10.48047/jocaaa.2023.31.04.76

Study Material Classification in Bengali

Piyal Roy[1*], Rajat Pandit[1] and Sudip Kumar Naskar[2]

[1]Computer Science, West Bengal State University, West Bengal, India.
[2]Computer Science and Engineering, Jadavpur University, West Bengal, India.

*piyalroyOO@gmail.com

Abstract:
Study Material classification is an important task in Natural Language Processing (NLP). One of the most important and crucial components of intelligent tutoring and educational systems is classification of the study materials for learners. It is the task of categorizing study materials or text documents for the students of various grade levels. A text that seems adequate for a student at the under graduate level may not be appropriate for a student in school level, and vice versa. Therefore, in this study, we concentrated on the task of classifying study materials for students at various levels, starting in grade 4 and going all the way up to post-graduate level. We incorporated both fine-grained and coarse-grained approaches for classification task. For the task, we created a large dataset in Bengali from scratch. We deployed 11 machine learning models, 4 deep neural network models and 6 BERT models for the task. Each model that has been used produces a reliable classification. The top performing model we found is NeuralspaceReverie-Bert, which beats all other models and produces precision, recall, F-score and accuracy of 0.95, 0.95, 0.95 and 95.09%, respectively.

Keywords: NLP, Text Classification, Machine Learning, Neural Network, Bert

1. INTRODUCTION

Researchers over the past decade have been relentless in doing enormous research on NLP. Study material classification or finding the readability level of a study material for a student is one of the most significant and essential subfields of Natural Language Processing. In this task, study materials or text documents are classified for different level of readers. The readability level measures how simple it is for a reader to comprehend a text. The difficulty of reading a text is a significant aspect that affects its readability. One will need more time to grasp the meaning of the text the more challenging a sentence is. The choice of headings, paragraphs, the length of the sentences, and even the writing style all affect how readable a text is. The natural flow of the text and how simple it is to read are both impacted by each of these elements.

It can be challenging to approach writing objectively. Certainly, the majority of people can write a few sentences or an article on a particular subject. The more difficult part is making sure the text is appropriate for different reading levels. The text which is suitable for a post-graduate student may not be suitable for a grade 4 student. Extended phrases, the overuse of complex terms, and the failure to tie everything together into a coherent narrative chunk of text are only a few examples that make a book difficult to read and turn the readers away. Therefore, it is crucial to determine the readability level of a text so that it can be delivered to a suitable level of readers.

Researchers have explored two unique yet related approaches for classifying study materials: fine-grained classification and coarse-grained classification. In fine-grained classification, text documents are classified into a number of closely related grade levels by more in-depth research. This method makes it possible to distinguish between items meant for readers with different levels of skill more precisely. On the other hand, coarse-grained classification offers a more generic categorization and often divides texts into fewer general grade levels. Numerous researches are being conducted to explore the various applications of NLP. In English and other low resource languages including German[1], Russian, Dutch, Spanish, and Chinese [2], numerous works have been done to date to address the readability classification issue. The researchers have also worked on the readability classification issue in Bengali [3] [4] [5] [6] in

10.48047/jocaaa.2023.31.04.76

a number of their studies. However, there is a lack of research regarding the classification of study materials for the Bengali medium students at various levels, which is extremely important to their education.

This paper provides both fine-grained and coarse-grained approaches for classification of study materials for students at various academic levels, from grade 4 to post-graduate. To the best of our knowledge, there is still no dataset for this task of classifying study materials. So, starting from scratch, we created our own dataset. In order to create a large dataset for the classification of study materials, we gathered the original books offered to the students in grade 4 to post-graduate level. The study materials for various grade levels have been classified using a total of 21 machine learning and deep learning based models. We also aim to make our dataset public after the publish of this research work.

The contributions made in the paper are summarized below.

- We present a rigorous and in-depth study on classification of Bangla literature study materials across different academic levels, ranging from grade 4 to post-graduate, using 21 machine learning and deep learning models.
- A novel dataset has been created from scratch by collecting original books used by students in these academic levels to support the classification task. We plan to make the dataset freely available to the research community upon publication of this research.

Section 2 discusses some work related to study material classification in various languages including English and Bengali. Next, Section 3 depicts the dataset used in this experiment. The approaches for this paper are introduced in Section 4. Then the experimental outcomes have been presented in Section 5. Finally, Section 6 draws our conclusions and outlines for future research.

2. RELATED WORK

Study material classification problem is the process of automatically categorizing text study materials into predefined categories based on their content. The problem of text document classification has been the subject of numerous research studies. The majority of research have been carried out for the English language, however the quantity of research done for the Bengali language is still very low.

Hancke et al. [1] introduced a method for the classification problem of German text readability. They added language-specific morphological complexity indicators an extra set of features to pre-trained classifiers that had been trained with syntactic, lexical, and language model characteristics. To increase the precision of readability categorization, Vajjala et al. [7] came up with an idea of conducting their experiment employing WEKA's Multi-Layer Perceptron method using various combinations of lexical and syntactic data. Feng et al. [8] developed a collection of cognitively driven discourse-level variables that can predict adult readability of text document using a linear regression model. A method that can automatically differentiate between web pages for children and adults was proposed by Eickoff et al. [2] with the intention of improving the efficiency of a web search engine for children. They described a number of different 80 aspects from the perspective of the research done in the field of cognitive science and psychology of children. Furthermore, they conducted a considerable amount of user testing regarding website suitability and gave detailed information on their findings. A method of estimating reading difficulty based on both lexical and grammatical traits was proposed by Heilman et al. [9]. They contrast their integrated strategy with separate grammar and language-based strategies. However, the results indicate that the grammar-based method worked well compared to the vocabulary-based model. Predictions from the grammar-based model and those of the vocabulary-based model may be integrated using the concept of confidence scores to make more accurate predictions of the reading difficulty of first and second languages. Kate et al. [10] proposed a prediction framework for predicting the readability of natural-language documents based on regression with various linguistic and language modeling features. Their method outperforms naive human judges in its ability to anticipate readability as determined by linguistically trained expert human judges. Syntactic and lexical aspects were also beneficial, but they discovered that language-model based features were more effective for their task.

