BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla

Anonymous ACL submission

Abstract

In this paper, we introduce 'BanglaBERT', a BERT-based Natural Language Understanding (NLU) model pretrained in Bangla, a widely spoken yet low-resource language in the NLP literature. To pretrain BanglaBERT, we collect 27.5 GB of Bangla pretraining data (dubbed 'Bangla2B+') by crawling 110 popular Bangla sites. We introduce a new downstream task dataset on Natural Language Inference (NLI) and benchmark on four diverse NLU tasks covering text classification, sequence labeling, and span prediction. In the process, we bring them under the first-ever Bangla Language Understanding Evaluation (BLUE) benchmark. BanglaBERT achieves state-of-the-art results outperforming multilingual and monolingual models. We will make the BanglaBERT model, the new datasets, and a leaderboard publicly available to advance Bangla NLP.

1 Introduction

Despite being the sixth most spoken language in the world with over 300 million native speakers constituting 4% of the world's total population,[1] Bangla is considered a resource-scarce language. Joshi et al. (2020b) categorized Bangla in the language group that lacks efforts in labeled data collection and relies on self-supervised pretraining (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019) to boost the natural language understanding (NLU) task performances. To date, the Bangla language has been continuing to rely on fine-tuning multilingual pretrained language models (PLMs) (Ashrafi et al., 2020; Das et al., 2021; Islam et al., 2021). However, since multilingual PLMs cover a wide range of languages (Conneau and Lample, 2019; Conneau et al., 2020), they are large (have hundreds of millions of parameters) and require substantial computational resources for fine-tuning. They also tend to show degraded performance for low-resource languages (Wu and Dredze, 2020) on downstream NLU tasks. Motivated by the triumph of language-specific models (Martin et al. (2020); Polignano et al. (2019); Canete et al. (2020); Antoun et al. (2020), *inter alia*) over multilingual models in many other languages, in this work, we present **BanglaBERT** – a BERT-based (Devlin et al., 2019) Bangla NLU model pretrained on 27.5 GB data (which we name '**Bangla2B+**') we meticulously crawled 110 popular Bangla websites to facilitate NLU applications in Bangla.

We also introduce a Bangla Natural Language Inference (NLI) dataset, a task previously unexplored in Bangla, and evaluate BanglaBERT on four diverse downstream tasks on sentiment classification, NLI, named entity recognition, and question answering. We bring these tasks together to establish the first-ever **B**angla **L**anguage **U**nderstanding **E**valuation (**BLUE**) benchmark. We compare two widely used multilingual models to BanglaBERT using the BLUE benchmark and find that BanglaBERT excels on all the tasks.

We summarize our contributions as follows:

1. We present BanglaBERT, a pretrained BERT model for Bangla, and introduce a new Bangla natural language inference (NLI) dataset.
2. We introduce Bangla Language Understanding Evaluation (BLUE) benchmark and provide a set of strong baselines.
3. We release code and provide a leaderboard to spur future research on Bangla NLU.[2]

2 BanglaBERT

2.1 Pretraining Data

A high volume of good quality text data is a prerequisite for pretraining large language models. For instance, BERT (Devlin et al., 2019) is pretrained on the English Wikipedia and the Books corpus (Zhu et al., 2015) containing 3.3 billion tokens. Subsequent works like RoBERTa (Liu et al., 2019)

[1] https://w.wiki/Psq

[2] https://github.com/hidden/hidden

and XLNet (Yang et al., 2019) used more extensive web-crawled data with heavy filtering and cleaning.

Bangla is a rather resource-constrained language in the web domain; for example, the Bangla Wikipedia dump from July 2021 is only 650 MB, two orders of magnitudes smaller than the English Wikipedia. As a result, we had to crawl the web extensively to collect our pretraining data. We selected 110 Bangla websites by their Amazon Alexa rankings[3] and the volume and quality of extractable texts by inspecting each website. The contents included encyclopedias, news, blogs, e-books, stories, social media/forums, etc.[4] The amount of data totaled around 35 GB.

