End-To-End Neural Network for Paraphrased Question Answering Architecture with Single Supporting Line in Bangla Language

Md. Mohsin Uddin, Nazmus Sakib Patwary, Md. Mohaiminul Hasan, Tanvir Rahman, and Mir Tanveer Islam

***Abstract*—Recent studies on QA (Question Answering) system in English language have been emerged extensively with the composition of Natural Language Processing (NLP) and Information Retrieval (IR) by amplifying miniature sub tasks to accomplish a whole AI-system having capability of answering and reasoning complicated and long questions through understating paragraph. In our proposed study, we present a general heuristic framework, an end-to-end model used for paraphrased question answering using single supporting line which is the initial appearance ever in Bangla language. Corpus dataset was scrapped from Bangla wiki and then questions were generated corresponding context have been used to learn the model. Translated bAbI dataset (1 supporting fact) in Bangla language has been also incorporated with to experiment the proposed model manually.**

To predict appropriate answer, model is trained with question-answer pair and a supporting line. For comparing our task applying variation of basic Recurrent Neural Network (RNN): Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) different accuracy has been found. For further accomplishment, synthetic and semantic word relevance in high dimension vector space: Bangla word embedding system(word2vec) is added to the system as sentence representation along with Positioning Encoding (PE) and which outperforms both memory network GRU and LSTM precisely.

***Index Terms*—Machine learning, natural language processing, information retrieval, long short time memory, gated recurrent unit.**

I. Introduction

Intelligent question answering system that offers tasks like asking automated machine any questions and getting appropriate answer from computer automatically has been very important end-user task in recent age which ease human's life drastically. To enable communication with computer, asking question to computer is indispensable task. Many researches have been done over this particular field of question answering in near years. However, it's an important task which uses a combination of both NLP [1] and IR that shortens the distance between IR-based search and intelligent assistants that uses information extracting process from data [2]. Different answers can be yielded with a modicum variation in semantically equivalent questions. For example, questions like "who did created Microsoft" and "who did started Microsoft" yield same identity. The question answering model ought to acknowledge the answer from its knowledge base considering both questions semantically equivalent [3].

There has been a lot of research in machine learning that are intended to reasoning and intelligently answering questions. It's a comprehensive area [4]. Towards answering question and reasoning two grand challenges in intelligence system have been arrived in numerous research that make models which is able to make multiple computational steps for question answering and to make model that adopts and working ability considering long term dependencies in case of sequential data as well as unstructured data [5]. In the circumstance of semantic parsing for answering question in recent time, researchers are highly focused on complicated and long question answering [6].

Abundant number of researches has been performed considering English language related to different question answering task like search-based QA, factoid QA etc. in order to accomplish AI-complete question answering in isolated ways using end-to-end neural network. But in Bangla language no such research and task have been conducted regarding question answering which would contribute AI-complete question answering in Bangla language.

In our study we propose a system having consideration of close domain dataset and develop algorithms with variety for understanding language and paraphrased question answering. Different architecture of RNN like LSTM, GRU has been used to build satisfactory model. Question-answer pair from Bangla close domain dataset has been used as training data as well as with one support line and multiple related line. It can answer paraphrased question's answer too and questions containing 'who' ('কে'), 'where' ('কোথায়'), 'when' ('কখন '), 'what' ('কি') are answered also whereas [1], [2] considered so straight forward simple dataset that only contain questions containing 'where' ('কোথায়') for their single supporting fact category.

II. Related Works

In recent years, various research has been applied on question answering over unstructured paragraph data using end-to-end suitable deep neural network model. It's a part of NLP as well as use the sub part of information retrieval. In the paper [7], authors brought in a Recurrent neural

Manuscripts received May 30, 2020; revised July 25, 2020.

Md. Mohsin Uddin is with East West University, Dhaka Bangladesh, (e-mail: mmuddin@ewubd.edu).

Nazmus Sakib Patwary, Md. Mohaiminul Hasan, and Tanvir Rahman were with East West University, Dhaka, Bangladesh (e-mail: nazmus.ewu@gmail.com, mohaiminul.hasan.ewu@gmail.com, tanvir.rahman.ewu@gmail.com).

