Heliyon 10 (2024) e25467

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

An ensemble novel architecture for Bangla Mathematical Entity Recognition (MER) using transformer based learning

Tanjim Taharat Aurpa [a,*], Md Shoaib Ahmed [b,c]

[a] Department of Data Science, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh
[b] Department of Computer Science, Boise State University, Boise, ID, USA
[c] Jahangirnagar University, Dhaka, Bangladesh

ARTICLE INFO

Dataset link: https://github.com/JUDataMiningResearch/Bangla_MER

Keywords:
Bangla mathematical entity
Entity recognition
Transformer-based learning
BERT
Bangla language
Bangla NLP

ABSTRACT

Mathematical entity recognition is indispensable for machines to accurately explain and depict mathematical content and to enable adequate mathematical operations and reasoning. It expedites automated theorem proving, speeds up the analysis and retrieval of mathematical knowledge from documents, and improves e-learning and educational platforms. It also simplifies translation, scientific research, data analysis, interpretation, and the practical application of mathematical information. Mathematical entity recognition in the Bangla language is novel; to our best knowledge, no other similar works have been done. Here, we identify the mathematical operator, operands as numbers, and popular mathematical terms (complex numbers, real numbers, prime numbers, etc.). In this work, we recognize Bangla Mathematical Entity Recognition (MER) utilizing the ensemble architecture of deep neural networks known as Bidirectional Encoder Representations from Transformers (BERT). We prepare a novel dataset comprising 13,717 observations, each containing a mathematical statement, mathematical entity, and mathematical type. In our recognition process, we consider our proposed architectures using accuracy, precision, recall and f1-score as the performance metrics. The results have shown a satisfactory accuracy percentage of 97.98 with BERT and 99.76% with ensemble BERT.

1. Introduction

During the COVID-19 pandemic, people have become accustomed to online systems, including education systems, and the injection of artificial intelligence(AI) has made these systems more user-friendly. Academicians and students have become dependent on automated systems to create question papers, solve them, take classes, etc. COVID-19 has been cured, but online education is still a trend as people find it a way to share knowledge around the globe. Mathematics, one of the essential basics of education, is also can be analyzed for the same purpose. Mathematical statements and entity analysis can play a significant role in question-making, solution generation, and other mathematics-related academic tasks. Mathematical Entity Recognition (MER) from multilingual text can be serviceable from different points of view. Resembling the Name Entity Recognition (NER), MER detects mathematical terms from textual math questions. Identifying these terms or entities can lead a system to generate mathematical functions and equations to obtain the solution to that question easily for any system.

* Corresponding author.
E-mail addresses: taurpa22@gmail.com (T.T. Aurpa), shoaibmehrab011@gmail.com (M.S. Ahmed).

https://doi.org/10.1016/j.heliyon.2024.e25467
Received 7 October 2023; Received in revised form 26 November 2023; Accepted 27 January 2024

Available online 5 February 2024

Bangla, the sixth-largest native language on the earth, is spoken as the first language by 234 million people around the globe, including Bangladesh and India.[1] It also has historical significance. In recognition of the historical sacrifice made by the Bangla language martyrs who battled to have Bangla as their mother tongue, UNESCO proclaimed February 21 as International Mother Language Day. Since then, it has developed into the mother tongue of the Bangladeshi people and the major language of instruction for many kids. Therefore the demand for automated systems based on the Bangla language is expanding inchmeal. In this case, the powerful present-day architectures like transformers are ruling.

One of the most trendy deep learning architectures, termed transformers, has been demonstrated for handling sequential data without the help of recurrent networks such as GRU or LSTM. Large datasets can be processed more quickly due to their parallelization capability and attention mechanism using encoder-decoder stacks. The use of transformers can be observed significantly in many works such [32,8]. Transformer architectures are prevalent for text-processing tasks. Bidirectional Encoder Representations from Transformers (BERT) is one of the most prominent architectures based on transformers. The use of such a deep learning model in the determination of mathematical entities can be beneficial to performing different mathematical academic tasks efficiently. Therefore, we intend to use a deep transformer-based model in our research for better accuracy and performance.

Since BERT was first introduced to the Natural Language Processing (NLP) world, it has been immensely popular with researchers. It is a transformer-based Pretrained language model that performs classification and prediction tasks using the rebellious self-attention mechanism. This cutting-edge language model performs better than practically all NLP tools. BERT is utilized for the classification of textual data in [18], [39], [30]), Question Answering based systems [20], the extraction or recognition of divergent entities e.g. ([37], [29], [1]). At researchers' own convenience, many customized BERT architectures have been proposed and fine-tuned. mBERT is one of those variations which have brought the blessing to the multilanguage research. mBERT(multilingual BERT) enable the state at of art performance in other languages over English. It has been utilized for multilingual classification of text data [23,17,14], Word Sense Disambiguation [42], Estimating the quality of Translation [6,9], etc.

1.1. Research objectives

The main research objective is to propose a unique method for recognizing the Mathematical Entity. We have ensembled two BERT models using different input sequences of Bangla mathematical statements end entities. Moreover, we have composed a novel dataset for this task with real-world Bangla mathematical statements and entities. The main objective of this research is mentioned below:

- Recognition of Bangla mathematical entities form real-world Bangla mathematical statements.
- Utilization of the latest transformer-based model BERT by ensembling BERT with different input sequences for the Bangla Mathematical Entity Recognition task.
- Provision and Composition of a novel dataset for this research and manoeuvring that for the proposed ensemble method using Bangla Mathematical statements and entities.
- Determining convenient machine learning metrics, for instance, Accuracy, Loss, Macro Average F1 Score and Micro Average F1 score for our proposed methodology and Compare the model to evidence the state-of-the-art performance.

1.2. Paper outline

The paper is arranged as follows. Section 2 is all about the literature review of this work. In section 3, we have mentioned the preliminary concept and the proposed methodology in detail. Section 4 is about the Environmental Setup for this research. Next, in section 5, we discussed the findings of our research work. Then an analytical discussion of this work is provided in Section 6. The final Section is about the Conclusion and the Future work of this research.