There are also noisy sources of Bangla data dumps, a couple of prominent ones being OSCAR (Suárez et al., 2019) and CCNet (Wenzek et al., 2020). However, they contained lots of offensive texts; we found them infeasible to clean thoroughly. Fearing potential harmful impacts (Luccioni and Viviano, 2021), we opted not to use them.[5]

2.2 Pre-processing

We performed thorough deduplication on the pretraining data, removed non-textual contents (e.g., HTML/JavaScript tags), and filtered out non-Bangla pages using a language classifier (Joulin et al., 2017). After the processing, the dataset was reduced to 27.5 GB in size containing 5.25M documents having 306.66 words on average.

We trained a Wordpiece (Wu et al., 2016) vocabulary of 32k subword tokens on the resulting corpus with a 400 character alphabet, kept larger than the native Bangla alphabet to capture code-switching (Poplack, 1980) and allow romanized Bangla contents for better generalization. We limited the length of a training sample to 512 tokens and did not cross document boundaries (Liu et al., 2019) while creating a data point. After tokenization, we had 7.18M samples with an average length of 304.14 tokens and containing 2.18B tokens in total; hence we named the dataset '*Bangla2B+*'.

2.3 Pretraining Objective

Self-supervised pretraining objectives leverage unlabeled data. For example, BERT (Devlin et al., 2019) was pretrained with masked language modeling (MLM) and next sentence prediction (NSP). Several works built on top of this, e.g., RoBERTa (Liu et al., 2019) removed NSP and pretrained with longer sequences, SpanBERT (Joshi et al., 2020a) masked contiguous spans of tokens, while works like XLNet (Yang et al., 2019) introduced objectives like factorized language modeling.

We pretrained BanglaBERT using ELECTRA (Clark et al., 2020b), pretrained with the Replaced Token Detection (RTD) objective, where a generator and a discriminator model are trained jointly. The generator is fed as input a sequence with a portion of the tokens masked (15% in our case) and is asked to predict them using the rest of the input (i.e., standard MLM). The masked tokens are then replaced by tokens sampled from the generator's output distribution for the corresponding masks, and the discriminator then has to predict whether each token is from the original sequence or not. After pretraining, the discriminator is used for fine-tuning. Clark et al. (2020b) argued that RTD back-propagates loss from all tokens of a sequence, in contrast to 15% tokens of the MLM objective, giving the model more signals to learn from. Evidently, ELECTRA achieves comparable downstream performance to RoBERTa or XLNet with only a quarter of their training time. This computational efficiency motivated us to use ELECTRA for our implementation of BanglaBERT.

2.4 Model Architecture & Hyperparameters

We pretrained the base ELECTRA model (a 12-layer Transformer encoder with 768 embedding size, 768 hidden size, 12 attention heads, 3072 feed-forward size, generator-to-discriminator ratio $\frac{1}{3}$, 110M parameters) with 256 batch size for 2.5M steps on a v3-8 TPU instance on GCP. We used the Adam (Kingma and Ba, 2015) optimizer with a 2e-4 learning rate and linear warmup of 10k steps.

3 The Bangla Language Understanding Evaluation (BLUE) Benchmark

Many works have studied different Bangla NLU tasks in isolation, e.g., sentiment classification (Das and Bandyopadhyay, 2010; Sharfuddin et al., 2018; Tripto and Ali, 2018), semantic textual similarity (Shajalal and Aono, 2018), parts-of-speech (PoS) tagging (Alam et al., 2016), named entity recognition (NER) (Ashrafi et al., 2020). However, Bangla NLU has not yet had a comprehensive unified study. Motivated by the surge of NLU research brought

[3] www.alexa.com/topsites/countries/BD

[4] The completely list of the sources can be found in the Appendix.

[5] We cover other ethical considerations in the Appendix.

