LowResource at BLP-2023 Task 2: Leveraging BanglaBert for Low Resource Sentiment Analysis of Bangla Language

Aunabil Chakma
Bangladesh University of Engineering and Technology
`0419052075@grad.cse.buet.ac.bd`

Masum Hasan
University of Rochester
`m.hasan@rochester.edu`

Abstract

This paper describes the system of the LowResource Team for Task 2 of BLP-2023, which involves conducting sentiment analysis on a dataset composed of public posts and comments from diverse social media platforms. Our primary aim is to utilize BanglaBert, a BERT model pre-trained on a large Bangla corpus, using various strategies including fine-tuning, dropping random tokens, and using several external datasets. Our final model is an ensemble of the three best BanglaBert variations. Our system has achieved overall 3rd in the Test Set among 30 participating teams with a score of 0.718. Additionally, we discuss the promising systems that didn't perform well namely task-adaptive pertaining and paraphrasing using BanglaT5. Training codes and external datasets which are used for our system are publicly available at https://github.com/Aunabil4602/bnlp-workshop-task2-2023

1 Introduction

In the field of Natural Language Processing, Sentiment Analysis has earned significant attention as a research area dedicated to the analysis of textual content. A considerable body of research on Sentiment Analysis in Bangla has been conducted. Some of these works (e.g. Islam et al. (2021), Kabir et al. (2023)) are based on introducing new datasets. In parallel, other works(e.g. Amin et al. (2019), Al-Amin et al. (2017)) are done on novel approaches. In spite of these numerous works, different opportunities still exist to improve the Analysis of Sentiments.

In this paper, we describe our system for task 2 of the Bangla Language Processing Workshop @EMNLP-2023 (Hasan et al., 2023a). We employ various systems based on BanglaBert and BanglaBert-Large (Bhattacharjee et al., 2022). Our experimental systems include fine-tuning, increasing the generalization based on dropping ran-

Rank	Team	Micro-f1
1	MoFa_Aambela	0.731
2	yangst	0.727
3	LowResource(ours)	0.718
4	Hari_vm	0.717
5	PreronaTarannum	0.716

Table 1: Showing top 5 of the final standings of the BLP-2023 Task 2. Our team stands 3rd among 30 participants.

dom tokens, using open-source external data during pre-training, and other methods described in section 4. Utilization of random token drop and external datasets has benefited our systems by improving micro-f1 scores around 0.006 to 0.01. Our best model, an ensembled model from three top models based on the Development Test-Set score, has scored a micro-f1 score of 0.7179, standing overall 3rd among 30 participants. Table 1 shows the final standings of the task.

Additionally, we describe alternate potential methods that have not scored well in the result section 6. To illustrate, we explore Task Adaptive Pre-Training (Gururangan et al., 2020), in fact, has been used by this year's winner of SemEval Task 12 (Muhammad et al., 2023) on sentiment analysis of African Language, and generating paraphrases using BanglaT5 (Bhattacharjee et al., 2023). Moreover, we notice a significant drop in our score in the final test set of our best model. We describe this as our limitations in the section 7.

2 Related Works

Many of the related works are primarily focused on novel datasets covering diverse domains. Islam et al. (2022) have developed a dataset comprised of various public comments from social media platforms. Rahman and Dey (2018) have created their datasets based on Cricket and Restaurant

reviews. Most recently, (Kabir et al., 2023) have published a dataset entirely comprised of book reviews from online bookshops.

Existing approaches to Sentiment Analysis on Bangla Language primarily rely on machine learning and deep learning techniques. For example, Arafin Mahtab et al. (2018) have used Support Vector Machine(SVM) for Sentiment Analysis on public opinions on Cricket. Recurrent Neural Network based models are also highly used. (e.g. Hassan et al. (2016)). Irtiza Tripto and Eunus Ali (2018) have explored a variety of approaches including LSTM, SVM, and Naive Bayes. Moreover, convolutional neural network based models are also used for Sentiment Analysis on Bangla(e.g. Alam et al. (2017)).

