BengaliLCP: A Dataset for Lexical Complexity Prediction in the Bengali Texts

Nabila Ayman, Md. Akram Hossain*, **Abdul Aziz***,
Rokan Uddin Faruqui, and **Abu Nowshed Chy**
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{nabila.ayman.cu, akram.hossain.cse.cu, aziz.abdul.cu}@gmail.com
{rokan, nowshed}@cu.ac.bd

Abstract

Encountering intricate or ambiguous terms within a sentence produces distress for the reader during comprehension. Lexical Complexity Prediction (LCP) deals with predicting the complexity score of a word or a phrase considering its context. This task poses several challenges including ambiguity, context sensitivity, and subjectivity in perceiving complexity. Despite having 300 million native speakers and ranking as the seventh most spoken language in the world, Bengali falls behind in the research on lexical complexity when compared to other languages. To bridge this gap, we introduce the first annotated Bengali dataset, that assists in performing the task of LCP in this language. Besides, we propose a transformer-based deep neural approach with a pairwise multi-head attention mechanism and LSTM model to predict the lexical complexity of Bengali tokens. The outcomes demonstrate that the proposed neural approach outperformed the existing state-of-the-art models for the Bengali language.

Keywords: Lexical Complexity Prediction, LCP, Bengali, Dataset, Transformer, LSTM, XLM-RoBERTa

1. Introduction

Complex or rarely used words create difficulty while reading articles, that makes the reader misinterpret the context, or trudge on without understanding. Identifying and assigning the complexity score to particular words can lead to the solution by allowing the replacement of complex terms with simpler alternatives. The significance of this task lies in helping second language learners as well as native learners while anticipating complex literature. The first shared task in this arena, Complex Word Identification (CWI) is a binary task that identifies a word as complex or simple based on its context (Paetzold and Specia, 2016). In SemEval 2021, Shardlow et al. (2021) introduced the Lexical Complexity Prediction (LCP) task. It is the process of predicting the complexity score of a word or phrase in a sentence based on the contextual meaning of that word. Lexical simplification applications require selecting complex words to replace them with simpler synonyms and the LCP task can play an important role in selecting replaceable complex words. Moreover, there are other applications of LCP including readability assessment, language education, text generation, and cross-linguistic studies.

Bengali, also known as Bangla, is an Indo-Aryan language primarily spoken in the eastern region of the Indian subcontinent, including Bangladesh and the Indian states of West Bengal, Tripura, and Assam. With over 300 million speakers, it is the seventh most spoken language in the world. It is the language of education, government, media, and literature in Bangladesh and West Bengal. The task of conducting readability analysis in Bengali has been a significant focus of numerous research endeavors (Das and Roychoudhury, 2006; Islam et al., 2014; Chakraborty et al., 2021). Lexical complexity prediction can be an extended section of the readability analysis task that can assess the complexity of the word based on its context.

There is no standard resource in the Bengali language for conducting the work of Lexical Complexity Prediction employing NLP models. Text simplification techniques used in English are inapplicable to Bengali as there are numerous syntactic and lexical differences between Bengali and English. Some of these differences are listed below:

- Bengali and English are members of two distinct language families: Indo-European and West-Germanic.
- Almost all readability metrics in English consider polysyllabic words to be hard words, but in Bengali, polysyllabic words are familiar.
- Bengali is a language with a flexible word order that allows a variety of grammatically correct surface forms.
- The major distinction between Bengali and English syntax is in word order. The basic word sequence of English is subject-verb-object,

*Authors contributed equally to this work.

LREC-COLING 2024, pages 2227–2237
20-25 May, 2024.

whereas the basic word order of Bengali is subject-object-verb.

- Multi-word expressions (MWE) in the CompLex dataset (Shardlow et al., 2020) contain adjective-noun and noun-noun phrases whereas Bengali sentences have a small presence of these types of patterns.

These disparities between the English and Bengali languages motivated us to present the problem of predicting lexical complexity in Bengali. Following the SemEval 2021 Task 1: Lexical Complexity Prediction introduced by Shardlow et al. (2021), we formulated our Bengali LCP task into two sub-tasks: sub-task 1: Predicting the complexity score of single words (SW) and sub-task 2: Predicting the complexity score of multi-word expressions (MWE). The main contributions of this paper are listed below:

- Construct the first annotated lexical complexity prediction resource in the Bengali language, BengaliLCP dataset with 3033 sentences collected from newspapers and Wikipedia and make it publicly available[1].
- Develop a deep neural lexical complexity prediction system, PALCP (Pairwise Attention based Lexical Complexity Prediction), integrating contextual word embeddings of the XLM-RoBERTa model, pairwise multi-head attention features of sentence-word pair, and the LSTM model.

The remainder of the paper is structured as follows: We provide our analysis with related studies in section 2. Section 3 describes the process of creating the BengaliLCP dataset. In section 4, we describe our proposed system. The information about the evaluation and experimental result of the system is depicted in section 5. We conduct a discussion in section 6, and conclude our work in section 7.