Mir Tanveer Islam completed his BS in EEE from North South University, Dhaka, Bangladesh (e-mail: mirtanveerislam@Gmail.com).

doi: 10.18178/ijfcc.2020.9.3.565

network model that will predict person, any things or place related to given description about any entities. They have used huge amounts of unstructured and compositional data.

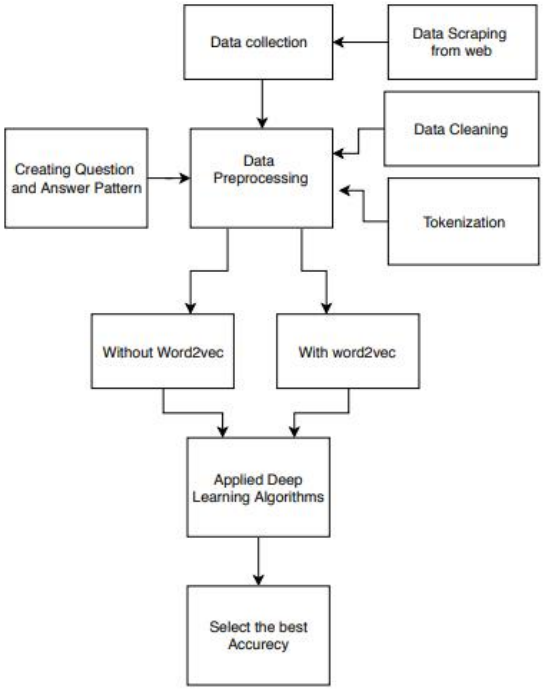

Fig. 1. Flow chart of proposed architecture.

They proposed dependency tree recursive neural network and they extended this to question answering neural network with trans-sentential averaging (QANTA) which can combined learn word and phrase-level representation to identify about entities. Their model takes dependency parse tree of sentences of questions and takes corresponding answer about entity as labeled input. Equivalent questions with slight difference may lead to different answers in case of question answering system [5]. A question answering system must interpret between those question and should give the appropriate answer from the knowledge base. In the paper, authors proposed a model that learns paraphrases for question answering and produces an estimation of probability distribution over all candidate answers.

Another paper [4] where researchers proposed an efficient neural model denoted as FastQA for question answering which outperforms existing model over very popular recent datasets named SQuAD, [18] NewsQA [19] and MsMARCO [20]. They used bidirectional RNN to learn input question and context. For answer span they used beam-search with a size K which reveals the top k-answers are predicted and the highest probability span is predicted as the appropriate answer based on learning model.

For accomplishing language understanding by the use of appealing QA dependent strategies has been emerged in near future [1]. In the paper authors have done experiment with 20 different tasks (based on supporting fact and reasoning fact) of question answering for getting suitable models applicable to detect symbolic sequence in language. Standard MemNN [21] outperforms over the LSTM baseline and N-gram baseline in their approach applied on simple bAbI dataset. Using the same dataset, PE representation enhances performance over bag-of-word (BoW) [8] for few particular tasks [2]. Raising hops also yields better accuracy for their given approach.

III. Methodology

In our study, several steps have been followed and the Fig. 1 illustrates the whole procedure. After executing all those steps, several outcomes have been found and best few accuracies have been taken for the evaluation of all the methods.

Firstly, we collected existing data set. 1 and translated into Bangla and the another data set we have created where the dataset contains many historic information. So we applied both dataset. After the data collection step, we have used several data preprocessing techniques. Then we applied two Deep Learning techniques using both position encoding and word embedding(word2vec). In this process, we have created several combination using activation function and optimizer of both LSTM and GRU algorithms.

A. Dataset

As, for the very first appearance, QA system for Bangla language has been conducted by us, we had to collect Bangla. corpus from Bangla wiki2 to assessment the new approach. Based on that corpus we have generated question-answer pair as well as supporting line by hand manually for the reason no preprocessed dataset in Bangla language has been found. In our study we only consider history domain related question in Bangla language. So, our model is based on close domain QA which is prerequisite motive for a robust AI-question answering system of Bangla language. We also evaluate our method on facebook bAbI translated– dataset [b1, b2].

TABLE I: Datasets Overview

Source of Dataset	Total Line	Total Words	Question-Answer Pair
Facebook bAbI	30000	125000	10000
History Corpus	3500	30000	1500

After having a dataset, we have purified our dataset removing unnecessary symbols and objects, words from other language. By doing all these cleaning stuffs our dataset got ready completely for the experimental and learning issues of the framework and it can be triggered for the further study concerning QA in Bangla language.

