance Analysis of Word Embeddings for Cyberbullying Detection, IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2021, p. 012008.
[33] M.M. Rahman, M.A. Pramanik, R. Sadik, M. Roy, P. Chakraborty, Bangla documents classification using transformer based deep learning models, in: 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), IEEE, 2020, pp. 1–5.
[34] T. Rahman, R. Ahmed, Combatting the impact of COVID-19 school closures in Bangladesh, 2021.
[35] T. Rahman, U. Sharma, A simulation of COVID-19 school closure impact on student learning in Bangladesh, 2021.
[36] M. Roemmele, D. Sidhpura, S. DeNeefe, L. Tsou, AnswerQuest: a system for generating question-answer items from multi-paragraph documents, arXiv preprint, arXiv:2103.03820, 2021.
[37] A. Saha, M.I. Noor, S. Fahim, S. Sarker, F. Badal, S. Das, 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), IEEE, 2021, pp. 1–6.
[38] S.S. Sarkar, P. Das, M.M. Rahman, M. Zobaer, Perceptions of public university students towards online classes during COVID-19 pandemic in Bangladesh, in: Frontiers in Education, Frontiers, 2021, p. 265.
[39] S. Singh, P. Jawale, U. Tiwary, silpa_nlp at SemEval-2022 tasks 11: transformer based NER models for Hindi and Bangla languages, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 1536–1542.
[40] F. Souza, R. Nogueira, R. Lotufo, Portuguese named entity recognition using BERT-CRF, arXiv preprint, arXiv:1909.10649, 2019.
[41] E. Stroh, P. Mathur, Question answering using deep learning, 2016.
[42] J. Su, S. Yu, D. Luo, Enhancing aspect-based sentiment analysis with capsule network, IEEE Access 8 (2020) 100551–100561.
[43] T. Tahsin Mayeesha, A. Md Sarwar, R.M. Rahman, Deep learning based question answering system in Bengali, J. Inf. Telecommun. 5 (2021) 145–178.
[44] I.V. Tetko, P. Karpov, R. Van Deursen, G. Godin, State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis, Nat. Commun. 11 (2020) 1–11.
[45] M.M. Uddin, N.S. Patwary, M.M. Hasan, T. Rahman, M. Tanveer, End-to-end neural network for paraphrased question answering architecture with single supporting line in Bangla language, Int. J. Future Comput. Commun. 9 (2020).
[46] A. Utka, et al., Pretraining and fine-tuning strategies for sentiment analysis of Latvian tweets, in: Human Language Technologies–the Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020, IOS Press, 2020, p. 55.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[48] H. Xu, B. Liu, L. Shu, P.S. Yu, Dombert: domain-oriented language model for aspect-based sentiment analysis, arXiv preprint, arXiv:2004.13816, 2020.
[49] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, P. He, Fine-tuning BERT for joint entity and relation extraction in Chinese medical text, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 892–897.
[50] J. Yu, J. Jiang, Adapting BERT for target-oriented multimodal sentiment classification, in: IJCAI, 2019.
[51] S. Yu, J. Su, D. Luo, Improving BERT-based text classification with auxiliary sentence and domain knowledge, IEEE Access 7 (2019) 176600–176612.
[52] M. Zhou, M. Huang, X. Zhu, Robust reading comprehension with linguistic constraints via posterior regularization, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2020) 2500–2510.
[53] X. Zhu, Cross-lingual word sense disambiguation using mBERT embeddings with syntactic dependencies, arXiv preprint, arXiv:2012.05300, 2020.
[54] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 19–27.