2. Related Work

Lexical complexity causes a text to reduce its readability and predicting lexical complexity has been the subject of extensive research in the field of NLP. A number of related works took place in the arena of the English language that created resources and developed models. Paetzold and Specia (2016) contributed to Task 11 of SemEval 2016 with the CWI task where they unveiled the first CWI dataset for identifying complex words in the text which is the first stage of text simplification. Instead of employing binary categories, a continuous annotation enables a ranking to be assigned to words, allowing us to determine not only whether a word is difficult for a reader but also how difficult that word is likely to be. In SemEval 2021, Shardlow et al. (2021) introduced a task to predict the complexity score for specific words in the sentence in the English language. This task utilized the CompLex dataset (Shardlow et al., 2020). Participants of this task have proposed various feature-based systems, deep learning systems, and systems that use a concatenation of the former two approaches.

Mosquera (2021) implemented 51 hand-crafted features (HCF) and utilized Light GBM implementation of gradient tree boosting. Ortiz-Zambrano and Montejo-Ráez (2021) proposed a system with 15 HCF based on the frequency of the words, and a supervised random forest regression algorithm is trained over these features. Along with these feature-based systems deep learning systems have been explored. Aziz et al. (2021) proposed pairwise learning with a transformers-based system where they exploited sentence pair regression with BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models. Pan et al. (2021) composed fine-tuning of several pre-trained language models including BERT, ALBERT (Lan et al., 2020), RoBERTa, and ERNIE (Zhang et al., 2019) with different training strategies such as pseudo labeling and data augmentation, and stacked them with a simple linear regression model. Stodden and Venugopal (2021) proposed a system that combines HCF, contextualized character embedding, a sense of relative normalization, and a neural network for regression.

Related work on the Bengali language in this arena is very limited. Numerous studies have addressed the broader field of readability assessment while predicting lexical complexity represents a more specific subdomain within this context. Das and Roychoudhury (2006) explored readability modeling and compared one and two parametric fit models for Bengali. They evaluated their system using several readability indices, including the Flesch Reading Ease score and the Automated Readability Index. Islam et al. (2012) implemented a readability classifier of Bengali textbook documents. They implemented a baseline system using various lexical features and three traditional readability formulas Gunning fog readability index, Dale–Chall readability formula, and the automated readability index (Senter and Smith, 1967). Islam et al. (2014) focused on the readability classification of Bangla texts and developed a machine learning-based approach using a corpus of Bangla texts and a set of manually selected features. Sinha and Basu (2016) applied classification and regression approaches to predict the readability of the Bengali language with SVM and SVR. Phani et al. (2019) conducted a readability analysis on Bengali passages and proposed eleven readability analysis models based on regression.

Some recent works implemented text simplifica-

[1] https://csecu-dsg.github.io/resources/

tion in the Bengali language. Hossain and Ahnaf (2021) proposed a system that simplifies the text, keeping the context unchanged, and achieved the best performance by fine-tuning the BERT model in their corpus. Mahata et al. (2022) proposed a method for improving the quality of machine translation between English and Bengali through the use of sentence simplification. The authors developed a hybrid system that combines rule-based and machine learning-based approaches for identifying complex sentences and simplifying them. Chakraborty et al. (2021) used document-level datasets from NCTB textbooks and performed supervised binary sentence classification of whether a sentence is complex or not. They obtained the best performance from the combination of BiLSTM, CC (Conjunction Count), CL (Character Length), and embeddings from the Language-agnostic BERT sentence embedding model.

Despite notable advancements in the field of lexical complexity prediction, particularly within the realm of the English language, a discernible gap becomes evident when we turn our attention to the Bengali language. Although there has been some scrutiny of readability analysis in Bengali, especially concerning educational materials, the task of predicting lexical complexity in Bengali has remained relatively unexplored. To broaden the horizons of the LCP task in Bengali, we have created a Bengali annotated corpus. We also propose a pairwise multi-head attention-based neural approach to exploit the intricate relationship between the target token and its context.

3. BengaliLCP Dataset

In this section, we describe the dataset creation, and annotation process along with dataset statistics and inter-annotator correlation.

3.1. Dataset Creation

For the standard LCP dataset, there are some standard qualities: continuous annotations, context-based complexity, multiple token instances, and diverse genre (Shardlow et al., 2022). We have followed these specifications and selected sources that contain an adequate amount of complex Bengali vocabulary along with standard articles. To build the corpus for the BengaliLCP dataset, we chose the Prothom Alo[2] news portal and the Bengali Wikipedia[3]. The number of contexts we collected from these sources is listed in Table 1.

[2]https://www.prothomalo.com/
[3]https://g.co/kgs/xb7AfK

Domain	Number of contexts
Newspaper	1496
Wikipedia	1537
Total	3033

Table 1: Number of contexts in each sub-corpus.

Prothom Alo: Bangladeshi daily newspaper, Prothom Alo is written in Bengali and published in Dhaka, Bangladesh. From newspapers, we can collect sentences that are constructed with formal vocabulary. The literature used in newspapers tends to be precise, straightforward, and accessible to a wide range of readers. We collected texts from news between 13th June 2022 and 30th June 2022. We selected 12 different categories in this domain that are listed in Figure 1.