B. Data Preprocessing

In our study, we have applied two types of data, one is translated data and another one is wiki data. In dataset section, we have discussed in details about datasets. In second step, we have cleaned our data, mostly unnecessary characters or word or URL link have been removed. For each and every cases we have created pattern and simply removed the unnecessary word or characters. For the final step, we have preparing data for the training for example creating question, query pattern and tokenization.

Web scraping: For wiki dataset, firstly we scraped history of Bangladesh from wiki. Web scraping is very efficient technique to collect the data from website. It saves lots of time. Using Web scraping, can be collect lots of data from websites. In this paper we scraping history of Bangladesh from wiki.

Tokenization: Tokenization is a process by which corpus sentence is converted into meaningful pieces where each word of the corpus is considered as a token. By dividing sentences into collection of tokens following splitting by space data gets ready for the next step called word representation. For Example Table II).

TABLE II: EXAMPLE OF TOKENIZATION

Sentence	Tokenization
I love Bangladesh.	'I', 'love', 'Bangladesh', '.'
আমি বাংলাদেশকে ভালবাসি।	"আমি", "বাংলাদেশকে", "ভালবাসি", "।"

C. Implementation

In our study, we have used positing encoding (PE) and word embedding (word2vec) [9], [10] for Bangla in both LSTM and GRU algorithms. Every unique word has a unique position in PE. Word embedding is a large amount of words where it's precisely syntactic and semantic relationship between words. Word2vec represents the word into a vector space where it helps the algorithms to achieve the high performance in natural language processing (NLP) works. Word2vec has dimensions where it can identify the similar or dissimilar word using those dimensions. Vocab size of the word2vec [9] mode is 436126 where it has 300-dimensional vectors.

LSTM [11]-[13] is an upgraded version of RNN [14] where LSTM makes easier to remember to previous data in memory. RNN has problem with vanishing gradient. Here LSTM resolved the problem. LSTM is suited to classify and predict the time series based on lags of unknown duration or time. In this paper, we have used two input vector, one of them is input sequence (story) and another one is question (query). The 'answer' as a final model where it combines the response model and encoded the query as input and after that it predicts the answer word.

$$\text{Gate}_{\text{input}} = \sigma(W_{input}[h_{t-1}; X_t] + b_{input})$$
$$\text{Gate}_{\text{forget}} = \sigma(W_{forget}[h_{t-1}; X_t] + b_{forget})$$
$$\text{Gate}_{\text{output}} = \sigma(W_{output}[h_{t-1}; X_t] + b_{output})$$

where, bx denotes bias for the respective gates(x), Xt denotes input at current timestamp, Wx denotes weight for the respective gates(x) neurons, ht-1 denotes output of the previous lstm block (at timestamps t-1) and Gate x represents different types of gates.

GRU [15], [16] is the variant of LSTM. GRU's are simpler and faster than LSTM. GRU keep possession of the resisting vanishing gradient properties. GRU have two gates, one of them is update gate and another one is reset gate. In below, There have two types of formula where one is for reset gate and another one is for update gate.

$$\text{Gate}_{\text{reset}} = \sigma(W_{input_{reset}} * X_t + W_{hidden_{reset}} * h_{t-1})$$
$$\text{Gate}_{\text{update}} = \sigma(W_{input_{reset}} * X_t + W_{hidden_{update}} * h_{t-1})$$

D. Predicting Answer

Every question has a supporting number which help to find out appropriate answer. For example,

TABLE III: QUESTION ANSWER SAMPLE - 1

Question and Answer Pattern	Answer	Support
1 ক্রমবর্ধমান গণআন্দোলনের মুখে পাকিস্তানের কেন্দ্রীয় সরকার শেষ পর্যন্ত নতি স্বীকার করতে বাধ্য হয় এবং ১৯৫৪ সালের ৭ই মে পাকিস্তান গণপরিষদে বাংলা অন্যতম রাষ্ট্রভাষা হিসেবে গৃহীত হয়।		
2 ১৯৭১ সালে বাংলাদেশ স্বাধীন হলে একমাত্র রাষ্ট্রভাষা হিসেবে বাংলা প্রবর্তিত হয়।		
3 কত সালে বাংলাদেশ স্বাধীন হলে একমাত্র রাষ্ট্রভাষা হিসেবে বাংলা প্রবর্তিত হয়?	১৯৭১ সালে	2