10.48047/jocaaa.2023.31.04.76

According to research by Pitler et al. [11], combining lexical, syntactic, and discourse variables results in a model that can accurately predict how human readers will rate the readability of a text. Their research explored a wide range of linguistic characteristics and used empirical evidence to show that discourse relations are closely related to the way a text is considered to be written. They demonstrated that a number of surface metrics that are typically thought to be connected to readability are not very accurate predictors of readability evaluations. On the other hand, Thompson et al. [12] have redefined the concept of readability in the context of statistical language modeling and introduced a new way of research approach for predicting the difficulty level of a text passage. They designed a metric using the extended version of multinomial naive Bayes classification model that uses the various language models. They argued that such a classifier can be trained on the basis of very little amount of labeled data and is subject-independent. Another group of researchers Schwarm et al. [13] have recommended that the use of SVMs to combine statistical languagemodels along with other features is more effective.

Islam et al. [4] [5] worked on Bengali textbook readability classification problem. They presented an information-theoretic and lexical readability classifier for Bengali textbook documents. Out of the 30 quantitative features they proposed, 25 are based on information theory. These characteristics are not required to be processed linguistically in any way. However, the outcomes of their experiments demonstrate that lexical and information theoretic features work incredibly well. In another study for Bengali text readability classification, Das et al. [3] developed a small readability model with one and two parametric fits for the multiple regression readability analysis of Bengali text. From seven different documents, they only used seven paragraphs. Two factors average sentence length and the number of syllables in every 100 words were used to determine whether a passage was easy or challenging. Another Bengali text readability classification approach was introduced by Sinha et al. [6]. They first explained how some of the popular readability metrics in English could not be used to Hindi and Bengali, due to their limitations. Then, they used user experiments to pinpoint crucial Bengali and Hindi structural elements that influence the reading of texts written in these two languages. As a result, they suggested two distinct reading models for both Hindi and Bengali.

A comparative literature review is given in Table 1.

Work	**Language**	**Methods**	**Accuracy %**
Hancke et al. [1] 2012	German	Sequential Minimal Optimization (SMO) algorithm	89.7
Vajjala et al. [7] 2012	English	WEKA's Multi-Layer Perceptron algorithm	93.3
Thompson et al. [12] 2004	English	Multinomial naive Bayes	93
Eickhoff et al. [2] 2011	English, Chinese, Dutch, French, German, Russian Spanish	SVM with combined topical aspects expressed by language modelling approaches with non-topical notions	83
Feng et al. [8] 2009	English	Linear regression model	
Heilman 2007 et al. [9]	English	KNN	93
Kate et al. [10] 2010	English	Bagged Decision Trees, Decision Trees, Linear Regression, SVM Regression, Gaussian Process Regression	81
Work	**Language**	**Methods**	**Accuracy %**

10.48047/jocaaa.2023.31.04.76

Pitler et al. [11] 2008	English	SVM	88.88
Schwarm et al. [13] 2005	English	SVM	89
Das et al. [3] 2006	Bengali	Regression Model	67
Islam et 2012 al. [4] [5]	Bengali	SVM	75
Sinha et al. [6] 2012	Bengali, Hindi	Regression Analysis	-

Table 1. Related Work

3. DATASET

Benchmark datasets for NLP works in Bengali are scarce, especially for the task of study materials classification, because the majority of researchers do not make their datasets available to the public alongside their research work. As a consequence, researchers tend to create their own dataset(s) from scratch, which is an extremely time-consuming and grueling process.

For this research, we started from zero to make our own dataset. We focused on the textbooks for students of 4^{th} to 12^{th} grade and from under-graduate (UG) to post-graduate (PG) level. We used the West Bengal Board of Secondary Education (WBBSE) textbooks for the students in 4th to 10th grade and the West Bengal Council of Higher Secondary Education (WBCHSE) textbooks for the students of 11th to 12th grade. Most universities do not provide books or study materials for their UG and PG level courses. As a result, we considered the Netaji Subhas Open University (NSOU), which offers books for UG and PG students, all of which are fortunately available to the public.

There are three stages to the development of our dataset, namely Books Collection, Text Extraction and Processing. The first stage, Books Collection, involves gathering books from trustworthy sources, including WBBSE and WBCHSE websites and the official NSOU website. The second stage extracts every word from the acquired PDF files using OCR engine. In the final stage, we divided the entire content of each text file into numerous smaller chunks after converting the pdf files into text files.

3.1 Books Collection

Gathering the actual books from trustworthy sources was the primary goal in creating the dataset. The WBBSE and WBCHSE books are publicly available on a number of websites. From one of them, we took the books for 4th to 12th grade level students. Both UG and PG books were collected from the official NSOU website. Each book has been digitally scanned and saved as a pdf. For this experiment, we considered only the books related to Bengali literature. Table 2, Table 3 and Table 4 contains information about each of the gathered books.