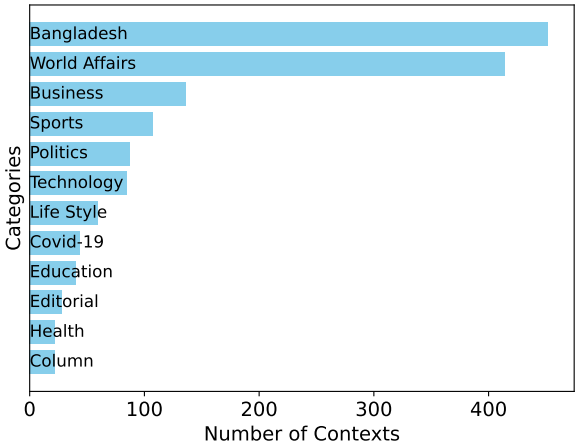


Figure 1: No. of context in newspaper categories.

Bengali Wikipedia: The Bengali Wikipedia is the version of the free online encyclopedia Wikipedia that is composed in Bengali. Compared to news portals, it contains more primitive words, which aids in the construction of a diversified dataset for lexical complexity prediction. Wikipedia covers a broad range of domains and contains standardized articles that allow for a comprehensive assessment of lexical complexity across different topics. We collected data from the Bengali version of Wikipedia during the time frame of May-July, 2022. Figure 2 shows the number of collected data from different categories in Wikipedia. The selected categories were chosen to represent a diverse range of linguistic contexts in Bengali.

Every corpus has its distinctive language features and patterns. For each context in the BengaliLCP corpus, we have selected single-word tokens and multi-word expressions as target tokens. In some instances, repetition of target tokens occurs as the same word can have differ-

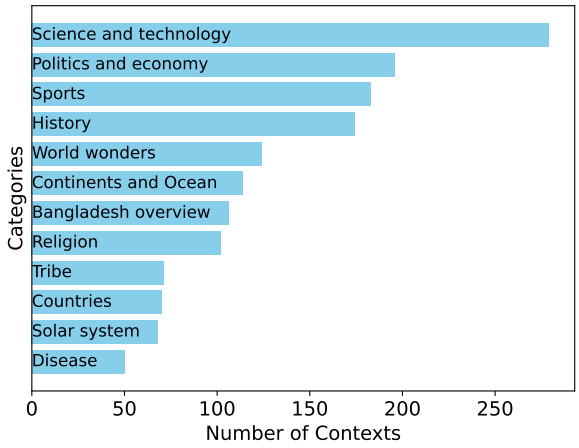


Figure 2: No. of context in Wikipedia categories.

ent complexity scores depending on its context. Parts of speech have a contribution while anticipating the meaning of a particular sentence. For generalizing, we selected nouns and verbs as single-word tokens. We constructed MWEs with two-word and three-word tokens. In CompLex (Shardlow et al., 2020), compound nouns such as adjective-noun and noun-noun patterns tend to be the most common phrase patterns used in English. In the case of Bengali, we have observed that noun phrases are not as common as in English. The common phrase patterns we found in Bengali are starting with nouns or adjectives. So we selected noun-noun, adjective-adjective, noun-adjective, and adjective-noun patterns for two-word tokens. For three-word tokens, we selected phrases that have a combination of adjectives and nouns. We used BNLP Bengali pos tagger (Sarker, 2021) to identify the parts of the speech tag of our target token. We selected target tokens based on their frequency of use as the use of words is correlated with the complexity of the word. From some contexts, we selected one word that is most frequent among all other words of that sentence and another word that is least frequent and can be labeled as complex by annotators. This characteristic of our dataset makes it diverse in complexity level so that NLP models can learn both complex and easy literature features.

3.2. Dataset Annotation

For predicting the lexical complexity of a specific token based on its context, a level-wise scoring tends to be more appropriate than binary annotation protocol (Shardlow et al., 2022). The reason behind this is binary annotation tends to label the word as simple or complex, but it can not express the level of complexity. Defining the complexity of words is very subjective, so it causes difficulty when people from different backgrounds have to agree on the binary level. Likert scale is a type of survey scale that ranges from one extreme level to another. In this work, we have used a five-point Likert scale where each point is defined as follows:

1. *Very Easy:* Words that the annotator recognized immediately.
2. *Easy:* Words whose meanings were known to the annotator.
3. *Neutral:* Words that were neither simple nor complex.
4. *Difficult:* Words whose definitions were ambiguous to the annotator but whose meanings could be deduced from the sentence.
5. *Very Difficult:* Words that annotators had never encountered before or that were incredibly ambiguous.

In the BengaliLCP dataset, three annotators annotated each SW token and MWE token based on the five-point Likert scale. To normalize the complexity score in the range (0-1), we conducted the following transformations of each annotation: 1 → 0, 2 → 0.25, 3 → 0.5, 4 → 0.75, 5 → 1. The final label for each token was then determined by averaging the normalized scores of all annotators. The eventual complexity score is a continuous score that indicates the level of complexity of the token based on its context. We included some instances of annotations in Table 2 along with their English translation that was obtained from Google Translate. More examples of annotated instances are provided in appendix A.

Text	Score
যদি ব্যক্তিদের উপরে স্বেচ্ছায় পরীক্ষা জন্য এগিয়ে আসার নির্দেশ দেওয়া হয়, তাহলে অন্তরণ পদক্ষেপটি সাধারণত সফল হয় না। (If individuals are ordered to come forward for voluntary testing, the isolation step is usually not successful.)	0.44
শীর্ষক অভিসন্দর্ভে তিনি লিখেছেন, 'সামাজিক সম্পর্কগুলোর যূথবদ্ধতাই মনুষ্যচরিত্রের সার।' (In the preface, he wrote, 'The cohesion of social relations is the essence of human character'.)	0.75

Table 2: Annotated instances of BengaliLCP dataset. Target token is underlined.