TABLE IV: PREDICTED ANSWER SAMPLE -1

Testing Question	Actual Answer	Predicted Answer
কোন সালে বাংলাদেশ স্বাধীন হলে একমাত্র রাষ্ট্রভাষা হিসেবে বাংলা প্রবর্তিত হয়?	১৯৭১ সালে	১৯৭১ সালে
কখন বাংলাদেশ স্বাধীন হলে একমাত্র রাষ্ট্রভাষা হিসেবে বাংলা প্রবর্তিত হয়?	১৯৭১ সালে	১৯৭১ সালে

In the above Table III has some rows and columns. In column, one is for question and another is one for supporting. In table-4 has two rows have same question but have different structure. But both diffident scenario, the model gives accurate result. Another example for the better understanding where first table (Table V) is an example of how we have trained the data and second table (Table VI) is an example of paraphrase question patterns.

TABLE V: QUESTION ANSWER SAMPLE - 2

Question and Answer Pattern	Answer	Support
1 মদনপালের এই বংশের শেষ রাজা।		
2 তার পত্নী মন্ত্রীর সহযোগে বিষপ্রয়োগে স্বামী-হত্যা করেছিলেন।		
3 ১৯৪৬ সালে উড়িষ্যা রাজ্য গঠন হয়।		
4 উড়িষ্যা রাজ্য কত সালে গঠন হয়?	১৯৪৬ সালে	3

TABLE VI: PREDICTED ANSWER SAMPLE -2

Testing Question	Actual Answer	Predicted Answer
উড়িষ্যা রাজ্য কত সালে গঠন হয়?	১৯৪৬ সালে	১৯৪৬ সালে
উড়িষ্যা রাজ্য কবে গঠন হয়?	১৯৪৬ সালে	১৯৪৬ সালে

IV. MODEL EXPERIMENTS AND RESULT

For the experiment and analysis, we have used anaconda, which is an environment of python 3.7 with a lot of essential packages for machine learning and other useful tools.

A. Experiments

First of all, for our experiment over the proposed frame-work, we examine two separate designs choice: (i) Positioning Encoding; (ii) Word2vec (dimension d =300) for the word representation of the corpus which goes inside the

model for the learning purpose in the training times.

We also contrasted two more major variation by using RNN based: (iii) LSTM; (iv) GRU - for comparing both given results from the architectural view. In LSTM, it uses three gates: input, forget, output and it uses internal memory to reminisce the past sequential effects of early layers whereas GRU uses only two gates such as update gate and reset gate which is easier to implement compare to LSTM.

By considering different activation function in the sequential model: (v) Softmax; (vi) Linear; (vii) Relu - we have got diversity of results. For separate design choice of word representation, the results of the model vary from activation function to function. For an example, for the word2vec Softmax outperforms than other activation function that reveals that Softmax deals properly with word representation in higher dimension.

B. *Result Analysis*

In this section, we compare our very first approach for the Bangla language paraphrased QA based on single support fact. We evaluate on different variation to compare the results precisely.

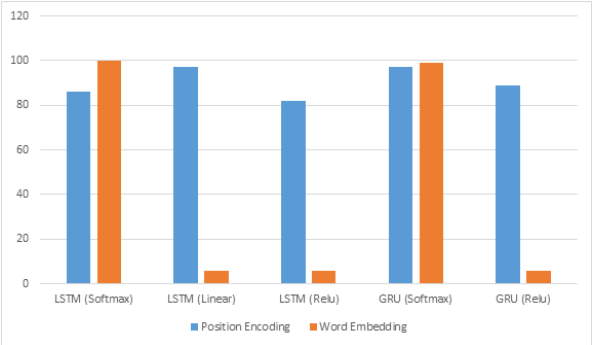

Fig. 2. Comparison view of average accuracy.