Book Name	**Grade Level**	**Sentences**	**Words**
Bhasha Path Class IV Bengali	4	1,476	11,882
Patabhar Class IV Bengali	4	3,013	27,028
Bhasha Path Class V	5	1,287	9,654
Patabahar Class V	5	2,571	21,310
Bhasa Charcha Class VI	6	1,951	17,911
Ha Ja Ba Ra La Class VI	6	564	5,509
Sahityamela Class VI Bengali	6	3,273	27,875
Sahityamela Class VII	7	4,027	34,306

10.48047/jocaaa.2023.31.04.76

Bhasa Charcha Class VII	7	1,421	12,790
Maku Class VII	7	1,654	15,909
Book Name	**Grade Level**	**Sentences**	**Words**
Bhasa Charya Class VIII	8	1,684	18,109
Pather Panchali Class VIII Rapied Bengali	8	2,713	33,083
Sahitya Mela Class VIII Bengali	8	4,486	40,861
Sahitysanchayan IX Bengali	9	2,431	23,964
Sahitysambhar-IX	9	2,135	20,855
Professor Shankur Dairy Class IX	9	3,736	36,120
Koni Class X Rapid Reader	10	3,345	25,392
Sahitya Sanchayan Class X Bengali	10	4,096	36,509

Table 2. WBBSE Books for Grade 4 to 10

Book Name	**Grade Level**	**Sentences**	**Words**
Bengali Bhasha O Sanaskrity Class XI	11	4,539	58,096
Sahitya Charya Class XI	11	3,139	24,157
Sahitya Katha for Class XI	11	1,338	13,744
Sahitya Charya Class XII	12	5,365	42,709
Sahitya Katha for Class XII	12	2,873	25,202

Table 3. WBCHSE Books for Grade 11 to 12

Book Name	**Grade Level**	**Sentences**	**Words**
Sahitya O Sanaskritir Itihas	UG	8,940	99,742
Bhashatatta Ebong Chondo O Olonkar	UG	6,867	73,141
Kabya	UG	15,509	1,25,467
Uponyash Ebong Chondo Golpo	UG	7,882	93,320
Probandha O Rommyarachana	UG	5,272	63,299
Natok O Nattyamancha Ebong Sahitter Rupved	UG	17,372	1,60,048
Loksahittya Ebong Adhunik Bharatiya Sahittya	UG	12,430	1,22,643
Anubad Bananbidhi Sampadana Anucched Protibedan	UG	8,317	89,485
Bangla Sahitter Itihas O Bhashatatta	UG	8,616	93,648
Chondo Alankar O Kabya Kabita	UG	9,927	93,135
Uponyash Chotogolpo Probandha Natok	UG	10,402	1,37,653
Bangla Sahitter Itihas	PG	27,376	3,08,381
Bhasha Bigyan O Sahittyatatta	PG	7,805	81,132
Prachin Moddhajugia Adhunik Kabita	PG	21,217	2,19,868
Probandha Sahittya O Saili Bigyan	PG	63,623	6,27,195
Bhasha O Sahittya Bisoyak Gobeshona Ebong Khetrasomikkha Paddati	PG	729	9,085
Anubad Charchar Proyojoniyota Bikashdhara Ebong Prokoron Pronali	PG	1,948	24,265

Table 4. NSOU Books for UG and PG

10.48047/jocaaa.2023.31.04.76

3.2 Text Extraction

Each page of the pdf files contained an image made from scanned pages from the source books. The next step was to extract every word from each and every one of the acquired pdf files. To extract from the pdf files, we used Google's Tesseract engine [14]. It is a Google-sponsored optical character recognition (OCR [15]) engine that is free and open-source for different operating systems. The retrieved text from each book is statistically represented in Table 2, Table 3 and Table 4 in detail.

3.3 Processing

We divided the entire content of each text file into numerous smaller chunks after converting the pdf files into text files. Every time we discovered two or more consecutive newlines in a text file, we divided the content into separate sections and we call it as segments. The segments that included fewer than 10 words were removed. This was done to avoid the risk of introducing noise and inadequate information into the dataset, which can compromise the quality and accuracy of the machine learning model. Short segments may not contain sufficient contextual information, which may negatively affect the model's ability to predict and classify data accurately. Additionally, removing short segments reduces the risk of overfitting or under fitting the model, which can result in poor generalization. It is important to note that balancing the removal of short segments with maintaining adequate diversity in the dataset is crucial to ensure that the model remains representative of the target population.

Finally, the segments were saved in a CSV file with the book title and appropriate grade level in the second and third columns, respectively. There is a total of 11 class labels in our final dataset. Table 5 provides statistics of the prepared dataset.

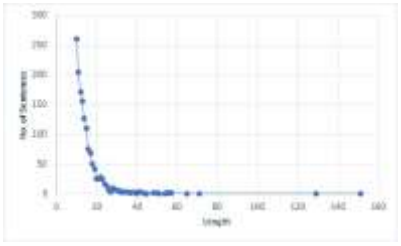
Grade	# Segments	# Sentences	# Words	# Sentences / Segment	# Words / Segment	#Words / Sentence	#Unique Words
4	989 (1.75 %)	3,785	35,814	3.83	36.21	9.46	9,729
5	900 (1.60 %)	3,168	27,860	3.52	30.96	8.79	7,650
6	1,318 (2.34 %)	4,800	46,783	3.64	35.50	9.75	12,144
7	1,621 (2.88 %)	5,823	57,340	3.59	35.37	9.85	14,516
8	2,157 (3.83 %)	7,779	87,301	3.61	40.47	11.22	20,279
9	1,929 (3.42 %)	7,434	77,490	3.85	40.17	10.42	18,480
10	1,458 (2.59 %)	6,665	58,678	4.57	40.25	8.80	15,514
11	1,448 (2.57 %)	7,944	91,785	5.49	63.39	11.55	25,175
12	1,479 (2.62 %)	7,350	65,047	4.97	43.98	8.85	17,521
UG	21,896 (38.85 %)	1,08,182	12,10,780	4.94	55.30	11.19	1,38,791
PG	21,171 (37.56 %)	1,10,025	12,14,688	5.20	57.38	11.04	1,98,971
Overall	**56,366**	**2,72,955**	**29,73,566**	**4.84**	**52.75**	**10.89**	**3,33,842**