3.3. Corpus Statistics

From the statistics depicted in Table 3, we can incorporate that both the newspaper and Wikipedia domains contain an adequate amount of unique words. The corpus of Wikipedia has a slightly large number of unique words in their contexts.

However, the mean complexity of our SW dataset is 0.1379, and the MWE dataset is 0.2387. We included the labeling criteria in the earlier section that depicts 0.0 to be “very easy” and 0.25 to be “easy” labels for the target token. In our dataset, words are falling more into an easier scale. The mean complexity of Wikipedia data is slightly higher than the newspaper data that is 0.1391 in the SW dataset and 0.2479 in the MWE dataset. The standard deviation for the SW and MWE datasets are 0.2075 and 0.1931, respectively. This disparity suggests that texts from Wikipedia exhibit more complex features compared to newspaper data and cover a wide range of specialized and technical topics, demanding more intricate language patterns and vocabulary.

Task	Contexts	Unique words	Mean score	Standard deviation
Newspaper data				
SW	2292	1800	0.1366	0.1977
MWE	2159	2061	0.2281	0.1598
Wikipedia data				
SW	2425	1963	0.1391	0.2164
MWE	2457	2413	0.2479	0.2179
BengaliLCP dataset				
SW	4717	3504	0.1379	0.2075
MWE	4616	4459	0.2387	0.1931

Table 3: Statistics of BengaliLCP dataset.

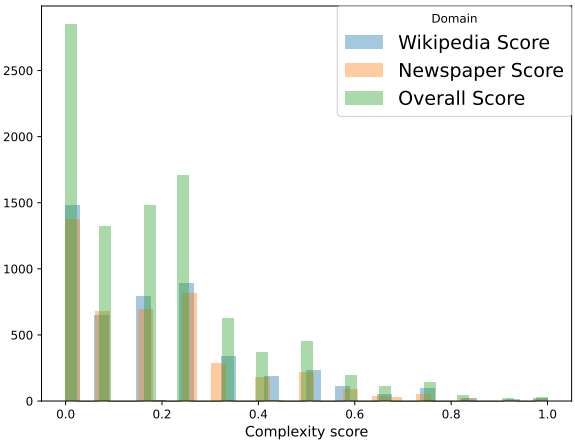


Figure 3: Distribution of complexity in the corpus.

In Figure 3, a ridge line plot showing the probability density function of the full dataset (overall) as well as each of the genres: Newspaper and Wikipedia, contained within the dataset. The plot shows that even though the majority of the probability mass is located to the left of the mid-point, there are still a large number of annotations on either side of the mid-point for both the sub-corpus. Given that the annotators are native Bengali speakers and that both the newspaper and Wikipedia utilize a straightforward form of Bengali, the tokens tend to fall into the easy category.

3.4. Inter-annotator Correlation

Three annotators participated in the BengaliLCP dataset annotation processing. The annotators come from three distinct study groups including undergraduate, graduate, and postgraduate, and speak Bengali as their first language, hence their backgrounds are similar. The scores assigned by each annotator can differ for the same instance because predicting complexity can depend on personal perspective. One annotator might rate a term as complex, whereas another annotator might identify the word and rate it as easy in their opinion. There are also cases where words may seem complex or ambiguous because of their context, but other readers can understand them easily. In order to check the correlation, we performed Cohen’s Kappa test of the annotations of the dataset as the following equation:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

Here, P_o is the observed agreement between raters, and P_e is the expected agreement due to chance.

Annotators	SW	MWE
A1 and A2	0.91	0.83
A1 and A3	0.57	0.47
A2 and A3	0.57	0.44
Average	0.68	0.58

Table 4: Cohen’s Kappa between annotators.

We illustrated the Cohen’s Kappa test result in Table 4. Here we can depict that the average of Cohen’s kappa score between annotators is in the range of 0.5-0.7 in the SW and MWE datasets, that indicates moderate to a fair agreement. It implies that the annotators have consistency between their observations. We have already discussed that predicting the complexity of a word based on its context is a highly perceptual task that depends a lot on individual perception. For this reason, it is not easy to have a high level of agreement. However, the moderate agreement observed in this case indicates a reasonable level of consistency and suggests that the annotations contribute significantly to the task.

