Data in Brief 54 (2024) 110407

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Bangla_MER: A unique dataset for Bangla mathematical entity recognition

Tanjim Taharat Aurpa [a,*], Samiha Maisha Jeba [d], Md Shoaib Ahmed [b,c], Mohammad Aman Ullah [d], Maria Mehzabin [d], Md Musfique Anwar [c]

[a] Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh
[b] Boise State University, United States
[c] Jahangirnagar University, Bangladesh
[d] IUBAT - International University of Business Agriculture and Technology, Bangladesh

ARTICLE INFO

Article history:
Received 7 October 2023
Revised 20 February 2024
Accepted 4 April 2024
Available online 12 April 2024

Dataset link: Bangla_MER (Original data)

Keywords:
Bangla NLP
Deep transformers
Bangla mathematical task

ABSTRACT

Mathematical entity recognition is essential for machines to define and illustrate mathematical substance faultlessly and to facilitate sufficient mathematical operations and reasoning. As mathematical entity recognition in the Bangla language is novel, to our best knowledge, there is no available dataset exists in any repository. In this paper, we present state of the art Bangla mathematical entity dataset containing 13,717 observations. Each record has a mathematical statement, mathematical type and mathematical entity. This dataset can be utilized to conduct research involving the recognition of mathematical operators, renowned mathematical terms (such as complex numbers, real numbers, prime numbers, etc.), and operands as numbers. The findings mentioned above, and their combination are also feasible with a modest tweak to the dataset. Furthermore, we have structured this dataset in raw format and made a CSV file, incorporating three columns: text, math entity, and label. As an outcome, researchers may easily handle the data, facilitating a variety of deep learning and machine learning explorations.

DOI of original article: 10.1016/j.heliyon.2024.e25467
* Corresponding author.
E-mail address: aurpa0001@bdu.ac.bd (T.T. Aurpa).

https://doi.org/10.1016/j.dib.2024.110407

Specifications Table

Subject	Artificial Intelligence
Specific subject area	The Bangla MER dataset contributes to the systems of Bangla NLP. These tasks are a part of Artificial Intelligence which can be used for creating automated systems.
Data format	Raw
Type of data	Table (String/Text)
Data collection	To create the dataset, first, we collected real-world mathematical statements. Next, we have extracted three different types of mathematical entities. Then we named them Numbers, Operators and Common Mathematical Terms (CMT)—the remaining words in the sentences we have classified as Others.
Data source location	International University of Business Agriculture and Technology, Dhaka, Bangladesh
Data accessibility	Repository name: Github DOI: https://doi.org/10.5281/zenodo.8323342 URL to data: https://github.com/JUDataMiningResearch/Bangla_MER
Related Research	[1] Aurpa, Tanjim Taharat, and Md Shoaib Ahmed. "An ensemble novel architecture for Bangla Mathematical Entity Recognition (MER) using transformer based learning." Heliyon (2024).

1. Value of the Data

- After COVID-19, automated educational systems have become increasingly popular, focusing on mathematics as a fundamental education component. Beyond AI-generated math question sets, the need for solutions has increased, leading to the possible use of Mathematical Entity Recognition (MER) from multilingual text to cater to various solution viewpoints.
- Bangla, the world's sixth most spoken language, is the predominant language of 228.7 million people in Bangladesh and India [4]. It has enormous historical significance. UNESCO recognized February 21 as International Mother Language Day to honour Bangla language martyrs who heroically fought it, elevating it to a critical significance. It is presently the mother tongue of Bangladeshis and an important instructional language. In this world, advanced structures like transformers are the most powerful in the field of NLP.
- Currently, MER gathers data from real-world mathematical statements, enabling the compilation of valuable solutions in real-time. This dataset could indirectly support the existing Bangla educational system.
- Mathematical Entities are essential in generating mathematical expressions and functions. Therefore this dataset can help math instructors create and solve different mathematical problems automatically.
- To address this, we created a new dataset of 13,717 Bangla MER instances taken from real-world arithmetic expressions and categorized them into three categories. By selectively excluding raw data from direct categorization with 3,430 different mathematical assertions, we improved the effectiveness and real-time application of deep learning model training.