2. Related work

Due to BERT's outstanding performance, many researchers have chosen it to solve numerous problems. The paper [18] presented the finding of contextual modeling of embeddings through pre-trained architectures like BERT. They simply modify the layers of BERT architecture by adding a linear layer for classification and this modification contributed to improving the performance. Another contextualized work with BERT is conducted in [20]. They showed that their work is able to outperform existing CNN and BiLSTM-based methods with an average of 1.65%. Another BERT base work was done by [33] on the Twitter-based dataset where authors pre-trained the BERT model to improve the Latvian SA-based tweet's performance. The use of BERT in fake news detection is also chart-topping. Authors in [24] proposed a combined method using LSTM and BERT where the output of the BERT layer is connected to a LSTM layer. The architecture is able to improve 2.50% accuracy on PolitiFact dataset, and the improvement for GossipCop dataset is 1.10%. To break BERT's language constraints in [13] authors proposed Bangla-BERT. Here, they are able to increase the results by at most 5.3%. BERT's significant use can be seen in Bangla hate speech detection [2,11].

Deep learning has been maneuvered in different mathematical tasks precisely. Authors [38] have implemented different deep-learning methods for mathematical expression extraction, processing, and designing. They developed a public dataset for various

[1] https://en.wikipedia.org/wiki/Bengali_language.

math-based contents. Suleiman et al. [31] conducted research to determine the abstractive summaries from mathematical texts. They implemented the work with attention-based Recurrent Neural Network and LSTM. Here the ROUGE1 score is 43.85, the ROUGE2 score is 20.34, and finally, the ROUGE-L score is 39.9. PDF2LaTeX a new technique was implemented by authors in [36]. They implemented a CNN-LSTM-based architecture which first takes the image from pdf, applies OCR to extract text, and then converts the mathematical text into LATEX statements. Authors in [34] proposed a CNN RNN-based method for finding the mathematical definitions from the text.

One of the latest mathematical tasks is Mathematical Expression Recognition. Numerous research has been executed on this topic utilizing divergent technologies. Authors in [40] proposed an encoder and decoder-based network for hand-written mathematical expression detection. They implemented the Second Order Attention Network (SAN), which provided an ExpRate of 81.0. Another piece on this same domain is implemented as a Bi-directional fashioned mutual learning network with aggregated attention in [3]. They achieved the highest 56.85% accuracy with CROHME 2014 dataset. Sakshi [15] employed Support Vector Machine (SVM) and Convolutional Neural Network (CNN) for recognizing and categorizing math expression recognition. They used the Hasyv2 dataset, where they acquired an accuracy of 62.3% and 76.21% for the SVM and CNN, respectively. Another Convolutional Neural Network (CNN) based architecture came up with [26]. This approach successfully enhanced the accuracy in this domain on the same dataset, and the accuracy was 76.71%. Shinde et al. [28] created an equation solver with CNN on a different dataset MNIST. For complex equations, CNN provided 85% accuracy. A combined CNN-SVC-based model provided 89.76% accuracy for operator categorization and 91.48% for predicting the numbers in [16], creates another combination by ensembling contrastive and supervised learning. Authors here are able to bring 3.4% improvement on public datasets CROHME.

Entity Recognition is one of the famous Natural Language Processing Tasks. On different topics, various types of entities are recognized using deep learning. Authors in [27] manoeuvred different multilingual models such as mBERT for Name entity recognition (NER). They claimed their work is a state of art methodology with an 85.77% F1 score. A Deep Learning-based system in [10] used hybrid embedding for named entity recognition. The model outperformed with an F1 score of 83.50% and 75.99% for the Panjabi and Hindi NER tasks. Biomedical Entity Recognition also becomes a well-liked research focus nowadays. The paper [12] implemented a BiLSTM-CNN-Char Deep Learning model for Biomedical and clinical entity recognition. The authors here didn't use any heavy model. Therefore, the system is faster. Another paper [5] utilized transfer learning hierarchically and combined multi-task learning and fine-tuning for biomedical entity detection. They successfully improve accuracy on five gold standard Entity Recognition datasets from .42 to .98. Even so, research on Entity Recognition has been focused on various fields. Still, there needs to be more work on mathematical statements. Only work [41] was conducted based on only Chinese text. They implemented a combined model named BERT-BiLSTM-IDCNN-CRF, which provides 93.91% of the F1 score. Therefore we proposed a novel architecture which outperforms the existing methods.

3. Proposed methodology

3.1. Preliminary concepts

Before explaining the methodology, we intend to introduce some preliminary concepts. This section is all about the preliminary concepts of this research work.

3.2. Transformer-based learning

Transformer-based learning [35] introduced an attention mechanism in NLP research and brought a revolutionary change. Besides, the attention mechanism transformer consists of the encoder and decoder. The encoder inside the transformer follows some autoregressive steps. The sustained sequence z $(z_1,..,z_n)$ is obtained after mapping the input $(x_1,..,x_n)$ and assists to generate the output $(y_1,..,y_m)$ using the decoder. We have drawn the basic of the transformer in Fig. 1.

Two significant concepts of transformers are:

Encoder-Decoder Stacks: The encoder stack of a transformer contains $N = 6$ layers, which have two sublayers. These sublayers are responsible for maintaining the multi head self attention and a feed-forward network that is point wise fully connected. The sublayer's dimension is 512 and is *LayerNorm (x + Sublayer (x))*.

Multi-Head Attention: This mechanism has three different tasks to be conducted. Once the decoder layer's output is passed to the next encoder layers. Then the self-attention layer generates the queries, memory keys, and values using the output of the previous encoder layer's outcome. Decoder uses the autoregressive manner and masks out the input of softmax. With the keys and values in dimension d_k and the queries, the self-attention is calculated. 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 the matrix queries Q, keys K, and values V the attention calculated with Equation (1) and (2)-

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

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

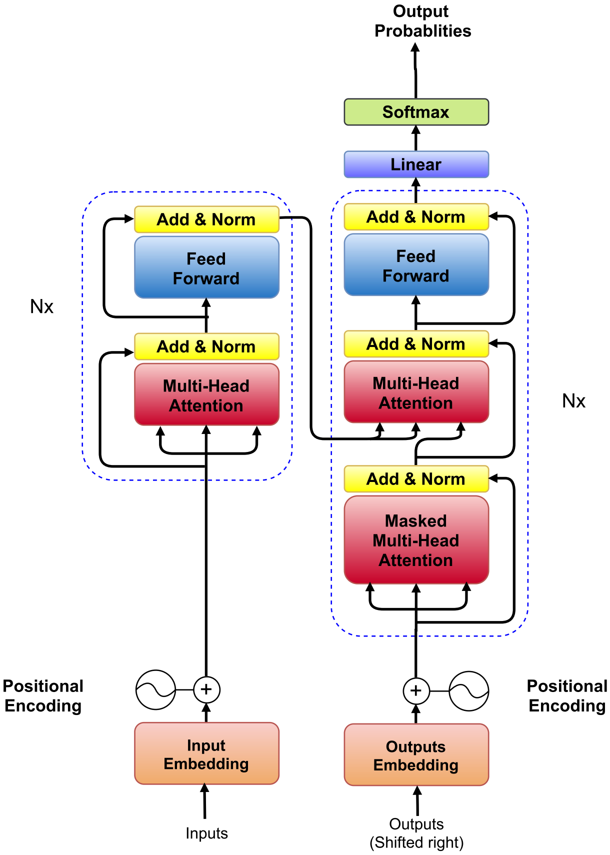

Fig. 1. The Transformer - model architecture. (The left and right side of the figure represents the working principle of the encoder-decoder stacks using the fully connected layers and self attention.)