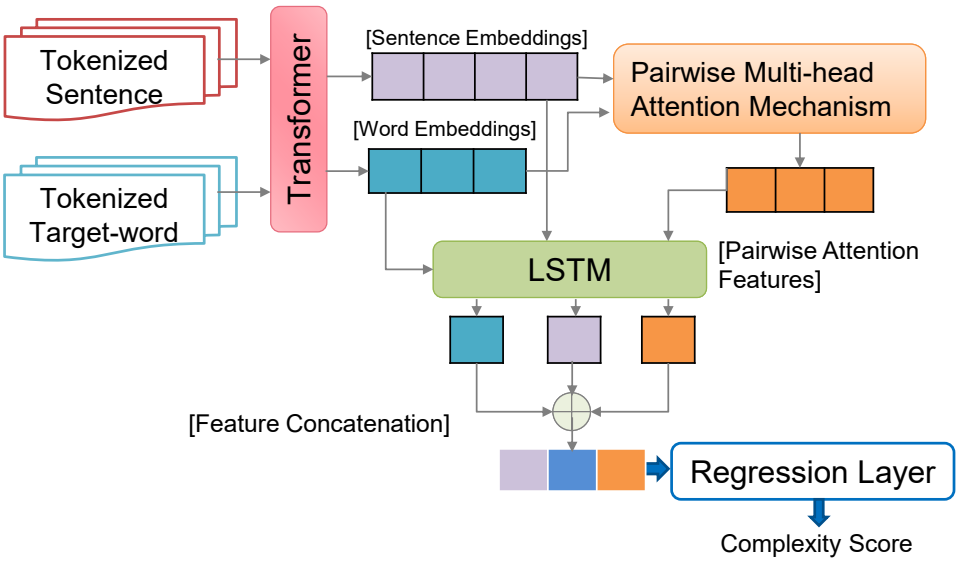


Figure 4: Proposed framework: The system employs XLM-RoBERTa to obtain embeddings for sentences and target words, then applies pairwise attention. These features are processed using LSTM, and the final regressor layer takes the concatenation of these features and predicts the complexity score.

4. Methodology

We implemented the Pairwise Attention Lexical Complexity Prediction (PALCP) system employing a transformer language model, pairwise multi-head attention mechanism, and LSTM network. Figure 4 illustrates the system for predicting lexical complexity. At first, we obtain sentence embeddings and target token embeddings from the XLM-RoBERTa transformer model. XLM-RoBERTa (Conneau et al., 2019) is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. Because of its pre-training in a wide range of languages, including Bengali, XLM-RoBERTa is better able to identify linguistic patterns and contextual information that is effective for downstream Bengali NLP tasks. We also utilized the XLM-RoBERTa tokenizer for breaking our input contexts into tokens. We extract contextual features of sentences and words separately from the last hidden layer of the transformer output representation. Then we apply a pairwise multi-head attention mechanism with the sentence and target token embeddings. After that, we process those embeddings with the prediction module and it produces the final predicted complexity score.

4.1. Pairwise Multi-head Attention

Vaswani et al. (2017) implemented a multi-head attention mechanism in the self-attention encoder decoder method of the transformer. In our model, we implement this attention mechanism to extract the dependency features between the sentence and the target token. The target token from which the complexity score needs to be predicted can have different dependencies with different parts of the sentence. This method allows our model to attend to different parts of the input sentence with different weights. That can achieve highly complex dependency features and the relationship between the target token with its context. The equation below represents the calculation of the pairwise multi-head attention mechanism with contextual sentence and word embeddings:

$$A = \text{softmax}(\frac{q \cdot k^\top}{\sqrt{head\_dim}}) \cdot v \quad (2)$$

Here, q and k are the linear representations of contextual sentence embeddings, and v represents the embeddings of the target token. The number of head dimensions is represented with $head\_dim$. Utilizing the softmax function, we get the attention weight. A is the attention feature of the target token based on the sentence.

4.2. Prediction Module

LSTM (long short-term memory) is a special type of recurrent neural network (Hochreiter and Schmidhuber, 1997). In our model, we processed the sentence embeddings and target embeddings from the transformer and their pairwise attention features with the LSTM network. LSTM captures sequential patterns and dependencies in the attended output, providing the model with a more fine-grained understanding of the contextual information of the input data. LSTM processes sentence embeddings, target token embeddings, and pairwise attention features and then we fed the concatenation of these features to the regression

layer. The equation of the regression module can be defined as follows:

$$Complexity\_score = W \cdot [T; A; S] + b \quad (3)$$

Here, T is the feature of the target token, S is the feature of the sentence, and A is the pairwise attention feature. The concatenation between features is represented by ';'. The output of the regression layer represents the predicted complexity score.

5. Experiments and Evaluations

This section depicts the result of our proposed model while predicting the lexical complexity of tokens in the Bengali language.

5.1. Evaluation Metric

To measure the performance of our system, we selected the evaluation measure according to the SemEval 2021, Task 1: Lexical Complexity Prediction (Shardlow et al., 2021). We utilized different evaluation metrics to measure the performance of systems: Pearson's correlation, Spearman's correlation, mean absolute error (MAE) and mean squared error (MSE). We chose Pearson correlation as the primary evaluation metric as it is effective when dealing with continuous output data.

5.2. Model Configuration

We experimented with different hyperparameters and versions of transformer models. We utilized Google Colaboratory GPU for training and evaluating our system. We employed huggingface library[4] versions of transformer models. Table 5 highlights the configuration of our best-performing system.

Hyperparameter	Configuration
learning rate	3e-5
batch size	16
epoch	2
head dimension	48
number of heads	16
LSTM input size	768
LSTM hidden size	60

Table 5: Hyper-parameter settings.

For tokenizing and encoding the Bengali data, we utilized xlm-roberta-base, bert-base-multilingual-cased, distilbert-base-multilingual-cased, sagorsarker/bangla-bert-base, and csebuetnlp/banglabert from the huggingface library. We conducted tuning with parameters epoch, learning rate, and batch size to procure the appropriate configuration. We tuned our system for epochs with a set of {2, 6, 9, 20}; learning rate with a set of {2e-5, 3e-5, 4e-5, 5e-5, 3e-6}; batch size with a set of {8, 16, 32}. Our system achieved the best performance with epoch 2, learning rate 3e-5, and batch size 16.