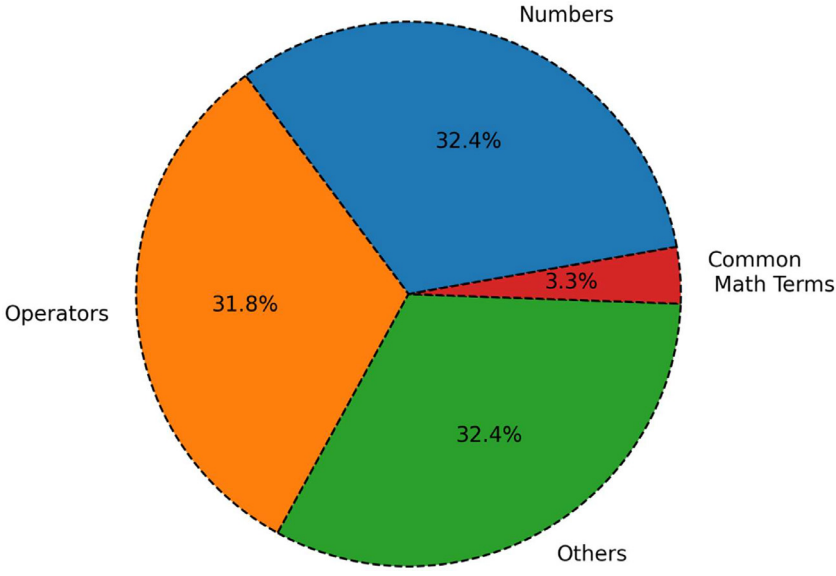

Fig. 1. Percentage of Bangla mathematical entities in the dataset.

2. Data Description

We created a new dataset of 13,717 Bangla MER instances taken from real-world arithmetic expressions and categorized them into four categories. By selectively excluding raw data from direct categorization with 3,430 different mathematical assertions, we improved the effectiveness and real-time application of deep learning model training.

1. **Numbers:** These indicate numerical entities in the text, such as 'one' and 'two.'
2. **Operators:** Words like 'addition' and 'factorial' were taken directly from the text and used as operators.
3. **Common Mathematical Terms (CMT):** This category includes expressions like 'complex number' and 'prime number' that are widely used in mathematical statements.
4. **Others:** The text's other diverse elements fall under this category.

We have derived these four object kinds from a unique count of 3,430 mathematical statements. Fig. 1 represents the number of mathematical entities in our dataset.

The English translations of the Bangla data are contained within a separate file. Using our dataset, we used Google Translate (https://translate.google.com/) to speed up the translation process. An overview of the Bangla Mathematical Entity Dataset is given in Table 1.

3. Experimental Design, Materials and Methods

3.1. Experimental environment

To facilitate data collection, we employed the Google cloud-based platform known as Google Sheets and stored the data in the CSV (Comma-Separated Values) format. The local machine used for data collection comprises an AMD Ryzen 7 5700U CPU and 16 GB of RAM. Google Colab, a cloud-based notebook service, trains deep learning models. The system offers GPU and TPU capabilities and is compatible with the Ubuntu operating system. It specifically supports the Tesla K-80 GPU manufactured by NVIDIA, equipped with 2 GB of GPU memory.

Table 1
This Table includes exemplary samples from our MER Dataset with Google Translation into English.

	Text	Math entity	Label
Bangla Text	শ্রীলংকার জাতীয় ক্রিকেট দলের আঠারো জন খেলোয়াড় বাংলাদেশে খেলতে এসেছেন। বাংলাদেশ দলেও আঠারো জন খেলোয়াড় আছেন। দুই দলে মোট ছত্রিশ জন খেলোয়াড় আছেন।	আঠারো, ছত্রিশ	Number
English Translation	Eighteen players of the Sri Lankan national cricket team have come to play in Bangladesh. Bangladesh team also has eighteen players. There is total thirty-six players in two teams.	Eighteen, Thirty-Six	Number
Bangla Text	সুজন বার্ষিক ক্রীড়া প্রতিযোগিতার জন্য প্রতিদিন একশো মিটার , চারশো মিটার ও আটশো মিটার দৌড়ায়। প্রতিদিন সে কত মিটার দৌড়ায়	যোগ	Operator
English Translation	Sujan runs 100m, 400m and 800m daily for annual sports competitions. How many meters does he run every day?	Addition	Operator
Bangla Text	দশ কোনো মৌলিক সংখ্যা নয়।	মৌলিক সংখ্যা	Common Mathematical Terms
English Translation	Ten is not a prime number.	Prime Number	Common Mathematical Terms
Bangla Text	মাইশার জন্মদিনে তাদের বাড়িতে তেরো জন বন্ধু এবং পাঁচ জন আত্মীয় এসেছিল। জন্মদিনে তাদের বাড়িতে মোট আঠারো জন অতিথি এসেছিল	মাইশার জন্মদিনে তাদের বাড়িতে জন বন্ধু এবং জন আত্মীয় এসেছিল। জন্মদিনে তাদের বাড়িতে মোট জন অতিথি এসেছিল	Others

(*continued on next page*)

Table 1 (*continued*)