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

3.3. Bidirectional Encoder Representations from Transformers (BERT)

The multilayered bidirectional transformer encoder with a self-attention mechanism is called a BERT [7]. A token sequence generated from input text that is unambiguous and contains one or more phrases is the input for BERT. BERT consists of the following two steps:

- **Pre-training BERT:** MLM, Mask Language modeling, and NSP, the Next Sentence Prediction, are two unsupervised tasks on which the BERT model is pretrained. To get the pre-trained bidirectional model, MLM involves masking a set of random tokens and making predictions about them. NSP aims to anticipate the subsequent sentence in a sentence pair. These models are known as autoencoding language models and can perform better than autoregressive models. BERT has received pre-training using the 800 million word BooksCorpus database and the 2,500 million word English Wikipedia text passages (without lists, headers, or tables) [43].
- **Fine-tuning BERT:** BERT is recognized for a number of downstream tasks such as Question Answering, Named Entity Recognition, Classification etc. It includes the ability to select appropriate inputs for both single and linked phrases. BERT starts fine-tuning the model with the initial value of pre-trained parameters. After propagating over labelled data, BERT adjusts the parameter value and finalizes them for the predictions.

Fig. 2 is the visualization of the BERT model where we have passed two input sequences to the model.

In Fig. 2, two inputs are passed to the BERT, and they are the Text and the Entity name. These two inputs are separated using a special BERT token [SEP].

3.4. Multilingual BERT (mBERT)

mBERT [19] is a variation of BERT which is pretrained to bring out a state-of-the-art performance for NLP tasks in various languages. For the accomplishment of that mission, it has utilized 10,000 sentences from one hundred and four different languages.

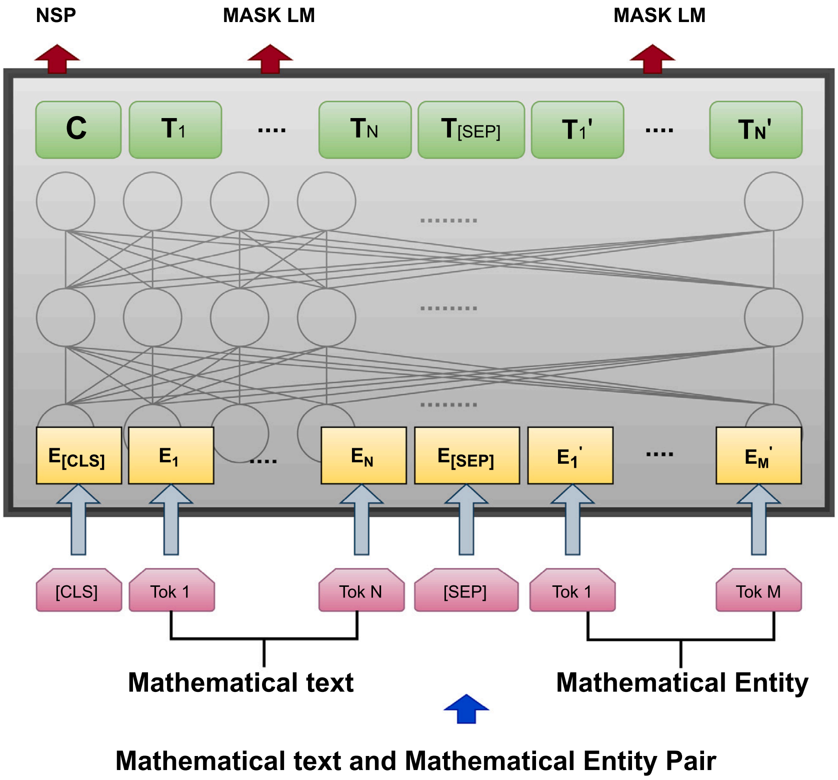

Fig. 2. BERT architecture. (Two steps of the BERT model- pretraining tasks and fine-tuning model).

Bangla is one of them. The collections of sentences are filtered from Wikipedia based on a minimum character length of 20. The dataset was divided into two equal parts for the training and testing of the model. mBERT is able to differentiate between language neural and language-specific parts of the data. Some probing tasks evaluated in mBERT are given below:

- **Language Identification:** When the linear classifier has been trained over sentence representations, it tries to distinguish the language of that sentence. Apart from other tasks, it fulfills the required fittings for the model.
- **Language Similarity:** Usually, the languages that are similar achieve very close Parts of Speech (POS) tagging representations [22]. In mBERT, these similar representations are using V-measure [25] on hierarchical language clustering with language families.
- **Parallel Sentence Retrieval:**
  The cosine distance within its representation and all other sentences' representations on the same parallel edge is calculated for each in the parallel-coupled sentence. The sentence with the smallest distance is selected. Here mBERT fitted linear regression for each language to project representations with English representations.
- **Word Alignment:** In mBERT, word alignment is identified as a bipartite graph's minimum weighted edge cover. This graph joins the tokens of two languages' sentences. Their edges are weighted with token representation's cosine distances.
- **MT (Machine Translation) Quality Estimation:** For MT tasks, mBERT justifies how the cosine distance of both source and translated sentences mirror the quality of MT. The cosine distance of the source sentence's representation and MT output's reflection is used to evaluate. Here also, a fully supervised regression and bilingual projection have been trained for centered and plain representations.