5.3. Experimental Result and Analysis

We implemented an HCF-based baseline system that included features such as word length, word frequency, and zipf frequency. Here, zipf frequency returns the word frequency on a human-friendly logarithmic scale. For word length feature, we calculated the total number of characters in target words. For word frequency and zipf frequency, we used the Wordfreq[5] API. We fed these features to various regressors such as xgboost, SVR, LightGBM, CatBoost, Linear regression, and Bayesian Ridge. The Pearson scores of the best three models along with our proposed system in both tasks of Bengali LCP are reported in Table 6. We also implemented another baseline with n-gram and TF-IDF features with various regressor models including xgboost, SVR, LightGBM, CatBoost, Linear regression, and Bayesian Ridge. Table 6 shows the Pearson scores of the best three n-gram-based baseline models for SW and MWE LCP along with our proposed system.

Model	SW LCP	MWE LCP
PALCP	0.6096	0.5460
HCF-based baseline systems		
CatBoost	0.5183	0.4509
Bayesian Ridge	0.5175	0.4897
Linear Regression	0.5167	0.4907
N-gram-based baseline systems		
CatBoost	0.1972	0.1356
Bayesian Ridge	0.1727	0.2032
Linear Regression	0.1935	0.2061

Table 6: Comparison of PALCP system with baseline systems.

We evaluated the BengaliLCP dataset using our proposed system and compared the performance with different state-of-the-art language models including XLM-RoBERTa (Conneau et al., 2019), Multilingual-BERT (Devlin et al., 2019), Bangla BERT-20 (Sarker, 2020), Multilingual-DistilBERT (Sanh et al., 2019), and BanglaBERT-22 (Bhattacharjee et al., 2022). We fine-tuned each language model using the specifications that we em-

[4]https://huggingface.co/docs/transformers/index

[5]https://pypi.org/project/wordfreq/

ployed in our PALCP model for fine-tuning XLM-RoBERTa, mentioned in Table 5.

Table 7 depicts the experimental outcome of these models in both sub-tasks. The result shows that the proposed model, PALCP obtained the best performance in terms of all of the evaluation metrics. It outperformed XLM-RoBERTa with 9.18% in sub-task 1 and 4.76% in sub-task 2, in terms of Pearson correlation. It also achieved 38.36% and 35.53% higher scores than Bangla BERT-20 in sub-task 1 and sub-task 2, respectively.

Model	P	S	MAE	MSE
Sub-task 1: Bengali SW LCP				
PALCP	**0.610**	**0.544**	**0.125**	**0.028**
XLM-RoBERTa	0.554	0.496	0.129	0.033
Multilingual-BERT	0.408	0.403	0.130	0.042
Bangla BERT-20	0.376	0.359	0.180	0.046
Multilingual-DistilBERT	0.436	0.418	0.132	0.044
BanglaBERT-22	0.444	0.420	0.128	0.038
Sub-task 2: Bengali MWE LCP				
PALCP	**0.546**	**0.564**	**0.099**	**0.019**
XLM-RoBERTa	0.520	0.553	0.109	0.020
Multilingual-BERT	0.378	0.408	0.111	0.023
Bangla BERT-20	0.352	0.371	0.115	0.026
Multilingual-DistilBERT	0.390	0.405	0.118	0.030
BanglaBERT-22	0.408	0.412	0.127	0.026

Table 7: Performance of proposed system. The best result is in boldface; P stands for Pearson's and S stands for Spearman's correlation.

To analyze the performance of individual components of our system, we performed an ablation study in Table 8. The PALCP system outperformed the individual models in both sub-tasks. While predicting lexical complexity from single words, the system without LSTM degrades performance by 12.63% in terms of Pearson correlation; whilst, pairwise multi-head attention (PMA) is omitted from the system, the Pearson score is reduced by 12.61%. In the task of predicting the lexical complexity of MWE, the integration of PMA and the LSTM as individual components degraded the performance. However, employing the integration of these components increased the performance by 4.76% in terms of Pearson correlation.