Table 5. Statistics of Final dataset

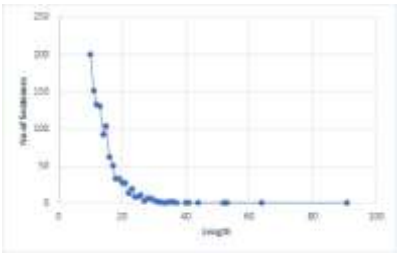
Figure 1 plots the number of sentences against sentence length for the different in the dataset. These graphs offer valuable insights into the distribution of sentence lengths. For the lower grade levels, such as class 4 and 5, we typically observe higher number of shorter sentences. This indicates a focus on shorter sentences, perhaps using simpler sentence construction, tailored to the comprehension level of young readers. As we progress to higher grades, like class 9, 10, and 11, the graphs exhibit a more diverse distribution of sentence lengths. This suggests an increasing emphasis on developing complex sentence structures and expanding vocabulary (cf. Table 5), aligning with the advanced

10.48047/jocaaa.2023.31.04.76

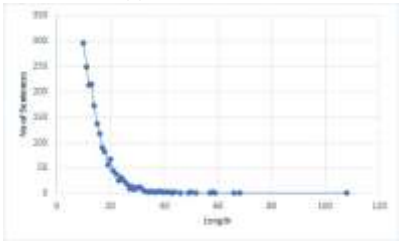
reading and writing skills expected at these levels. Moving on to UG and PG levels, we observe a further shift towards longer sentences and a broader range of word counts. This reflects the need for more detailed and comprehensive academic writing, incorporating refined vocabulary and complex sentence constructions.



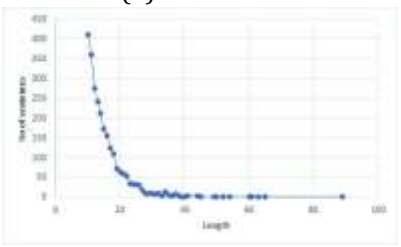
(a) Class 4



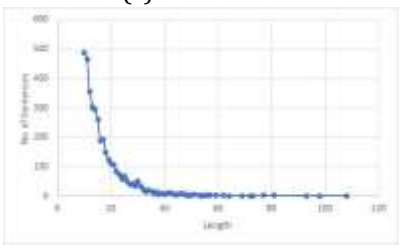
(b) Class 5



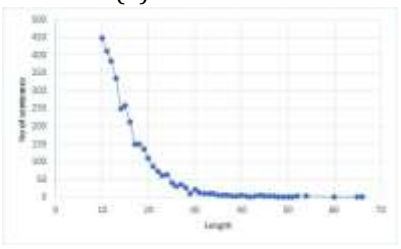
(c) Class 6



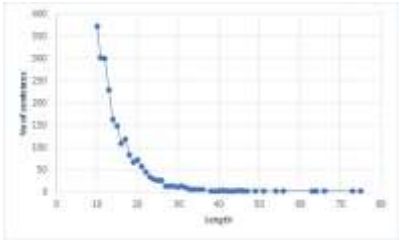
(d) Class 7



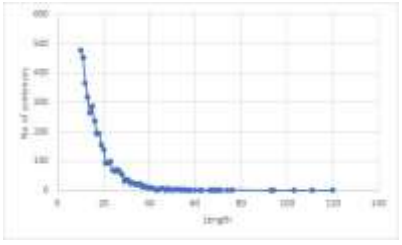
(e) Class 8



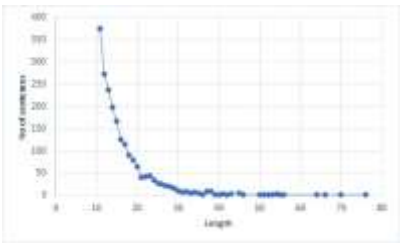
(f) Class 9



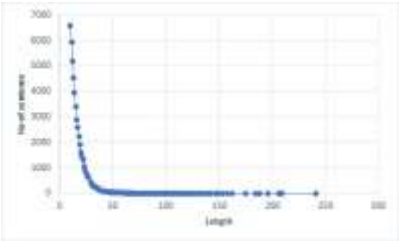
(g) Class 10



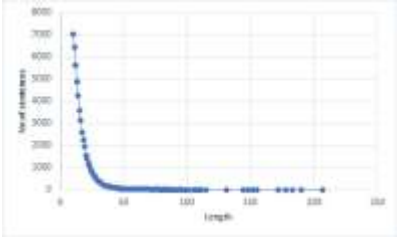
(h) Class 11



(i) Class 12



(j) UG



10.48047/jocaaa.2023.31.04.76

(k) PG

Fig. 1. Dataset Statistics

It can be observed from the graphs that the dataset contains some excessively long sentences which is very unlikely particularly for lower grades. We manually inspected some such cases and found that The Tesseract OCR based text extraction missed some sentence terminators in the Bengali language, resulting in those long sentences with excessively long lengths in lower grade levels like class 4.

4. METHODOLOGY

We performed three types of classification on the dataset to gain a comprehensive understanding of the task. First, we carried out fine-grained classification on the entire dataset, where the dataset was divided into 11 classes representing different grades ranging from class 4 to post-graduate levels. The fine-grained classes are class-4 (C4), class-5 (C5), class-6 (C6), class-7 (C7), class-8 (C8), class-9 (C9), class-10 (C10), class-11 (C11), class-12 (C12), undergraduate (UG) and post-graduate (PG). The second approach is a coarse-grained classification, which consolidated the classes into five broader categories: primary school (PS), middle school (MS), high school (HS), UG and PG. The PS category encompassed C4 and C5, the MS category included C6, C7 and C8, and the HS category comprised C9, C10, C11, and C12. UG and PG were considered as fine-grained classes themselves, i.e., they were not merged, since they are 3-year and 2-year courses, respectively. Finally, the third classification problem involved fine-grained classification within each of the coarse-grained classes.

For the classification task itself, there was a variety of machine learning algorithms that were used. In terms of classification algorithms, we considered both classical machine learning based as well as deep neural network based classification techniques. In case of the classical machine learning based classification algorithms, we considered Nu SVC, Linear SVC, SVC, Logistic Regression, Random Forest, Ridge CV, SGD, Passive Aggressive, Gradient Boosting, Decision Tree and Ada Boost. We also considered deep neural network based models like LSTM, Bi-LSTM, CNN and RNN [16].