3.5. Proposed ensemble BERT architecture

In this research, we have proposed an ensemble method that utilizes all input information in divergent ways. When the BERT model works for training or validating with two text sequences as input, it requires special tokens [CLS], [PAD], and [SEP]. The [SEP] token is used between two input sequences. Here, the text sequences are the mathematical statements and the mathematical entities. We have constructed two different combinations of input sequences as follows:

The First Sequence:
$[CLS][T_1^{MS}].....[T_n^{MS}][SEP][T_1^{ME}].....[T_n^{ME}][PAD]...$

The Second Sequence:
$[CLS][T_1^{ME}].....[T_n^{ME}][SEP][T_1^{MS}].....[T_m^{MS}][PAD]...$

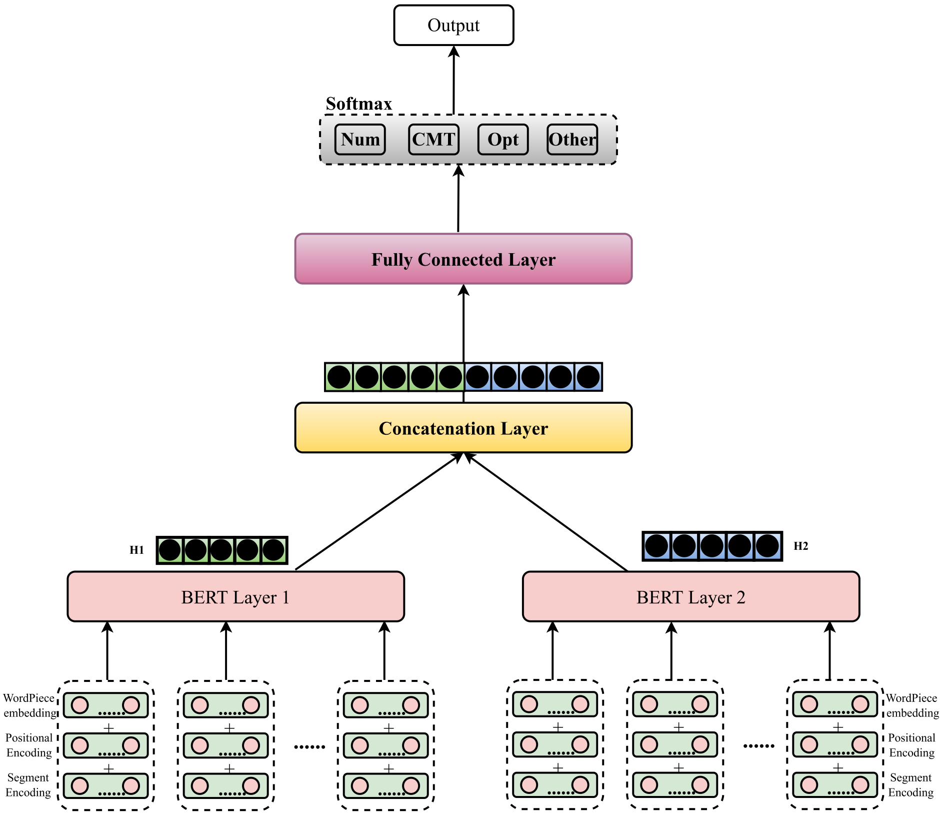

Fig. 3. The proposed ensemble BERT model with two different BERT layers.

In the first sequence, we keep the tokens of the math statement $[T_1^{MS}]$.....$[T_n^{MS}]$ After the [CLS] token, following the [SEP] token, we add the tokens of the math entity $[T_1^{ME}]$.....$[T_n^{ME}]$, and finally, we padded the sequence with the token [PAD]. Conversely, the second sequence has been organized in the opposite. We keep the tokens of the math entity $[T_1^{ME}]$.....$[T_n^{ME}]$ before the [SEP] token and the math statement $[T_1^{MS}]$.....$[T_n^{MS}]$ after it. Each of these sequences is sent to different BERT layers. Fig. 3 is the sketch of our proposed ensemble model. Here, the raw text is tokenized and provides different input tokens and an attention mask for different sequences. These two different input sequences passed to the input layers and followed to two different BERT layers. Next, we add a concatenated layer for combining the output of two BERT layers. Following a fully connected layer, the calculation ends in the softmax layer, which provides a probability for predicting entity class.

3.6. *Proposed framework*

Fig. 4 represents the workflow of the proposed method. The following steps are executed step by step.

- At first, the raw text, including the math statements and entities, is sent to the data processing step to get the preprocessed text and encoded entity names.
- The BERT tokenizer tokenizes the preprocessed mathematical statements and entities and precipitates the input sequence for the classifier.
- Then we train the model with input sequence and encoded labels. We trained both the BERT and the ensemble BERT model for comparison. We use the test data to identify the loss and accuracy for evaluating the models. We also use hyperparameter tuning during the model training.
- Finally, the trained model is ready to recognize the entity name using the unseen input statement and entity.

4. Experimental setup

4.1. *Experimental environment*

Deep Learning models need cutting-edge setups in order to handle data in parallel. As a result, we used Google Colab [4]. It is a cloud-based Jupyter Notebook platform that offers the essential tools for exploiting GPU and TPU. It is functional under Ubuntu