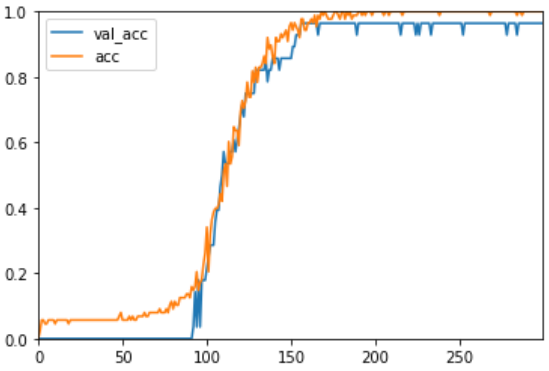

Fig. 3. Average accuracy of model using LSTM and Word2Vec with 300 epochs.

For our dataset, Fig. 2 illustrates the overall accuracy for different approaches. For the positioning encoding representation with LSTM neural architecture we have got average accuracy for Softmax, Linear, Relu activation function respectively 86%, 97%, 82%. Using word2vec word embedding representation for same sequence we have got average accuracy 100%, 5% and 5.6%. However, Word2vec with Linear and Relu activation function causes drastically collapse in case of average accuracy. For GRU based RNN we have got average accuracy of 96.59% and 88.6% respectively for Softmax and Relu activation function in case of positioning encoding. For word embedding GRU with Softmax has given average accuracy of 98.86%. For the bAbI dataset [1], [2] translated in Bangla language have gained 84% accuracy as the dataset contains only 24 unique words and word2vec doesn't fit good with that dataset. By applying their approach [1], [2] in our corpus we have got 86% accuracy whereas by adding word2vec in our framework we have got 100% accuracy.

Fig. 3 shows the graphical representation of our model which gives the highest accuracy (100%) with the combination of LSTM and word2Vec word embedding.

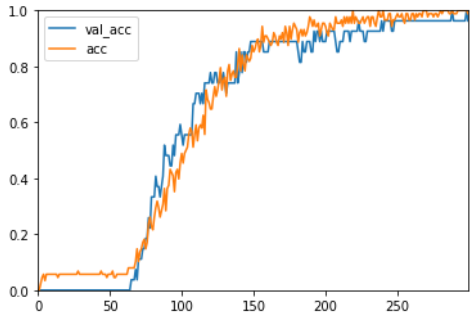

Fig. 4. Average Accuracy of model using GRU and Word2Vec with 300 epochs.

Fig. 4 illustrates the graphical representation of our model which gives the accuracy of 98.86% with the combination of GRU and word2Vec word embedding. Capitalize only the first word in a paper title, except for proper nouns and element symbols. For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [8].

V. Conclusion and Future Work

In this work, we effectively developed and trained heuristic neural framework applied on Bangla language that learns from question-answer pair with supporting fact and applicable to answer several paraphrased questions having single answer. This model with several contributions slightly approaches better execution using word2vec sentence representation. We have also experimented positioning encoding as sentence and question representation in the time of training the model which performs less accuracy than pre-trained Bangla word2vec with 300 dimensions' vector space representing each word of the Bangla corpus. By changing optimizer of GRU and LSTM based RNN variation on consequence have been come out.

However, as working with Bangla dataset is quit exhausting and originating Bangla question manually is tedious, our work could get further improvement by applying approach on a large question-answer dataset considering: chaining fact, induction, deduction etc. tasks. So, still we have much to work so for the further enhancement of complete QA system. Furthermore, our model is still unable to answer without supervision fact which is also a field of working in future.

Conflict of Interest

The authors declare no conflict of interest.

AUTHORS CONTRIBUTION

Md. Mohsin Uddin generated research idea, conducted the research and reviewed the work time to time while Nazmus Sakib Patwary, Md. Mohaiminul Hasan and Tanvir Rahman implemented the ideas with ML algorithm and LSTM.

The conception and design of this work were carried out by those four authors.

Nazmus Sakib Patwary, Md. Mohaiminul Hasan and Tanvir Rahman handled the data collection, analysis, and interpretation, with manuscript drafting. The paper was written by those three authors. Mir Tanveer Islam critically reviewed the study proposal and reviewed the final manuscript.

All authors discussed the results, commented on the manuscript and approved the final version.

REFERENCES

[1] A. Gelbukh, "Natural language processing," in *Proc. Fifth International Conference on Hybrid Intelligent Systems* (HIS'05), Nov 2005, p. 1.

[2] D. Weißenborn, G. Wiese, and L. Seiffe, "Fastqa: A simple and efficient neural architecture for question answering," 2017.