The BERT (Bidirectional Encoder Representations from Transformers) model [17] has been very successful for a lot of NLP applications and usually BERT based models give the best results for most of the NLP applications. Hence, we considered some BERT based models viz. Neuralspace-Reverie-Bert, Bangla-Bert-Base [18], Neuralspace-Reverie -Roberta, Indic-Bert [19], Multilingual-Bert [20] and Bangla-Electra-Bert.

4.1 Data Preprocessing

Data preprocessing is the initial step of our research, where we first filter the dataset of all empty segments. Stop words and punctuation are removed from the sentences to leave only the meaningful tokens in order to train the machine learning algorithms. Popular search engines like Google and Yahoo exclude stop words and punctuation to retrieve data from their database quickly and accurately. We did not remove stop words before running the neural network methods. It guarantees prospective model performance enhancements and improves classification accuracy.

We found that the segments in our dataset are not all the same length, which makes it difficult to use neural network-based models. Some of them have numerous words, while others have few words. This caused a noticeable fall in accuracy. The segments are changed into equal lengths to address this issue. We define this length as being equal to 100 which offers the highest level of accuracy. Zeros are padded to the shorter segments while words are trimmed from the longer portions. As a result, each segment will be 100 words in length. The first 100 words of a segment that are longer than 100 words are shortened to make them all the same length for modeling, while 0s are added to segments that are less than 100 words long. The dataset is required for two reasons. We should first use some data to train a model, and then test it with some new data to see how well it works. By dividing the original dataset into training and test datasets, we may achieve this.

So, we divide the dataset into two subsets in the following stage. The first subset is known as train data, whereas the second group is known as test data. The data used to train the model

10.48047/jocaaa.2023.31.04.76

contains 90% of the randomly selected data from the original dataset. The remaining 10% of the data is made up of test data, which is used to verify the model's predictions. Detailed distribution of the train and test dataset is presented in Table 6.

Grade	#Segments in Original Dataset (% of overall data)	#Segments in Train Dataset (% of data)	#Segments in Test Dataset (% of data)
C4	989 (1.75 %)	888 (89.79 %)	101 (10.21 %)
C5	900 (1.60 %)	801 (89.00 %)	99 (11.00 %)
C6	1,318 (2.34 %)	1,188 (90.14 %)	130 (9.86 %)
C7	1,621 (2.88 %)	1,482 (91.43 %)	139 (8.57 %)
C8	2,157 (3.83 %)	1,933 (89.62 %)	224 (10.38 %)
C9	1,929 (3.42 %)	1,742 (90.31 %)	187 (9.69 %)
C10	1,458 (2.59 %)	1,285 (88.13 %)	173 (11.87 %)
C11	1,448 (2.57 %)	1,301 (89.85 %)	147 (10.15 %)
C12	1,479 (2.62 %)	1,335 (90.26 %)	144 (9.74 %)
UG	21,896 (38.85 %)	19,669 (89.83 %)	2,227 (10.17 %)
PG	21,171 (37.56 %)	19,105 (90.24 %)	2,066 (9.76 %)
Overall	**56,366**	**50,729 (90.00 %)**	**5,637 (10.00 %)**

Table 6. Distribution of Train and Test Dataset

4.2 Experimental Setup

All the models are created using Python-3 in the Google Colab environment with the assistance of NVIDIA-SMI GPU and 32 GB of RAM. The hyper parameters of the evaluated models are as follows.

The vocabulary size is 332576 used for creating the models. For the machine learning models, we used TF-IDF vectorizer. In our case, the pre-trained word-to-vec model with the dimension of 300 is used for the neural network models. The pre-trained model already contains some of Google News data set training (around 100 billion words). The model represents three million words and phrases in 300-dimensional vectors.

The Deviational Dropout is performed using the parameter SpatialDropout1D equal to 0.2. In the current version, the whole 1D feature maps are dropped instead of dropping certain parts. In order to carry out the multi-class classification task, the activation function, which is nothing else but the node determining whether a neuron fires or not, is used as the softmax. It normalizes the input value into the vector of values and probability distribution, which sum up to 1. The loss function is sparse categorical crossentropy because our model needs to recognize the class of defined grades that our segment belongs to. This particular loss function allows measuring the difference between two probability distributions. 20 epochs are determined as the optimal quantity for obtaining the maximum accuracy. Small batches have a huge impact on improving the performance of model training with unseen data [21]. That is why, in our case, the batch size equals 16. Adam [22] optimizer is used for adaptive learning rate optimization. Having successfully built the models and got our data set ready, the next step will be to train and test them.

5. RESULTS

Several machine learning, deep learning, and BERT based models were employed in classifying the study materials into different categories. The models were trained using our training dataset and tested on the test set (cf. Table 6). To determine the relative efficiency of the models used, we assess model performance on the basis of accuracy, precision, recall, and F-score.

The performance for fine-grained classification into 13 classes is shown in Table 7 below. The ML models exhibited varied performances where Nu SVC recorded the best precision, recall, and F-score among them. Linear SVC, SVC, and Logistic Regression recorded similar performance. Models such as Random Forest, Ridge CV, SGD, and Passive Aggressive models performed better, but lower than the above mentioned models. Gradient Boosting, Decision Tree, and Ada Boost models performed poorer as compared to the rest of the models. In the category of neural network models, LSTM and Bi-LSTM performed excellently. RNN and CNN performed decently. BERT models, especially Neuraspace-Reverie-Bert performed better than the rest of the models in terms of precision, recall, and F-score.