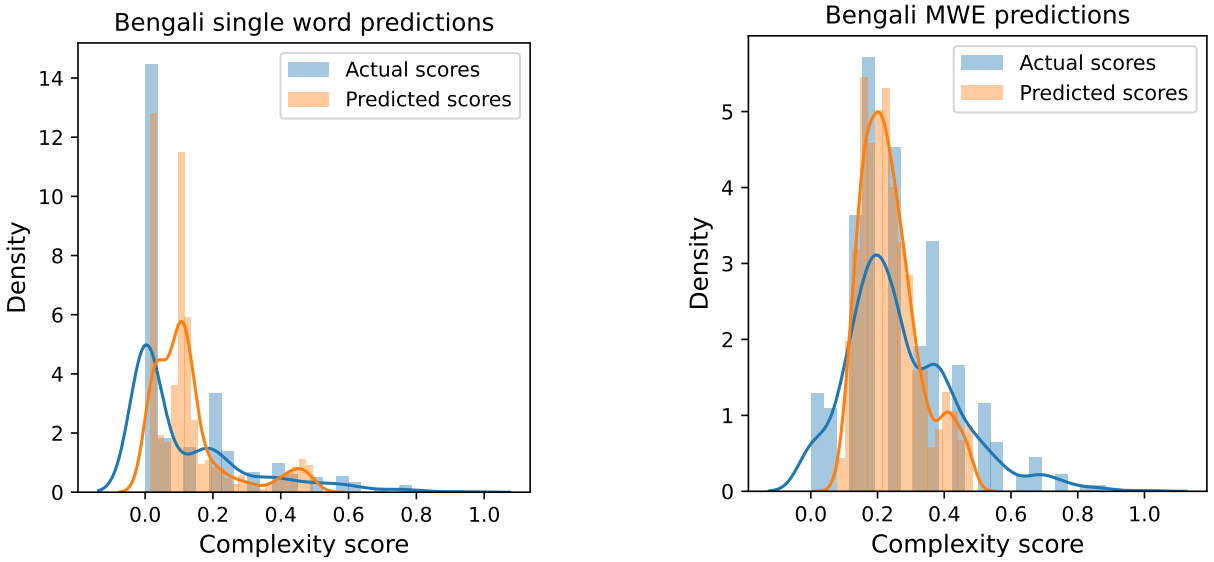
After analyzing the performance of individual models, we can observe that each component has a significant contribution to the system while performing lexical complexity prediction. XLM-RoBERTa is trained on a large-scale multilingual corpus, which includes data from multiple languages, including Bengali. This multilingual training approach allows XLM-RoBERTa to leverage knowledge and representations learned from other languages, potentially benefiting the Bengali language understanding. The LSTM network expands the capacity of the model to learn long-term sequential data, and the pairwise multi-head attention mechanism extracts the relationship between the target token and its context. With the integration of these components, our system achieved comparative performance in the LCP task.

Model	Pearson	Spearman
Sub-task 1: Bengali SW LCP		
PALCP	**0.6096**	**0.5435**
XLM-RoBERTa+PMA	0.5326	0.5057
XLM-RoBERTa+LSTM	0.5327	0.5268
XLM-RoBERTa	0.5535	0.4955
Sub-task 2: Bengali MWE LCP		
PALCP	**0.5460**	**0.5643**
XLM-RoBERTa+PMA	0.5209	0.5260
XLM-RoBERTa+LSTM	0.5005	0.4932
XLM-RoBERTa	0.5203	0.5531

Table 8: Performance analysis of individual components of the system.

6. Discussion

The histogram of the actual complexity score and predicted complexity score by the proposed system for BengaliLCP is depicted in Figure 5. In the case of SW complexity, we observe that the model provided predictions erroneously between 0.1 to 0.2 complexity score range. We can see this same tendency of giving wrong predictions in the MWE LCP. In both sub-tasks, the model has a limitation while predicting comparatively higher complexity scores. We have discussed earlier in Section 3.3 that our BengaliLCP dataset contains more low-complexity data than high-complexity data which caused the lack of performance of the model while predicting scores for highly complex input data. The performance of pre-trained language models heavily relies on the availability and quality of training data. Bengali, being a less-resourced language compared to languages like English, might have a scarcity of high-quality training data. A diminished performance can result from the inability of the model to understand the intricacies and complexities of the language due to limited data.



(a) SW complexity prediction.

(b) MWE complexity prediction.

Figure 5: Comparison between the predictions of the PALCP model and the actual score of the BengaliLCP test set.

Sentence	Score	Pred
Poor predictions by PALCP model		
E1: এতে অ্যালোইস নেগ্রেলির পরিকল্পনা অনুসারে সমুদ্রের পানি অবাধে প্রবাহিত হওয়ার বাধা দেওয়ার জন্য কোন জলকপাট ব্যবস্থা নেই। (It has no waterfall system to prevent the free flow of sea water as planned by Alois Negrelli.)	0.94	0.39
E2: শীর্ষক অভিসন্দর্ভে তিনি লিখেছেন, 'সামাজিক সম্পর্কগুলোর যূথবদ্ধতাই মনুষ্যচরিত্রের সার।' (In the preface, he wrote, 'The cohesion of social relations is the essence of human character'.)	0.75	0.29

Table 9: Unsuccessful test cases analysis; Score stands for actual complexity score and Pred stands for the predicted complexity score.

In Table 9, we listed some test cases where the proposed PALCP model failed to predict accurately. Here, examples of test sentences are given with their target token, actual, and predicted complexity scores. In example E1, the actual complexity of the target token is 0.94 but the prediction score is 0.39. The target token of the next example E2 contains 0.75 as the actual complexity score and 0.29 as the predicted complexity score. These examples contain comparatively more primitive words as their target token. In our proposed system, we utilized XLM-RoBERTa for encoding the texts of the Bengali language. The reason behind this inadequate performance of the Bengali LCP detection can be that these out-of-vocabulary words may not be well represented in the word embeddings of the multilingual language model. As a result, the model may fail to understand or generate accurate predictions for these terminologies. Another issue can be comparatively short ambiguous sentences that couldn't provide much contextual information to the system. To reduce these drawbacks we need to extend the Bengali resources so that NLP models can learn the intricate hidden features of the Bengali language and perform more accurately.