[3] L. Dong, J. Mallinson, S. Reddy, and M. Lapata, "Learning to paraphrase for question answering," *CoRR*, vol. abs/1708.06022, 2017.

[4] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. V. Merriënboer, A. Joulin, and T. Mikolov, "Towards ai-complete question answering: A set of prerequisite toy tasks," *Computer Science*, 2015.

[5] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "Weakly supervised memory networks," *CoRR*, 2015.

[6] M. Iyyer, W. Yih, and M.-W. Chang, "Search-based neural structured learning for sequential question answering," in *Proc. the 55th Annual Meeting of the Association for Computational Linguistics*, Jul. 2017, pp. 1821–1831.

[7] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III, "A neural network for factoid question answering over paragraphs," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014, pp. 633–644.

[8] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 12, pp. 43–52, 2010.

[9] F. Alam, S. Chowdhury, and S. Noori, "Bidirectional lstms — Crfs networks for bangla pos tagging," in *Proc. International Conference on Computer & Information Technology*, 2016.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 10 2013.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Eprint Arxiv*, 2014.

[12] A. Graves, "Generating sequences with recurrent neural networks," 2013.

[13] T. Mikolov, A. Joulin, S. Chopra, M. Mathieu, and M. Ranzato, "Learning longer memory in recurrent neural networks," *Computer Science*, 2014.

[14] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *CoRR*, 2018

[15] A. H. Mirza, "Variants of combinations of additive and multiplicative updates for gru neural networks," in *Proc. the 2018 26th Signal Processing and Communications Applications Conference* (SIU), May 2018, pp. 1–4.

[16] E. Nurvitadhi, J. Sim, D. Sheffield, A. Mishra, S. Krishnan, and D. Marr, "Accelerating recurrent neural networks in analytics servers: Comparison of fpga, cpu, gpu, and asic," in *Proc. the 2016 26th International Conference on Field Programmable Logic and Applications (FPL)*, Aug 2016, pp. 1–4.

[17] P. Rajpurkar, R. Jia, and P. Liang, *Know What You Don't Know: Unanswerable Questions for SQuAD*, 2018.

[18] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, 2006.

[19] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," in *Proc. 2nd Workshop on Representation Learning for NLP*, 2016.

[20] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, *A Human-Generated MAchine Reading COmprehension Dataset*, 2016.

[21] S. Sukhbaatar, J. Weston, and R. Fergus, "End-to-end memory networks," *Advances in Neural Information Processing Systems*, pp. 2440-2448, 2015.

Md. Mohsin Uddin is currently a lecturer of the Department of Computer Science and Engineering at East West University, Dhaka. He joined East West University in April 2018. He has obtained B.Sc. engineering. degree in computer science and engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka. He has obtained his MS degree in computer science from University of Lethbridge, Alberta, Canada. He worked as a software engineer in various multinational software companies in Canada and Bangladesh. He has 6+ years of professional experience in big data, NLP, machine learning systems design and software development. He has several research publications in renowned international conferences and journals. His research areas are NLP, machine learning, and big data.

Nazmus Sakib Patwary has completed his bachelor degree from the Department of Computer Science and Engineering at East West University (EWU), Dhaka, Bangladesh in December, 2019. He has several research publications in renowned international conferences.

His current interest is on machine learning, artificial intelligence and software development. He has been working around 2 years on machine learning and software development.

Md. Mohaiminul Hasan has completed his bachelor degree from the Department of Computer Science and Engineering at East West University, Dhaka, Bangladesh. He has around 2 years of experience on software development.

He has done his thesis on machine learning and natural language processing (NLP). His current interest is on machine learning and software development.

Tanvir Rahman has done his bachelor degree in computer science and engineering from East West University, Dhaka, Bangladesh. He has experiences on doing several projects on machine learning and web development.

He has done his thesis on natural language processing (NLP) and machine learning. He has an interest in doing research in machine learning field.

Mir Tanveer Islam is currently working as a FPGA design engineer in a semiconductor company, where he is developing hardware accelerators by using different neural network models.

He completed his BS in electrical and electronic engineering from North South University, Bangladesh in 2015 where he had training on microprocessors, computer system architecture and VLSI and verilog HDL. His research interests include computer vision, machine learning and artificial intelligence.