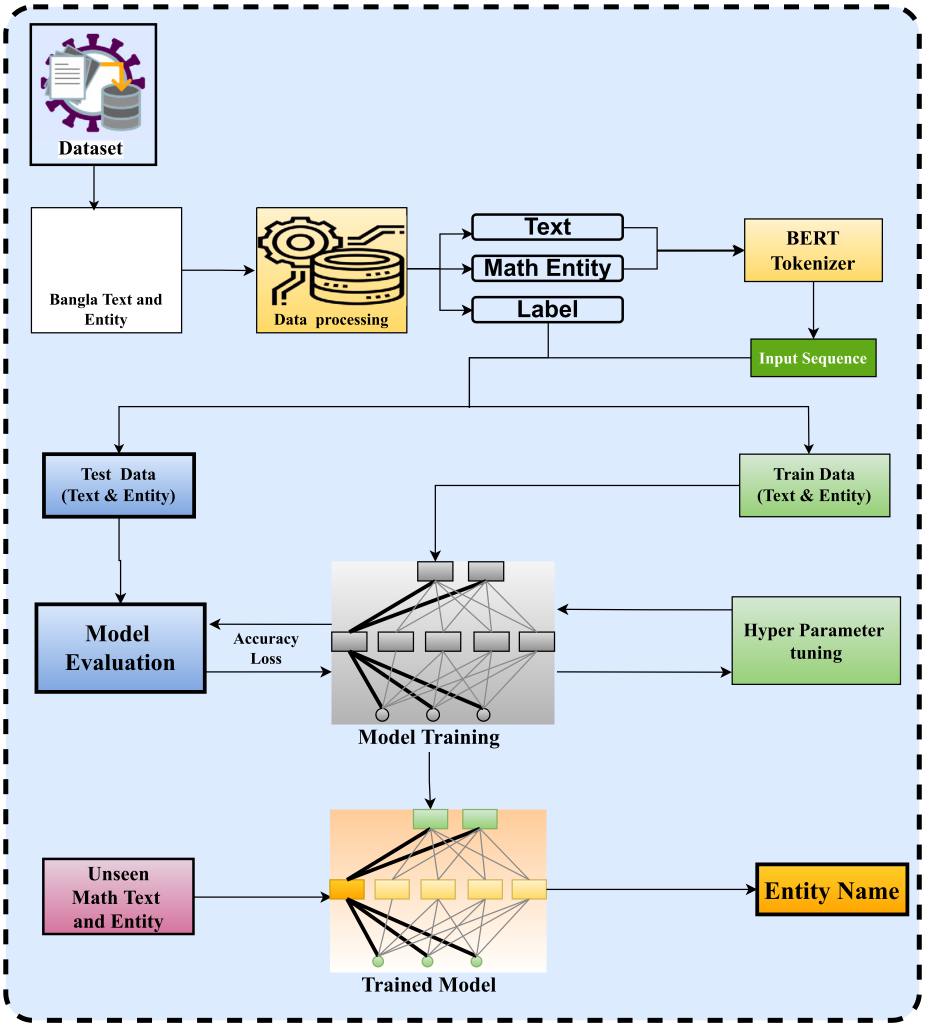

Fig. 4. Proposed framework for Bangla RC. (Here sketch the working procedure of the proposed model, starting from data preprocessing to predicting answers of RC.)

Table 1
The value of different hyperparameters for the proposed ensemble method.

Hyperparameters	Values for the proposed model
Learning rate (AdamW)	2e-04
Epoch	40
Batch_size	24
Verbose	1
max_len	60

OS and comes with 12 GB of GPU memory and an NVIDIA Tesla K-80 GPU. It included the necessary pre-configured libraries and packages for deep learning applications as well as the Python runtime.

4.2. Hyperparameter tuning

Data ordering and weight initialization of a deep learning model can be impacted by hyperparameter values. Determination of the hyperparameters' most essential values enables our model to make accurate predictions. We try out different values of Batch size, Maximum length of the sequence and Learning rate and observe which combination model performed best. In Table 1, we have provided the worthy values of our model's hyperparameters.

Here we obtained the best performance with Learning rate = 2e-04, max length = 60, verbose = 1 and batch size = 24. We consider 40 epochs to determine the accuracy. As the optimizer, AdamW [21] has been chosen in this research.

4.3. Dataset composition and provision

Math Entity Recognition (MER) is highlighted less in NLP research fields. It was challenging for us to manage sufficient informative data to make the model learn about this task. Therefore we have created our own dataset[2] for this task and trained the model. We Collected real-world mathematical statements and developed a novel dataset for Bangla MER with 13,717 unique observations. After the collection of the data, we extracted the mathematical entities from there and finally annotated them with four different entities mentioned as follows:

1. Numbers: This indicates the numerical entities from the text, for instance, one, two etc.
2. Operators: We extracted the operator from the text, such as addition, factorial etc.
3. Common Math Terms (CMT): Different math terms, such as prime number, complex number etc, often appear in mathematical statements.
4. Others: The other entities in the text are considered here.

The unique number of mathematical statements is 3430, from where we extracted four entities.

We partitioned the dataset into two parts: the training and testing dataset. Here, the training and testing dataset contains 11520 and 2197 observations, respectively.

4.4. Data prepossessing

Raw data incorporate extraneous words and characters. These might serve as barriers to classification. Moreover, these increase the complexity of the methodology. As a result, we skip putting raw data into the classifier directly. We preprocessed the data in different levels and then applied it to the model. These steps are mentioned below:

- 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[3] 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. So we apply stemming and lemmatization and use the root word to process the corpus.

5. Experimental evolution

5.1. Proposed model's performance

This section of the paper represents the result and other findings of this research. Here predominantly, we focus on different observations of BERT and our proposed ensemble method. We didn't have other transformer models for comparison except for BERT. The other transformers are not pretrained in the Bangla language. Therefore their performances are not significant compared to the BERT.

Before determining other scores, the first job is to select appropriate hyperparameters for the model. Consequently, we apply different combinations of hyperparameters and observe which combination of values is the most suitable for the proposed model. Table 2 broaches the accuracy of our model for different combinations of the hyperparameters and help to justify why we have moved forward with our chosen hyperparameters in this research. Here we consider five values of Learning rate - 1e-5, 2e-5, 3e-5, 4e-5 and 5e-5. Three different batch sizes- 12, 16, and 24 are utilized, and two different max lengths, 50 and 60, are used. After observing the accuracy of all the combinations, the model outperforms when learning rate = 2e-5, batch size = 24 and the max length = 60 with an accuracy score of 99.76%.

Here we train our proposed ensemble BERT model over 40 epochs and trace both the training and the testing accuracy. Fig. 5 indicates the curves of the training and the testing accuracy for all the epochs. The highest testing accuracy for the ensemble models is 99.76%, and the highest training accuracy value is 99.91%. We trace the training and testing accuracy for the BERT model too. Fig. 7 set forth the accuracy curves over forty epochs. Here the highest training accuracy is 99.79%, and the highest testing accuracy is 97.98%. Evidently, our proposed model accorded better results than the BERT-based classifier. .1199% training accuracy and 1.78% testing accuracy have been increased after ensembling the BERT model.

We also determine the loss over epochs for the ensemble BERT model. Fig. 6 delineates the training and testing loss curves over forty epochs. The minor training loss in these forty epochs is only 0.0021. The testing loss is reduced to 0.0154 in the testing loss curve. The training and testing loss for the BERT-based classifier is sketched in Fig. 8. Here the minimum training loss and testing loss are 0.0263 and 0.0265 over forty epochs consequently. The training loss is reduced by 0.0242 in the ensemble BERT model. The reduction of testing loss in our proposed model is 0.011. We have used the cross entropy loss using Equation (3).

[2] https://github.com/JUDataMiningResearch/Bangla_MER.

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

Table 2
Accuracy of the ensemble BERT model for various hyperparameter combinations.