7. Conclusion

Our study highlights the challenges faced by languages with limited resources, such as Bengali, in contributing effectively to the LCP task. Deep learning models often struggle to extract meaningful features from these languages due to the scarcity of annotated data. To address these issues and facilitate research in Bengali LCP, we have taken a step forward by introducing a new annotated dataset in Bengali. Furthermore, we proposed a pairwise neural model that leverages the power of transformer-based language models, pairwise multi-head attention mechanisms, and LSTM to conduct and outperform other state-of-the-art models in the LCP task. Building upon the findings and contributions of this study, several potential directions for future work emerge, that can further advance the field of LCP in the arena of multilingual application.

8. Bibliographical References

Abdul Aziz, MD Akram Hossain, and Abu Nowshed Chy. 2021. Csecu-dsg at semeval-2021 task 1: Fusion of transformer models for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 627–631.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Mubasshir, Md. Saiful Islam, Wasi Ahmad Uddin, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Lagnuage model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022*.

Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. 2021. Simple or complex? learning to predict readability of bengali texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12621–12629.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Sreerupa Das and Rajkumar Roychoudhury. 2006. Readability modelling and comparison of one and two parametric fit: A case study in bangla. *Journal of Quantitative Linguistics*, 13(01):17–34.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nahid Hossain and Adil Ahnaf. 2021. Bert-based text simplification approach to reduce linguistic complexity of bangla language. In *2021 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE)*, pages 1–5. IEEE.

Zahurul Islam, Alexander Mehler, and Rashedur Rahman. 2012. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

Zahurul Islam, Md Rashedur Rahman, and Alexander Mehler. 2014. Readability classification of bangla texts. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 507–518. Springer.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sainik Kumar Mahata, Avishek Garain, Dipankar Das, and Sivaji Bandyopadhyay. 2022. Simplification of english and bengali sentences for improving quality of machine translation. *Neural Processing Letters*, 54(4):3115–3139.

Alejandro Mosquera. 2021. Alejandro mosquera at semeval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559.

Jenny A Ortiz-Zambrano and Arturo Montejo-Ráez. 2021. Complex words identification using word-level features for semeval-2020 task 1. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 126–129.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. Deepblueai at semeval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584.

Shanta Phani, Shibamouli Lahiri, and Arindam Biswas. 2019. Readability analysis of bengali

literary texts. *Journal of Quantitative Linguistics*, 26(4):287–305.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading.

Sagor Sarker. 2021. Bnlp: Natural language processing toolkit for bengali language.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: the complex 2.0 dataset. *Language Resources and Evaluation*, pages 1–42.

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

Regina Stodden and Gayatri Venugopal. 2021. Rs_gv at semeval-2021 task 1: Sense relative lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 640–649.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Annual Meeting of the Association for Computational Linguistics*.

A. Annotated Instances of BengaliLCP

Table 10 shows annotated instances from the BengaliLCP dataset. Each text contains a target token which is underlined in the table and the complexity score of the target token. The table also shows the English translation of the sentences that are obtained using Google Translate. The range of the complexity score is between 0-1, with '0' being the lowest complexity level and '1' being the highest complexity level.

Text	Score
হঠাৎ একজন আদিবাসী বলল, অনেক বছর আগে নদীর কাছাকাছি পর্বতে সে এরকম একটি পাথর দেখেছিল। (Suddenly a native said, many years ago he had seen such a stone in the mountain near the river.)	0.06
আফগানিস্তান ও ইরাকে যুক্তরাষ্ট্রের অন্যায় কার্যক্রমকে প্রকাশ্যে এনে তিনি রাজনৈতিক নেতাদের বিরাগভাজন হয়েছেন। (He has drawn political leaders ire by exposing wrongdoing by the United States in Afghanistan and Iraq.)	0.68
তথ্যটি যে তত্ত্বটি ধরে রাখে তা আসে লোকদের একটি গল্প থেকে, সে আকাশের ওরিয়ন তারামণ্ডলীর অঞ্চলের দেখা-শোনা করতো। (The theory that holds the data comes from a story of people who used to watch the Orion constellation region of the sky.)	0.68
কিন্তু প্রায় ২৫ মিলিয়ন বছর আগে পর্যন্ত, উত্তর কুইন্সল্যান্ড এখনও গ্রীষ্মমন্ডলীয় অঞ্চলের দক্ষিণে নাতিশীতোষ্ণ জলের মধ্যে ছিল - প্রবাল বৃদ্ধি সমর্থন করার জন্য খুব শীতল। But until about 25 million years ago, north Queensland was still in temperate waters south of the tropics - too cold to support coral growth.	0.50
সামন্ততান্ত্রিক কৃষি ব্যবস্থার অর্থনৈতিক ভিত্তি ১৬ শতকের ইংল্যান্ডে উল্লেখযোগ্যভাবে পরিবর্তিত হতে শুরু করে কারণ ম্যানোরিয়াল সিস্টেম ভেঙ্গে গিয়েছিল এবং ক্রমবর্ধমান বৃহৎ এস্টেটসহ কম জমির মালিকদের হাতে জমি কেন্দ্রীভূত হতে শুরু করে। The economic basis of feudal agriculture began to change significantly in 16th-century England as the manorial system broke down and lands began to be concentrated in the hands of fewer landowners with increasingly large estates.	0.92

Table 10: Annotated instances of BengaliLCP dataset. The target token is underlined.