In recent years, Large Language Models(LLM), trained on huge corpus, have become popular for their capability to understand the language and can easily fine-tuned for any task like Sentiment Analysis. LLMs based on the Bangla language(e.g. BanglaBert (Bhattacharjee et al., 2022), shahajBert (Diskin et al., 2021), BanglaT5 (Bhattacharjee et al., 2023)) are also available, which opens opportunities to work on various tasks for Bangla.

3 Task Description

This is a multi-class classification task where the objective is to detect the sentiment of the given text into 3 different classes: Positive, Negative, and Neutral. The score will be calculated using the micro-f1. The task consists of two phases: a development phase followed by a test phase. The final standing is based on the score of the test set provided during the test phase.

3.1 Dataset Description

The dataset is comprised of MUBASE (Hasan et al., 2023b) and SentiNob (Islam et al., 2021) datasets. The SentiNob dataset consists of various public comments collected from social media platforms. It covers 13 different domains, for example, politics, education, agriculture, etc. On the other hand, the MUBASE dataset consists of posts collected from Twitter and Facebook. The sample sizes of different sets given for training, validation, and testing are shown in Table 2.

4 System Description

Here, we discuss several systems that we have experimented with for the task including the pre-

Set Name	Sample size
Training	32566
Development	3934
Development Test	3426
Test	6707

Table 2: Sample sizes of various sets provided in the Task 2.

processing of the dataset.

4.1 Fine-tuning Pre-trained LLMs

Fine-tuning Pre-trained Models can achieve high scores with fewer training steps. Top competitors of different shared tasks (e.g. Wang et al. (2022), Wang et al. (2023)) use these pre-trained models. For this task, we use several variations of BanglaBert for fine-tuning. Namely, we use BanglaBert and BanglaBert-Large. Besides, we also use XLM-Roberta-Large (Conneau et al., 2020), a multi-lingual model. We don't explore much on multi-lingual models, since we have found that monolingual models are more used than multi-lingual models on monolingual-specific tasks (Muhammad et al., 2023) due to high scores.

4.2 Task Adaptive Pre-training of LLMs

Gururangan et al. (2020) suggest that Domain Adapting Pre-Training(DAPT) And Task Adaptive Pre-Training(TAPT) improve the scores of the corresponding task. Here, we do TAPT on BanglaBert and BanglaBert-Large using the Electra pre-training method (Clark et al., 2020), which was originally used to pre-train these models. We don't perform DAPT since the models already cover the domains.

4.3 2-Stage Fine-Tuning of LLMs

In the first stage, we fine-tune BanglaBert using the external data only. Here, we don't include any given data from the task. In the next stage, we do regular fine-tuning on the train set. We use the term "2FT" as a short form of this approach. The list of the external datasets and sample sizes are shown in table 10.

4.4 Data augmentation

We experiment with 2 data augmentation techniques to improve the generalization. First, instead of dropping random words (Bayer et al., 2021), we drop random tokens(RTD) since dropping words

might change the meaning. We apply RTD on the fly during the training. Second, we employ paraphrasing as data augmentations using BanglaT5 (Bhattacharjee et al., 2023).

4.5 Preprocessing of Data

We remove the duplicates found in the training set and development set. We replace any url and username with URL and USER tag respectively similar to Nguyen et al. (2020). While using BanglaBert we normalize the sentence by their specific normalizer[1] as required by their model. All of the sentences are tokenized by the individual tokenizer required by each model. We set the max length of tokenization to 128 for each text.

We use several external data. However, most of the labels don't match the labels of this task. For the initial fine-tuning of the LLMs, we first map different labels to the three labels for this task. The label mapping is shown in table 11. For TAPT, we didn't need any of these labels since we do masked language modeling. Finally, we also remove the duplicates found in the external datasets.