Learning rate	Batch size	Max Length	Accuracy(%)
1e-5	12	50	99.45
		60	99.49
	16	50	99.05
		60	99.29
	24	50	98.96
		60	99.31
2e-5	12	50	98.92
		60	98.94
	16	50	98.93
		60	98.91
	24	50	98.96
		60	**99.76**
3e-5	12	50	98.89
		60	99.31
	16	50	99.20
		60	99.07
	24	50	98.65
		60	99.30
4e-5	12	50	98.91
		60	98.25
	16	50	99.08
		60	99.10
	24	50	99.01
		60	98.94
5e-5	12	50	98.29
		60	98.61
	16	50	99.20
		60	99.03
	24	50	98.76
		60	99.01

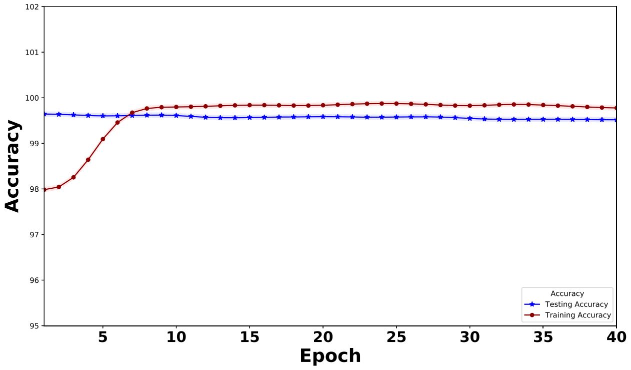

Fig. 5. The training and testing accuracy of the proposed ensemble BERT model over 40 epochs.

$$L = \sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \tag{3}$$

We have shown a few more performance outcomes of our proposed model. Fig. 9 depicts the confusion matrix of our proposed ensemble BERT model for the mathematical entities. The confusion matrix stipulates the correct classifications and misclassifications. It is observable in Fig. 9 that the misclassification rate is significantly less by the proposed classifier. Only 22 observations of Common Mathematical Terms (CMT) are misclassified as other entities. But all the operators and number entities are precisely recognized by the ensemble model. Therefore the misclassification rate for these two entities is zero.

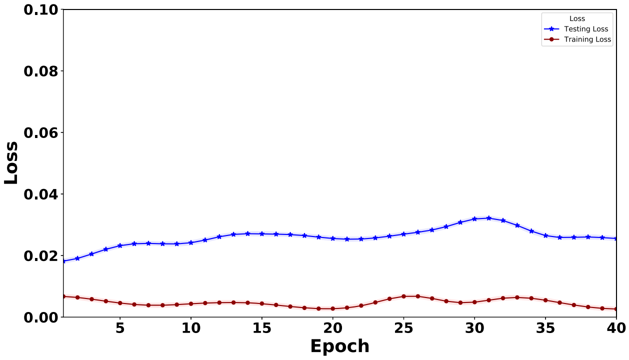

Fig. 6. The loss of the proposed ensemble BERT architecture over 40 epochs.

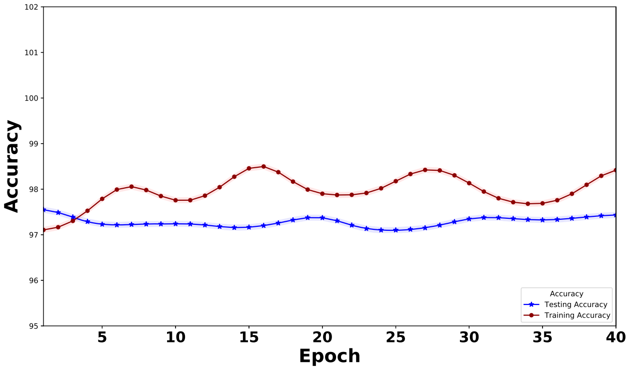

Fig. 7. The training and testing accuracy of the BERT model over 40 epochs.

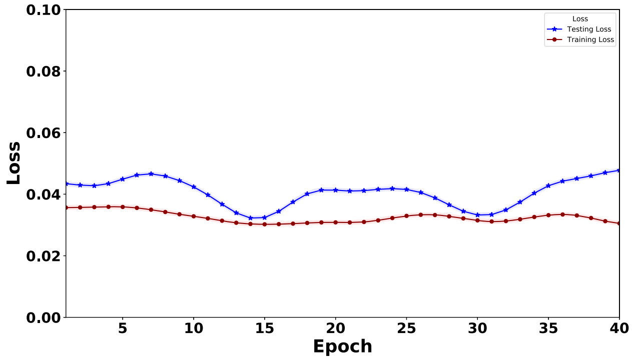

Fig. 8. The loss of the BERT architecture over 40 epochs.

Employing the confusion matrix, we determine impactful evaluation metrics such as Macro F1 score, Micro F1 score and Accuracy. The following Equations (4), (5), (6) are used to determine these metrics.

$$\text{Micro F1 Score} = \frac{\sum \text{TP}}{\sum \text{TP} + \frac{1}{2}(\sum \text{FN} + \sum \text{FP})} \tag{4}$$

$$\text{Macro F1 Score} = \frac{\sum_{i=1}^{\text{No of classes}} \text{F1\_Score}_i}{\text{No of classes}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

Here TP, TN, FP, and FN are the True Positive, True Negative, False Positive, and False Negative of the proposed ensemble model obtained from the confusion matrix. To calculate the Micro average F1 score, First, we determine the TP, TN, FP, and FN for each entity type. Then we sum the TP, TN, FP, and FN and determine the Micro average F1 score. For the Macro average F1 score, we

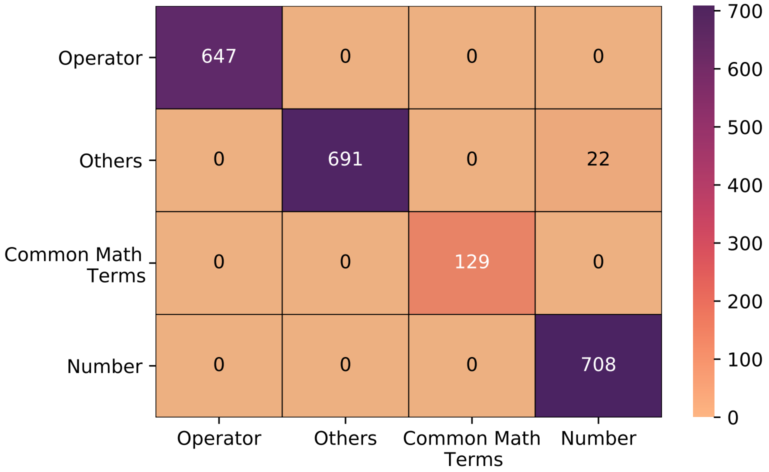

Fig. 9. The confusion metric of our proposed classifier.