10.48047/jocaaa.2023.31.04.76

	Models	Precision	Recall	F-Score	Accuracy
Machine Learning Models	Nu SVC	**0.68**	**0.69**	**0.69**	**69.13**
	Linear SVC	0.59	0.61	0.60	60.78
	SVC	0.60	0.60	0.60	60.09
	Logistic Regression	0.58	0.60	0.59	59.84
	Random Forest	0.59	0.59	0.59	59.50
	Ridge CV	0.59	0.59	0.59	59.16
	SGD	0.56	0.58	0.57	57.96
	Passive Aggressive	0.56	0.57	0.57	57.18
	Gradient Boosting	0.52	0.53	0.52	52.55
	Decision Tree	0.46	0.48	0.47	48.41
	Ada Boost	0.43	0.41	0.42	41.49
Neural Network Models	LSTM	**0.67**	**0.68**	**0.68**	**68.17**
	Bi-LSTM	0.66	0.65	0.66	66.32
	RNN	0.62	0.64	0.63	63.11
	CNN	0.64	0.62	0.63	63.02
BERT Models	Neuralspace-Reverie- Bert	**0.95**	**0.95**	**0.95**	**95.09**
	Bangla-Bert-Base	0.82	0.82	0.82	83.16
	Neuralspace-Reverie-Roberta	0.77	0.77	0.77	76.93
	Indic-Bert	0.71	0.72	0.72	74.77
	Bangla-Electra	0.67	0.70	0.68	72.78
	Multilingual-Bert	0.70	0.71	0.71	71.15

Table 7. Evaluation of Fine-Grained Classification Models

Table 8 presents the performance for classification on the 5 coarse-grained classes - PS, MS, HS, UG and PG. A similar trend is observed for coarse-grained classification as in fine-grained classification. Among the machine learning models, Nu SVC achieved the highest precision of 0.68, indicating its ability to accurately classify instances. Linear SVC, SVC, and Logistic Regression showed similar precision, recall, and F-score values.

	Models	Precision	Recall	F-Score	Accuracy
Machine Learning Models	Nu SVC	**0.68**	**0.68**	**0.68**	**68.47**
	Linear SVC	0.60	0.61	0.60	60.90
	SVC	0.60	0.60	0.60	60.47
	Logistic Regression	0.60	0.60	0.60	60.45
	Random Forest	0.59	0.59	0.59	59.42
	Ridge CV	0.60	0.61	0.60	60.50
	SGD	0.59	0.59	0.59	58.96
	Passive Aggressive	0.57	0.57	0.57	57.06
	Gradient Boosting	0.53	0.53	0.53	52.57
	Decision Tree	0.46	0.47	0.47	47.73
	Ada Boost	0.43	0.43	0.43	42.47
Neural Network Models	LSTM	**0.77**	**0.65**	**0.71**	**55.54**
	Bi-LSTM	0.67	0.65	0.66	53.03
	RNN	0.62	0.64	0.63	50.94
	CNN	0.66	0.62	0.64	50.87

10.48047/jocaaa.2023.31.04.76

BERT Models	Neuralspace-Reverie-Bert	**0.98**	**0.98**	**0.98**	**98.31**
	Bangla-Bert-Base	0.85	0.85	0.85	84.64
	Neuralspace-Reverie-Roberta	0.81	0.80	0.81	80.49
	Indic-Bert	0.82	0.82	0.82	82.10
	Bangla-Electra	0.77	0.77	0.77	76.88
	Multilingual-Bert	0.76	0.76	0.76	75.63

Table 8. Evaluation of Coarse-Grained Classification Models

Random Forest, Ridge CV, and SGD models performed well, while Passive Aggressive, Gradient Boosting, Decision Tree, and Ada Boost models had lower accuracy. Among neural network models, LSTM got the highest precision and F-score, followed by Bi-LSTM, RNN, and CNN models. Among the BERT models, Neuralspace-Reverie-Bert outperformed all other models in terms of both F-score and accuracy. Other BERT based models also showed strong performance producing better scores than the machine learning and neural network models.

Table 9 presents the results for fine-grained classification within the PS category (C4 and C5). Among the machine learning models, Nu SVC produced the highest precision, recall, F-score and accuracy. In case of neural network models, LSTM produced the best results across all the evaluation metrics. Among the BERT models, Neuralspace-Reverie-Bert outclassed all other BERT models by significant margins producing the highest scores in all metrics. Although fine-grained classification within the PS category involves binary classification between C4 and C5, it proves to be a relatively difficult task to distinguish between these 2 classes.

	Models	**Precision**	**Recall**	**F-Score**	**Accuracy**
Machine Learning Model	Nu SVC	**0.68**	**0.68**	**0.68**	**68.25**
	SVC	0.67	0.68	0.68	67.55
	Logistic Regression	0.67	0.67	0.67	67.20
	Linear SVC	0.65	0.65	0.65	65.43
	Ridge CV	0.65	0.65	0.65	65.43
	Random Forest	0.65	0.65	0.65	64.73
	SGD	0.63	0.63	0.63	63.32
	Passive Aggressive	0.63	0.63	0.63	62.96
	Gradient Boosting	0.63	0.63	0.63	62.96
	Ada Boost	0.62	0.62	0.62	61.73
	Decision Tree	0.57	0.57	0.57	56.61
Neural Network Model	LSTM	**0.58**	**0.55**	**0.57**	**55.03**
	Bi-LSTM	0.57	0.53	0.55	53.54
	RNN	0.55	0.51	0.53	50.94
	CNN	0.55	0.51	0.53	50.87
BERT Models	Neuralspace-Reverie-Bert	**0.94**	**0.94**	**0.94**	**94.18**
	Bangla-Bert-Base	0.80	0.80	0.80	79.89
	Neuralspace-Reverie-Roberta	0.80	0.80	0.80	79.89
	Indic-Bert	0.75	0.75	0.75	74.60
	Bangla-Electra	0.78	0.63	0.70	77.78
	Multilingual-Bert	0.65	0.63	0.64	62.96

Table 9. Evaluation of Fine-Grained Primary School Models

Table 10 and Table 11 present the performance for fine-grained classification within the MS category (C6, C7 and C8) and HS category (C9, C10, C11, and C12), respectively. Fine-grained classification of MS and HS categories show the exact same trend as fine-grained classification of the PS category,

10.48047/jocaaa.2023.31.04.76

with NUSVC, LSTM and Neuralspace-Reverie-Bert proving to be the best models among the machine learning models, neural network models and BERT models, respectively, for this task. Among the fine-grained classification within PS, MS and HS categories, the HS category proved to be the most difficult one, as the results suggest.