5 Experimental Setup

We have used Models and Trainer from Huggingface[2](PyTorch version). We employ mixed precision training (Micikevicius et al., 2017) that enables faster training and consumes low GPU memory. Moreover, we built a code such that the results are reproducible. All of the experiments are done using a single V100 GPU in Google Colaboratory[3]. We do hyper-parameters search on learning rate, batch size, dropout ratios, and total epochs. We start the search with the parameter settings as suggested Gururangan et al. (2020). Our best training parameters of fine-tuning and TAPT are shown in the table 8 and 9 respectively. Note that, we don't use samples from the development set, development-test set, and test set for fine-tuning and pre-training.

6 Results

To begin, we discuss the systems that have scored well on the Development-Test's score. The top individual model is BanglaBert-Large with a random token drop that has scored 0.733, and even without any enhancement, it can score 0.723. The

next best single model is BanglaBert with random token drop(RTD) and 2-stage fine-tuning that has scored 0.729. Table 3 shows the scores of our selected models in the Development-Test Set. Here, we see that both usages of external datasets and RTD have benefited the BanglaBert and BanglaBert-Large. We have built an ensemble of 3 best individual models(model ID 3, 5, and 6) that has scored 0.734, where we decide the class based on majority voting, and in case of a tie, we use the class predicted by the best model. We chose only the 3 best models for the ensemble because the other model's score was low and taking an odd number of models helps to decide the output class in case of a tie.

We have submitted the ensembled model as our best model in the test phase and has scored 0.718. Moreover, We have submitted the 3 individual best models. Our scores on the Test Set are shown in table 4. Here, we have found some inconsistency: BanglaBert-Large with random token drop, which we have considered the best model based on the Development-Test set, performed worst among the other 2 models, and BanglaBert with random token drop and pre-fine-tuned with external data, our 2nd best model, has performed the best. More importantly, every variant of BanglaBert-Large has scored low on the Test set. We discuss some analysis more in section 7. Finally, table 6 shows the confusion matrix of our ensembled model on Test set. We see that our model performed worst on detecting the Neutral class, i.e. only 412 out of 1277 samples have been correct having an accuracy of 32%, where the accuracy of Positive and Neutral classes are 78% and 83% respectively.

There are some systems that didn't achieve favorable performance from the beginning of our experiments. Firstly, TAPT didn't improve our results but rather declined the score by 0.039 with respect to simple fine-tuning as shown in table 5. What we can infer is that TAPT is supposed to help adapt the BanglaBert to the task domain, but it overfitted on the training samples, where the original model is already in a good optima that covered the task domain better.

Paraphrasing to create additional data using BanglaT5 also didn't work well. Its score is shown in table 5. The most perceptible reason is that paraphrased sentences, although good, were not diverse enough from the original sentences. Examples of generated paraphrases are shown in figure

---

[1] https://github.com/csebuetnlp/normalizer

[2] https://huggingface.co/

[3] https://colab.research.google.com/

1.

Other than BanglaBert, we try the XLM-Roberta-Large, a multi-lingual model, which is used by several task winners (e.g. (Wang et al., 2022)). However, it has scored low on the Development-Test set even with all enhancements. Its score is also shown in Table 3.

ID	System	Micro-F1
1	BBert	0.718
2	BBert+RTD	0.722
3	BBert+RTD+2FT	0.729
4	BBertL	0.723
5	BBertL+RTD	0.733
6	BBertL+RTD+2FT	0.725
7	XLM-Roberta-Large+RTD	0.713
8	Ensemble(3+5+6)	0.734

Table 3: Scores of our best models of various systems in the Development-Test Set. Here, BBert means BanglaBert, RTD means random token drop and 2FT means 2-stage fine-tuning.

System	Micro-F1
BBertL+RTD	0.711
BBert+RTD+2FT	0.719
BBertL+RTD+2FT	0.713
Ensemble(above 3 models)	0.718

Table 4: Scores of 3 of our best models(based on Development-Test Set) and ensembled model in Test Set. Here, BBert means BanglaBert, RTD means random token drop and 2FT means 2-stage fine-tuning.