Table 3
The score of the different evaluation metrices for the proposed ensemble model.

Metrices	Score of the Proposed Ensemble Model
Accuracy	99.76%
Micro F1 Score	98.99%
Macro F1 Score	99.23%

calculate the individual F1 score of each entity type and then divide the resultant value by the total number of entity types. Table 3 shows the Micro average F1 score, Macro average F1 score and the Accuracy of the proposed ensemble model.

5.2. Performance of the proposed ensemble technique on different transformers

The ensemble technique on the BERT model showed significant improvements in the results. Therefore, we applied the technique to other transformer models and justified the performance of our proposed ensemble techniques. We implement ELECTRA and XLNet on our dataset. After that, we have ensembled both ELECTRA and XLNet using our proposed methodology. For these two models, the performance increased also. XLNet provided 85.48% accuracy, while the ensembled XLNet model's accuracy was 86.34%. For this transformer model, accuracy improved by 0.86%. ELECTRA showed 82.01% accuracy. The ensemble techniques increased the accuracy by 1.27% and became 83.28%.

As these models are not trained in Bangla, the performance is comparatively poor compared to the BERT model. However, the ensemble technique significantly assists in increasing the performance. We sketched the comparison scenario in Fig. 10 for better visualization.

6. Discussion

Mathematical Entity Recognition (MER) is a unique problem that gets less research attention. However, mathematical entities have an essential contribution to the recognition or generation of mathematical expression. In this research, we intend to focus on that problem using the latest deep transformer model BERT. BERT has already evinced remarkable performance in Natural Language Processing for low-resource languages like Bangla. This model is pre-trained in several languages. mBERT portrayed a noteworthy performance. Moreover, we ensemble the proposed model by combining two different BERT layers, and both of these layers have taken different input combinations of the input sequences using the text, entity, and the [SEP] token. This method has significantly improved performance—the proposed method provided 1.78% more accuracy than the regular BERT classifier. The loss is reduced to 0.0154.

Another obstacle to this research was that there was no available dataset to train and test the model. Therefore we prepare a noble dataset for the Bangla Mathematical Entity Recognition (MER). We have collected real-world mathematical statements the extract different mathematical entities from there. We have worked with a total of four types of entities, and are Numbers, Operators, Common Mathe Terms and others. We believe this dataset will create a constructive impact on the research based on Entity Recognition. Moreover, it is able to contribute to Bangla Natural Language Processing (NLP) research.

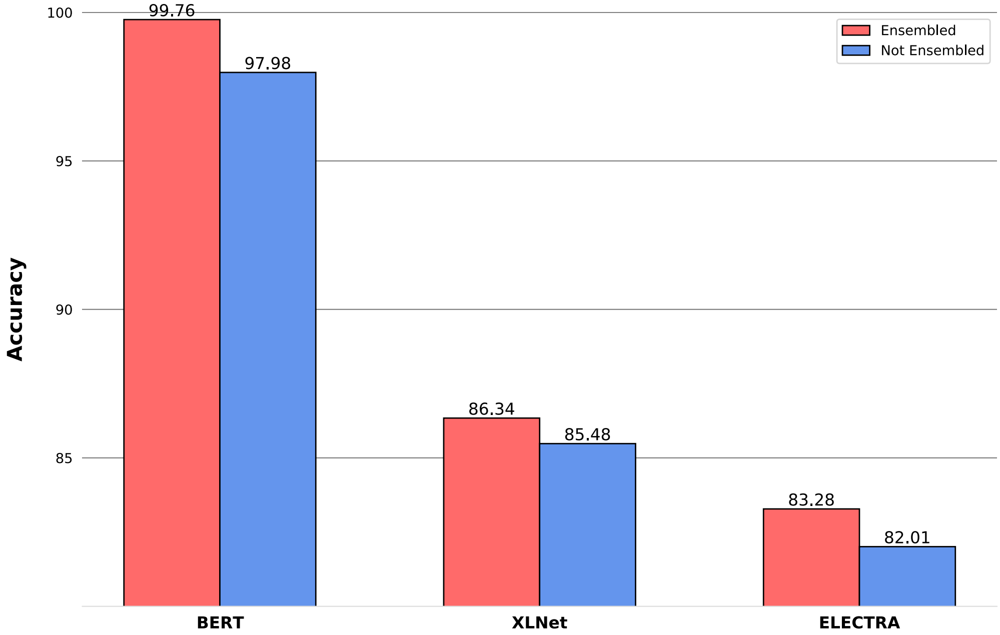

Fig. 10. Performance of ensemble technique on different transformers.

We ascertain the training and testing loss curves and the training and testing accuracy curves for the proposed model. We have used the Gaussian filter to smooth the curves in Figure. Besides, we find out the Accuracy, Loss, Macro Average F1 Score and Micro Average F1 score for the model. Moreover, we provided an analysis of the Confusion matrix of the proposed classifier.

We have determined different metrics for our proposed architecture to create a comparison scenario with BERT. When we applied this ensembled technique to other transformer models, such as XLNet and Electra, the ensemble technique performed better than a single-layered transformer. Each transformer (BERT, ELECTRA, or XLNet) layer captures different levels of contextual information from the input text. Combining multiple layers can result in a richer text representation, benefiting downstream tasks [7]. Moreover, the input sequences of the different layers are also differently organized. Another minor reason for this improved performance is that nonidentical instances of transformer models might have slightly different pre-trained weights due to variations in the training data or initialization. Ensembling them can help capture diverse patterns and information.

7. Conclusion and future work

The enormous importance and promising benefits of mathematical entity recognition in the Bangla language have been explored in this study. In order to expedite automated theorem proving, make it more straightforward to analyze and extract mathematical information from documents, as well as enhance e-learning and educational platforms, mathematical entity recognition is essential, as the study demonstrates. It has also shown how practical mathematical understanding is for empirical research, data processing, interpretation, and practical applications.

The use of the ensemble architecture of deep neural networks, notably Bidirectional Encoder Representations from Transformers (BERT), for Bangla Mathematical Entity Recognition (MER) is one of the work's key achievements. Additionally, the handiwork of a novel dataset, including 13,717 observations, each comprising a mathematical statement, mathematical entity, and mathematical type, is a worthwhile resource for further research and development in this domain.