The results of both fine-grained and coarse-grained classification presented in Tables 7–11 show that BERT models produce higher classification accuracy than other models for our task. In particular, Neuralspace-Reverie-Bert outperforms all other models, including the other BERT models. In order to comprehend the task better, we studied class-wise performance of the models. Table 12 presents the performance of the Neuralspace-Reverie-Bert model across different classification tasks. Class-wise performance scores reported in Table 12 suggest that the model identifies UG and PG level study materials pretty accurately. It is to be noticed that UG and PG classes dominate the dataset, together they constitute over 76% of the dataset. Table 12 also suggests that study materials belonging to lower grades are difficult to be distinguished, both in case of fine-grained classification (C6, C4) and coarse-grained classification (PS).

	Models	**Precision**	**Recall**	**F-Score**	**Accuracy**
Machine Learning Model	Nu SVC	**0.63**	**0.64**	**0.63**	**63.51**
	Logistic Regression	0.63	0.63	0.63	63.24
	SVC	0.61	0.62	0.61	61.74
	Ridge CV	0.61	0.62	0.61	61.74
	Linear SVC	0.61	0.61	0.61	61.22
	SGD	0.60	0.61	0.61	60.89
	Gradient Boosting	0.57	0.57	0.57	56.97
	Passive Aggressive	0.59	0.59	0.59	59.25
	Random Forest	0.58	0.59	0.59	59.06
	Ada Boost	0.54	0.54	0.54	53.70
	Decision Tree	0.50	0.51	0.50	50.56
Neural Network Model	LSTM	**0.52**	**0.50**	**0.51**	**50.20**
	Bi-LSTM	0.51	0.49	0.50	48.84
	RNN	0.48	0.46	0.47	46.47
	CNN	0.48	0.46	0.47	46.41
BERT Models	Neuralspace-Reverie-Bert	**0.96**	**0.96**	**0.96**	**95.88**
	Bangla-Bert-Base	0.80	0.80	0.80	79.89
	Neuralspace-Reverie-Roberta	0.80	0.79	0.80	79.41
	Indic-Bert	0.72	0.70	0.71	70.39
	Bangla-Electra	0.64	0.67	0.65	66.67
	Multilingual-Bert	0.71	0.69	0.70	69.22

Table 10. Evaluation of Fine-Grained Middle School Models

	Models	**Precision**	**Recall**	**F-Score**	**Accuracy**
Machine Learning Model	Nu SVC	**0.60**	**0.60**	**0.60**	**59.95**
	Logistic Regression	0.58	0.58	0.58	57.63
	Ridge CV	0.58	0.57	0.58	57.47
	Linear SVC	0.57	0.57	0.57	56.94
	SVC	0.57	0.57	0.57	56.73

10.48047/jocaaa.2023.31.04.76

	SGD	0.57	0.57	0.57	56.68
	Random Forest	0.54	0.54	0.54	54.04
	Gradient Boosting	0.55	0.53	0.54	53.03
	Passive Aggressive	0.52	0.52	0.52	51.98
	Ada Boost	0.48	0.47	0.47	46.65
	Decision Tree	0.40	0.40	0.40	40.42
Neural Network Model	LSTM	**0.51**	**0.50**	**0.50**	**50.47**
	RNN	0.47	0.47	0.47	46.73
	Bi-LSTM	0.50	0.48	0.49	49.11
	CNN	0.48	0.46	0.47	46.67
BERT Models	Neuralspace-Reverie-Bert	**0.93**	**0.93**	**0.93**	**92.55**
	Bangla-Bert-Base	0.81	0.81	0.81	80.51
	Neuralspace-Reverie-Roberta	0.78	0.78	0.78	77.97
	Multilingual-Bert	0.62	0.62	0.62	62.12
	Bangla-Electra	0.61	0.59	0.60	59.43
	Indic-Bert	0.37	0.41	0.39	41.36

Table 11. Evaluation of Fine-Grained High School Models

To obtain further information about the performance of the Neuralspace-Reverie-Bert classifier, especially regarding the misclassification cases, confusion matrices for the fine-grained and coarsegrained classification are presented in Tables 13 and 14, respectively. In a similar way, confusion matrices for fine-grained classification for the 3 coarse-grained categories (PS, MS, and HS) are presented in Tables 15, 16, and 17, respectively. The confusion matrices give us an insight into how the model performs its predictions. By analyzing these matrices, we will be able to see any pattern in the classifications made by the model.