7 Limitations and future work

As mentioned earlier, we find inconsistency in the score of our best model (BanglaBert-Large) between the Development-Test Set and the Test Set. Clark et al. (2020) have stated that variance in performance is observed with different seeds when the size of the dataset is small. We assume this might be the cause, although we didn't rely on

System	Micro-F1
Fine-Tuning	0.727
TAPT	0.688
Paraphrasing	0.674

Table 5: Performance of TAPT and Paraphrasing on BanglaBert-Large in comparison with fine-tuning on Development Set.

	Predicted		
	Neg	Neut	Pos
Neg	2770	244	324
Neut	598	412	267
Pos	331	128	1633

(True)

Table 6: Confusion Matrix of the Ensembled model on Test Set.

Seed	BBert	BBertL
1234	0.7156	0.7115
42	0.7179	0.7110
747	0.7197	0.7210
52467	0.7192	0.7122
2779	0.7135	0.7161
362	0.7185	0.7134
8194	0.7182	0.7127
avg.	0.7177	0.7140

Table 7: Scores from using different seeds for BanglaBert(BBert), BanglaBert-Large(BBertL) on Test Set.

other seeds since the distribution of Development-Test Set and Test Set should be similar as they come from the same datasets. To be more certain, we ran an experiment using different seeds for both BanglaBert and BanglaBert-Large on the Test set. As anticipated, models show varying performance when initialized with different seeds. Table 7 shows the results of this experiment. Moreover, we have found that the average score of the BanglaBert is better than the BanglaBert-Large. In fact, this result is consistent with the result found by the authors of BanglaBert that BanglaBert-Large performs lower than BanglaBert on Sentiment Analysis on SentiNob dataset[4]. BangalThus, before considering a model, the average score from different seeds needs to be evaluated when the training data is small.

TAPT is a popular method for pre-training, but it has been ineffective for our task. However, we have inferred this based on a few experiments. Thus, we suggest that more research needs to be done on the effectiveness of TAPT, as well as DAPT, on BanglaBert.

Our research has been mostly based on fine-tuning. As future work, we would like to explore using common data augmentation techniques (Bayer et al., 2021) for the given data. Besides, there are several multilingual Pre-trained

---

[4]https://huggingface.co/csebuetnlp/banglabert_large

Models that include the Bangla Language are need to be explored along with sophisticated methods and may even achieve better results.

8 Conclusion

In this paper, we stated our systems based on BanglaBert and BanglaBert-Large for that Sentiment Analysis task. We used simple techniques like, 2-stage fine-tuning, using external datasets, and dropping random tokens. Our system scored 3rd overall in the task. We also discussed some potential systems that didn't demonstrate satisfactory performance. More importantly, we have discussed the score inconsistency of our best model between Development-Test Set and Test Set as our limitation. Finally, we discussed directing some future research like applying TAPT and DAPT on BanglaBert and trying more data augmentations or sophisticated methods.

References

Md. Al-Amin, Md. Saiful Islam, and Shapan Das Uzzal. 2017. Sentiment analysis of bengali comments with word2vec and sentiment information of words. In *International Conference on Electrical, Computer and Communication Engineering (ECCE)*.

Md. Habibul Alam, Md-Mizanur Rahoman, and Md. Abul Kalam Azad. 2017. Sentiment analysis for bangla sentences using convolutional neural network. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6.

Al Amin, Imran Hossain, Aysha Akther, and Kazi Masudul Alam. 2019. Bengali vader: A sentiment analysis approach using modified vader. In *International Conference on Electrical, Computer and Communication Engineering (ECCE)*.

Shamsul Arafin Mahtab, Nazmul Islam, and Md Mahfuzur Rahaman. 2018. Sentiment analysis on bangladesh cricket with support vector machine. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *CoRR*, abs/2107.03158.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021. Emotion classification in a resource constrained language using transformer-based approach. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop:150–158.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitry Popov, Dmitriy Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas Wolf, and Gennady Pekhimenko. 2021. Distributed deep learning in open collaborations. In *Advances in Neural Information Processing Systems*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023a. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023b. Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis.

Asif Hassan, Mohammad Rashedul Amin, Abul Kalam Al Azad, and Nabeel Mohammed. 2016.

Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In *2016 International Workshop on Computational Intelligence (IWCI)*, pages 51–56.