There is undoubtedly an array of directions in which future research can build on this study's successes and advance the subject of mathematical entity recognition in the Bangla language, for instance, the expansion of the dataset, multilingual support, handling complex mathematical statements (mathematical equation recognition and generation), domain-specific applications (computer vision, LLMs), adapting to emerging mathematical concepts (explanation of mathematical solution), integration with mathematical assistants, Etc.

CRediT authorship contribution statement

Tanjim Taharat Aurpa: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Formal analysis, Conceptualization. **Md Shoaib Ahmed:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We uploaded our data to the public platform GitHub with the repository name Bangla_MER. The data is available at https://github.com/JUDataMiningResearch/Bangla_MER.

Acknowledgement

Author Tanjim Taharat Aurpa wants to thank Md Joynal Abedin, Mushrafa Hussain, and Jerin Isnat Abedin for their constant support while conducting this research, as well as throughout her entire research career.

References

[1] 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.
[2] A. Baruah, K. Das, F. Barbhuiya, K. Dey, Aggression identification in English, Hindi and Bangla text using bert, roberta and svm, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 76–82.
[3] X. Bian, B. Qin, X. Xin, J. Li, X. Su, Y. Wang, Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 113–121.
[4] 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.
[5] Z. Chai, H. Jin, S. Shi, S. Zhan, L. Zhuo, Y. Yang, Hierarchical shared transfer learning for biomedical named entity recognition, BMC Bioinform. 23 (2022) 1–14.
[6] S. Chowdhury, N. Baili, B. Vannah, Ensemble fine-tuned mbert for translation quality estimation, arXiv preprint, arXiv:2109.03914, 2021.
[7] 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.
[8] S.E. Friedman, I.H. Magnusson, S.M. Schmer-Galunder, Extracting qualitative causal structure with transformer-based nlp, arXiv preprint, arXiv:2108.13304, 2021.
[9] 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.
[10] A. Goyal, V. Gupta, M. Kumar, Deep learning-based named entity recognition system using hybrid embedding, Cybern. Syst. (2022) 1–23.
[11] A.J. Keya, M.M. Kabir, N.J. Shammey, M. Mridha, M.R. Islam, Y. Watanobe, G-bert: an efficient method for identifying hate speech in Bengali texts on social media, IEEE Access (2023).
[12] V. Kocaman, D. Talby, Accurate clinical and biomedical named entity recognition at scale, Softw. Impacts 13 (2022) 100373.
[13] 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.
[14] J. Krishnan, A. Anastasopoulos, H. Purohit, H. Rangwala, Cross-lingual text classification of transliterated Hindi and Malayalam, arXiv preprint, arXiv:2108.13620, 2021.
[15] V. Kukreja, S. Ahuja, et al., Recognition and classification of mathematical expressions using machine learning and deep learning methods, in: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, 2021, pp. 1–5.
[16] V. Kukreja, et al., A hybrid svc-cnn based classification model for handwritten mathematical expressions (numbers and operators), in: 2022 International Conference on Decision Aid Sciences and Applications (DASA), IEEE, 2022, pp. 321–325.
[17] 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.
[18] X. Li, L. Bing, W. Zhang, W. Lam, Exploiting bert for end-to-end aspect-based sentiment analysis, arXiv preprint, arXiv:1910.00883, 2019.
[19] J. Libovickỳ, R. Rosa, A. Fraser, How language-neutral is multilingual bert?, arXiv preprint, arXiv:1911.03310, 2019.
[20] 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.
[21] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint, arXiv:1711.05101, 2017.
[22] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, arXiv preprint, arXiv:1906.01502, 2019.
[23] 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.
[24] N. Rai, D. Kumar, N. Kaushik, C. Raj, A. Ali, Fake news classification using transformer based enhanced lstm and bert, Int. J. Cogn. Comput. Eng. 3 (2022) 98–105.
[25] A. Rosenberg, J. Hirschberg, V-measure: a conditional entropy-based external cluster evaluation measure, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pp. 410–420.
[26] Sharma C. Sakshi, V. Kukreja, Cnn-based handwritten mathematical symbol recognition model, in: Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021, Springer, 2021, pp. 407–416.
[27] R. Sharma, S. Morwal, B. Agarwal, Named entity recognition using neural language model and crf for Hindi language, Comput. Speech Lang. 74 (2022) 101356.
[28] R. Shinde, O. Dherange, R. Gavhane, H. Koul, N. Patil, Handwritten mathematical equation solver, Int. J. Eng. Appl. Sci. Technol. 6 (2022) 146–149.
[29] F. Souza, R. Nogueira, R. Lotufo, Portuguese named entity recognition using bert-crf, arXiv preprint, arXiv:1909.10649, 2019.
[30] J. Su, S. Yu, D. Luo, Enhancing aspect-based sentiment analysis with capsule network, IEEE Access 8 (2020) 100551–100561.
[31] D. Suleiman, A. Awajan, Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges, Math. Probl. Eng. 2020 (2020) 1–29.
[32] 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.
[33] 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.

[34] N. Vanetik, M. Litvak, S. Shevchuk, L. Reznik, Automated discovery of mathematical definitions in text, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 2086–2094.

[35] 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.

[36] Z. Wang, J.C. Liu, Pdf2latex: a deep learning system to convert mathematical documents from pdf to latex, in: Proceedings of the ACM Symposium on Document Engineering 2020, 2020, pp. 1–10.

[37] 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.

[38] A. Youssef, B.R. Miller, Deep learning for math knowledge processing, in: Intelligent Computer Mathematics: 11th International Conference, CICM 2018, Hagenberg, Austria, August 13-17, 2018, in: Proceedings, vol. 11, Springer, 2018, pp. 271–286.

[39] J. Yu, J. Jiang, Adapting bert for target-oriented multimodal sentiment classification, in: IJCAI, 2019.

[40] Y. Yuan, X. Liu, W. Dikubab, H. Liu, Z. Ji, Z. Wu, X. Bai, Syntax-aware network for handwritten mathematical expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4553–4562.

[41] Y. Zhang, S. Wang, B. He, P. Ye, K. Li, Named entity recognition method of elementary mathematical text based on bert, J. Comput. Appl. 42 (2022) 433.

[42] X. Zhu, Cross-lingual word sense disambiguation using mbert embeddings with syntactic dependencies, arXiv preprint, arXiv:2012.05300, 2020.

[43] 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.