	Grade	**Precision**	**Recall**	**F-Score**
Fine Grained	PG	**0.98**	**0.98**	**0.98**
	UG	0.97	0.99	0.98
	C8	0.89	0.93	0.91
	C10	0.92	0.9	0.91
	C9	0.94	0.82	0.88
	C12	0.85	0.92	0.88
	C7	0.83	0.86	0.84
	C5	0.79	0.86	0.83
	C11	0.79	0.78	0.78
	C4	0.86	0.65	0.74
	C6	0.84	0.64	0.73
Coarse Grained	PG	**0.99**	**0.99**	**0.99**
	UG	0.98	0.99	0.99
	HS	0.96	0.98	0.97
	MS	0.97	0.95	0.96
	PS	0.98	0.94	0.96
Fine Grained Primary	C4	**0.96**	**0.93**	**0.94**
	C5	0.93	0.96	0.94
Fine Grained Middle	C8	**0.98**	**0.98**	**0.98**

10.48047/jocaaa.2023.31.04.76

	C6	0.94	0.94	0.94
	C7	0.94	0.93	0.94
Fine Grained High	C9	**0.94**	**0.94**	**0.94**
	C10 C12	0.95 0.9	0.9 0.94	0.93 0.92
	C11	0.91	0.91	0.91

Table 12. Class-wise performance of Neuralspace-Reverie-Bert

Predicted \ Actual	C4	C5	C6	C7	C8	C9	C10	C11	C12	UG	PG
C4	57	10	6	7	0	2	2	1	2	0	1
C5	3	83	4	2	1	0	0	0	0	3	0
C6	2	7	74	14	8	1	1	1	1	2	4
C7	2	4	1	138	1	0	3	2	1	6	2
C8	0	1	1	0	213	0	0	5	3	6	0
C9	0	0	1	1	7	159	0	12	7	3	3
C10	1	0	0	1	1	2	121	1	5	2	1
C11	1	0	0	1	3	5	1	111	6	5	10
C12	0	0	0	2	3	0	4	3	144	1	0
UG	0	0	1	1	0	1	0	1	1	2166	14
PG	0	0	0	0	2	0	0	4	0	36	2094

Table 13. Confusion Matrix of Fine-Grained Neuralspace-Reverie-Bert

Predicted \ Actual	PS	MS	HS	UG	PG
PS	196	6	4	2	1
MS	5	446	14	4	2
HS	0	4	619	6	4
UG	0	0	7	2168	4
PG	0	6	0	26	2113

Table 14. Confusion Matrix of Coarse-Grained Neuralspace-Reverie-Bert

Predicted \ Actual	C4	C5
C4	89	7
C5	4	89

Table 15. Confusion Matrix of Fine-Grained Primary School Neuralspace-Reverie-Bert

Predicted \ Actual	C6	C7	C8
C6	110	6	1
C7	6	140	4
C8	1	3	239

10.48047/jocaaa.2023.31.04.76

Table 16. Confusion Matrix of Fine-Grained Middle School Neuralspace-Reverie-Bert

6. CONCLUSION

In this paper, we presented classification of Bengali literature study materials for students at various academic levels, from grade 4 to post-graduate level. Distribution of study materials in a customized manner is important to ensure an effective learning experience for students with varied levels of competence as the education field is adopting digital platforms. Classification of Bengali literary study materials into several grade levels will assist in satisfying the important demand for effectively matching content with the cognitive abilities of the learners of Bengali literature. Moreover, the incorporation of both fine-grained and coarse-grained classification strategies can further enrich the understanding of text complexity and readability. Fine-grained classification allows for a nuanced categorization, enabling educators to match study materials more precisely with the aptitude level of the students, while coarsegrained classification provides a wider but still meaningful view of text complexity, making it easier to distribute content to larger student groups.

Due to unavailability of suitable data, we created a substantially large dataset utilizing textbooks from various grade levels. All of these books are provided for the students by the respective boards and universities. We carried out experiments using a number of machine learning, neural networks, and BERT models. Neuralspace-Reverie-Bert produces the best overall performance for this task.

In future, we intend to expand this work by including other domains like science, history, geography, etc. We would also like to curate dataset for other languages and replicate the same experiments in those languages.

REFERENCE

[1]. Hancke, J., Vajjala, S., Meurers, D.: Readability classification for german using lexical, syntactic, and morphological features. In: Proceedings of COLING 2012, pp. 1063–1080 (2012)

[2]. Eickhoff, C., Serdyukov, P., De Vries, A.P.: A combined topical/non-topical approach to identifying web sites for children. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 505–514 (2011)

[3]. Das, S., Roychoudhury, R.: Readability modelling and comparison of one and two parametric fit: A case study in bangla. Journal of Quantitative Linguistics 13(01), 17–34 (2006)

[4]. Islam, Z., Mehler, A., Rahman, R.: Text readability classification of textbooks of a lowresource language. In: Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, pp. 545–553 (2012)

[5]. Islam, Z., Rahman, M., Mehler, A., et al.: Readability classification of bangla texts. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 507–518 (2014). Springer

[6]. Sinha, M., Basu, A.: A study of readability of texts in bangla through machine learning approaches. Education and information technologies 21(5), 1071–1094 (2016)

[7]. Vajjala, S., Meurers, D.: On improving the accuracy of readability classification using insights from second language acquisition. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 163–173 (2012)

[8]. Feng, L., Elhadad, N., Huenerfauth, M.: Cognitively motivated features for readability assessment. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 229–237 (2009)

[9]. Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M.: Combining lexical and grammatical features to improve readability measures for first and second language texts. In: Human Language Technologies 2007: the Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 460–467 (2007)

10.48047/jocaaa.2023.31.04.76

[10]. Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., Welty, C.: Learning to predict readability using diverse linguistic features. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 546–554 (2010)

[11]. Pitler, E., Nenkova, A.: Revisiting readability: A unified framework for predicting text quality. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 186–195 (2008)

[12]. Collins-Thompson, K., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 193–200 (2004)

[13]. Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 523–530 (2005)

[14]. Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 629–633 (2007). IEEE

[15]. Hasnat, M., Habib, S., Khan, M., et al.: A high performance domain specific ocr for bangla script. In: Novel Algorithms and Techniques in Telecommunications, Automation and Industrial Electronics, pp. 174–178. Springer, (2008)

[16]. Negnevitsky, M.: Artificial intelligence: a guide to intelligent systems third edition (2011)

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

[18]. Sarker, S.: Banglabert: Bengali mask language model for bengali language understading. (2020). https://github.com/sagorbrur/bangla-bert

[19]. Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M.M., Kumar, P.: IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In: Findings of EMNLP (2020)

[20]. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018) arXiv:1810.04805

[21]. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 (2016)

[22]. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)