MD Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2022. Bemoc: A corpus for identifying emotion in bengali texts. *SN Computer Science*, 3(2):1–17.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6.

Khondoker Ittehadul Islam, Md. Saiful Islam, and Md Ruhul Amin. 2020. Emonoba: A dataset for analyzing fine-grained emotions on noisy bangla texts. 23rd International Conference on Computer and Information Technology (ICCIT).

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Khondoker Ittehadul Islam, Tanvir Yuvraz, Md Saiful Islam, and Enamul Hassan. 2022. Emonoba: A dataset for analyzing fine-grained emotions on noisy bangla texts. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers):128–134.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. Findings of the Association for Computational Linguistics: ACL 2023:1237–1247.

Mahfuz Ahmed Masum, Sheikh Junayed Ahmed, Ayesha Tasnim, and Md Saiful Islam. 2020. An aspect-based sentiment analysis dataset for bengali and its baseline evaluation. In Proceedings of International Joint Conference on Advances in Computational Intelligence (IJCACI).

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2017. Mixed precision training. *CoRR*, abs/1710.03740.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Md Atikur Rahman and Emon Kumar Dey. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource bengali language. Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020):50–60.

Salim Sazzed. 2021. Abusive content detection in transliterated bengali-english social media corpus. Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching:125–130.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. NLNDE at SemEval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 488–497, Toronto, Canada. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

A   Appendix

Here, we show a figure and additional tables related to our descriptions.

Parameter	BBert(+L)	XLMR
Learning Rate(LR)	2e-5	-
LR Scheduler	Linear	-
Warmup Ratio	0.0	-
Train Batch Size	16	32
Train Epochs	3	5
Weight Decay	0.01	-
Token Drop Ratio	0.2	-
Classifier Dropout	0.1	-
Max Length	128	-

Table 8: Best hyper-parameter settings for fine-tuning. BBert(+L) means both the BanglaBert and BanglaBert-Large models, XLMR means the XLM-Roberta-Large model, and "-" means equal to the left column values.

Original	Converted
Love	Positive
Joy	Positive
Anger	Negative
Sad/Sadness	Negative
Fear	Negative
Disgust	Negative
Surprise	Neutral
Abusive	Negative
Non-Abusive	Positive

Table 11: Label conversions of external datasets for aligning to our task.

Parameter	BBertL
λ	50
MLM probability	0.25
Learning Rate(LR)	1e-4
LR Scheduler	Linear
Warmup Ratio	0.06
Train Batch Size	64
Train Epochs	100
Weight Decay	0.01
Token Drop Ratio	0.2
Max Length	128

Table 9: Best hyper-parameter settings for Electra Pre-Training of BanglaBert and BanglaBert-Large during Task Adaptive Pre-training(TAPT). λ is the weight of the loss for Discriminator

Example-1
Original: কফি এক্সপ্রেস , মিরপুর 11 নং বাস স্ট্যান্ড থেকে পশ্চিমে । ইয়ান তাই রেস্টুরেন্ট এর পেছনে
Paraphrased: কফি এক্সপ্রেস মিরপুর ১১ বাস স্ট্যান্ড থেকে পশ্চিমে, ইয়ান রেস্তোরাঁর পিছনে।

Example-2
Original: শুধু বেশি নম্বর পেয়ে কী হবে ? কেন এ প্রশ্ন শিক্ষামন্ত্রীর ?
Paraphased: শুধু বেশি নম্বর পেলে কি হবে? শিক্ষামন্ত্রীর প্রশ্ন কেন?

Figure 1: Showing 2 paraphrasing examples using BanglaT5.

Dataset	Samples	Total class
Islam et al. (2022)	22739	6
Sazzed (2021)	1000	2
Iqbal et al. (2022)	7000	6
Kabir et al. (2023)	158065	3
Sazzed (2020)	11807	2
Rahman and Dey (2018)	2979	3
Das et al. (2021)	4994	6
Masum et al. (2020)	9014	3
Islam et al. (2020)	14852	3

Table 10: The list of External Datasets used for our training.