English Translation	Thirteen friends and five relatives came to their house on Maisha's birthday. A total of eighteen guests came to their home on their birthday	Many friends and relatives came to their house on Maisha's birthday. Total number of guests came to their house on birthday	Others
Bangla Text	রাতুল ও মিতু শহীদ মিনারে ফুল নিয়ে এসেছে। রাতুল এনেছে তিনটি ফুল, মিতু এনেছে দুইটি ফুল । তারা মোট পাঁচটি ফুল নিয়ে এসেছে	তিন, দুই, পাঁচ	Number
English Translation	Ratul and Mitu brought flowers to Shaheed Minar. Ratul brought three flowers; Mitu brought two flowers. They brought five flowers in total	Three, two, five	Number
Bangla Text	আকাশ তার বাড়ির সামনে দিয়ে সকালে চল্লিশ টি গাড়ি ও বিকালে ছত্রিশ টি গাড়ি যেতে দেখেছে। ঐ দিন বাড়ির সামনে দিয়ে সে তিয়াত্তর টি গাড়ি যেতে দেখেছে	যোগ	Operator
English Translation	Akash saw forty cars passing in front of his house in the morning and thirty-six in the afternoon. That day he saw seventy-three cars passing in front of the house	Addition	Operator
Bangla Text	মাহিরের তেইশটি গল্পের বই আছে। অপূর্বের সতেরোটি গল্পের বই আছে। মাহিরের থেকে অপূর্বের ছয়টি বই কম আছে	বিয়োগ	Operator
English Translation	Mahir has twenty-three story books. Apoorva has seventeen story books. Apurba has six books less than Mahir	Subtraction	Operator
Bangla Text	ষাটের ফ্যাক্টরিয়ালের মান এক হাজার ছেষট্টির চেয়ে বড়	ষাট, এক হাজার ছেষট্টি	Number
English Translation	The factorial value of sixty is greater than one thousand and sixty-six	Sixty, one thousand and sixty-six	Number
Bangla Text	তিন এর চেয়ে বড় প্রত্যেক মৌলিক সংখ্যার বর্গকে বারো ধারা ভাগ করলে এক অবশিষ্ট থাকে ।	মৌলিক সংখ্যা	Common Mathematical Terms
English Translation	Dividing the square of every prime number greater than three by twelve leaves a remainder of one.	Prime number	Common Mathematical Terms

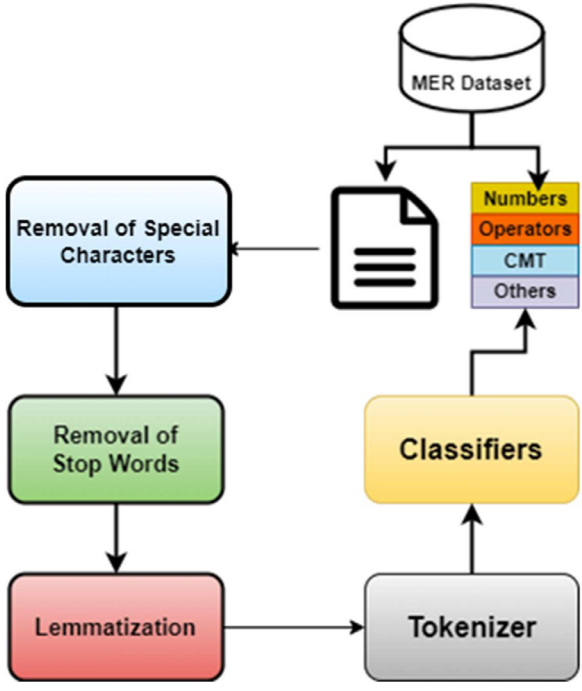

Fig. 2. The process of data preprocessing.

3.2. Data preprocessing

Before using the data for downstream operations, it is crucial to thoroughly clean it to eliminate any potential performance issues brought on by unnecessary letters, stop words, and other components in the raw text. The following preprocessing procedures, which are described below, support the improvement in classifier accuracy:

- The performance of downstream tasks is considerably improved by removing unnecessary punctuation ('.', '?', '|', etc.) and special characters ('#', '$', '&', etc.). Effectively, these superfluous characters have been removed from the dataset.
- Bangla stop words [5] are meaningless when used in the context of deep learning exercises. Therefore, their elimination is crucial before using the dataset.
- To identify word roots, we finally used lemmatization and stemming approaches. For instance, the words পড়েছি, পড়ছি, পড়েন, পড়ছিলাম, পড়বো, পড়বেন, পড়ি, পড়তাম etc. are all derived from the word পড়. This identification of root words or lemmas has excellent potential to improve the results.

The data preprocessing procedures for any regression or deep learning issue are outlined in Fig. 2.

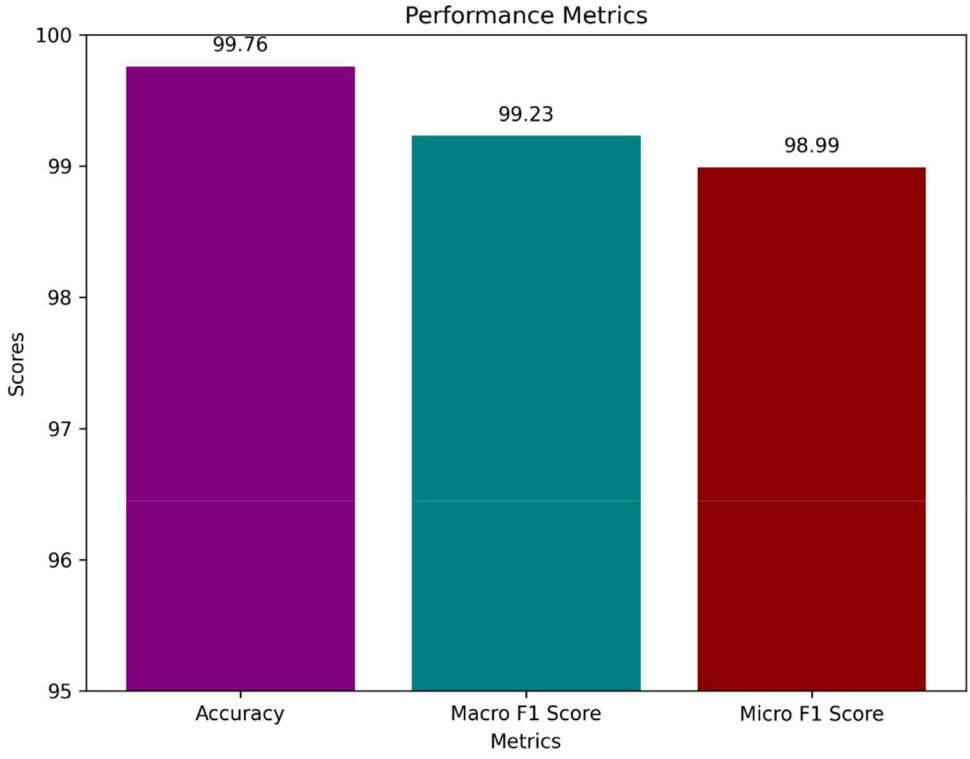

Fig. 3. Different Metrics for mBERT model in our MER dataset.

3.3. Data evaluation

We have applied a famous transformer-based architecture BERT (Bidirectional Encoder Representations from Transformers) to our Bangla MER dataset. Transformer models are state-of-the-art NLP models, and only the version of the BERT [2] known as mBERT [3] is trained in the Bangla Language. The model, trained with the dataset, is an ensemble model [1] where we combined two BERT layers with different input sequences. This ensemble model is the outperformer for this dataset.

Fig. 3 illustrates the results of mBERT on our dataset. The bar graph shows the different significant metrics for the mBERT model. The accuracy of the dataset is 99.76%. The macro and micro average scores are 99.23% and 98.99% respectively.

Limitations

The limitations of this dataset are enlisted below:

- The mathematical statements we have collected in this dataset are short, and the word count is below 100.
- The Common Mathematical Terms class has fewer observations than the other classes.

In future, we intend to overcome the limitations we have mentioned. Moreover, we will work to add entity relationships in the dataset.

Apart from these few limitations, this dataset represents a new idea and can be helpful to many researchers who are working on Bangla NLP. To our best knowledge, this is the very first dataset for Bangla Mathematical Entity Recognition.

Ethics Statement

The authors of this paper are aware of the ethical statements of this journal, and they agree with it.

Data Availability

Bangla_MER (Original data) (Github).

CRediT Author Statement

Tanjim Taharat Aurpa: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing, Data curation, Visualization, Supervision; **Samiha Maisha Jeba:** Writing – original draft, Visualization, Writing – review & editing; **Md Shoaib Ahmed:** Writing – original draft; **Mohammad Aman Ullah:** Writing – original draft; **Maria Mehzabin:** Writing – original draft; **Md Musfique Anwar:** Writing – review & editing.

Acknowledgements

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

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.

References

[1] T.T. Aurpa, M.S. Ahmed, An ensemble novel architecture for Bangla Mathematical Entity Recognition (MER) using transformer based learning, Heliyon 10 (2024) e25467.
[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[3] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. arXiv preprint arXiv:1906.01502.
[4] T.T. Aurpa, M.S. Ahmed, R.K. Rifat, M. Anwar, A. Alid, UDDIPOK: a reading comprehension based question answering dataset in Bangla language, Data Brief 47, 108933 (2023).
[5] R. Ul Haque, P. Mehera, M.F. Mridha, M.A. Hamid, A complete Bengali stop word detection mechanism, in: Proceedings of the Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), IEEE, 2019, pp. 103–107.