Task	Corpus	|Train|	|Dev|	|Test|	Metric	Domain
Sentiment Classification	SentNoB	12,575	1,567	1,567	Macro-F1	Social Media
Natural Language Inference	BNLI	381,449	2,419	4,895	Acc.	Misc.
Named Entity Recognition	MultiCoNER	14,500	800	800	Micro-F1	Misc.
Question Answering	TyDiQA	4,542	238	226	EM/F1	Wikipedia

Table 1: Statistics of the Bangla Language Understanding Evaluation (BLUE) benchmark.

about by benchmarks in other languages, e.g., English (Wang et al., 2018), French (Le et al., 2020), Korean (Park et al., 2021), we establish the first-ever Bangla Language Understanding Evaluation (BLUE) benchmark.

NLU generally comprises three types of tasks: text classification, sequence labeling, and text span prediction. Text classification tasks can further be sub-divided into single-sequence and sequence-pair classification. Therefore, we consider a total of four tasks for BLUE. For each task type, we carefully select one downstream task dataset. We emphasize the quality and open availability of the datasets while making the selection. We briefly mention them below:

1. **Single-Sequence Classification**: Sentiment classification is perhaps the most-studied Bangla NLU task, with some of the earlier works dating back over a decade (Das and Bandyopadhyay, 2010). Hence, we chose this as the single-sequence classification task. However, most Bangla sentiment classification datasets are not publicly available. We could only find two public datasets: *BYSA* (Tripto and Ali, 2018) and *SentNoB* (Islam et al., 2021). We found BYSA to have many duplications. Even worse, many duplicates had different labels. SentNoB had better quality and covered a broader set of domains, making the classification task more challenging. Hence, we opted to use the latter.

2. **Sequence-pair Classification**: In contrast to single-sequence classification, there has been a dearth of sequence-pair classification works in Bangla. We found work on semantic textual similarity (Shajalal and Aono, 2018), but the dataset is not publicly available. As such, we curated a new Bangla Natural Language Inference (*BNLI*) dataset for sequence-pair classification. We chose NLI as the representative task due to its fundamental importance in NLU. Given two sentences, a premise and a hypothesis as input, a model is tasked to predict whether or not the hypothesis is entailment, contradiction, or neutral to the premise. We used the same curation procedure as the XNLI (Conneau et al., 2018) dataset: we translated the MultiNLI (Williams et al., 2018) training data using the English to Bangla translation model by Hasan et al. (2020) and had the evaluation sets translated by expert human translators.[6] Due to the possibility of the incursion of errors during automatic translation, we used the Language-Agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2020) of the translations and original sentences to compute their similarity and discarded all sentences below a similarity threshold of 0.70. Moreover, to ensure good-quality translation, we used similar quality assurance strategies as Guzmán et al. (2019).

3. **Sequence Labeling**: In this task, all words of a text sequence have to be classified. Named Entity Recognition (NER) and Parts-of-Speech (PoS) tagging are two of the most prominent sequence labeling tasks. We chose the Bangla portion of SemEval 2022 *MultiCoNER*[7] dataset for BLUB.

4. **Span Prediction**: Extractive question answering is a standard choice for text span prediction. We used the Bangla portion of the *TyDiQA*[8] (Clark et al., 2020a) dataset for this task. We posed the task analogous to SQuAD 2.0 (Rajpurkar et al., 2018): presented with a text passage and a question, a model has to predict whether or not it is answerable. If answerable, the model has to find the minimal text span that answers the question.

We present detailed statistics of the BLUE benchmark in Table 1.

4 Experiments & Results

We fine-tuned BanglaBERT on the downstream tasks and compared with three multilingual models: mBERT (Devlin et al., 2019), XLM-R base and large (Conneau et al., 2020), and sahajBERT (Diskin et al., 2021) (an ALBERT-based (Lan et al.,

[6] We present more details in the Appendix.

[7] The test set of MultiCoNER will be released in December. We used the dev set for test and a portion of the training set for dev. We will redo all evaluations with the released test set.

[8] The test set of TyDiQA is not publicly available. Like MultiCoNER, we used the validation set for test purposes and a portion of the training set for validation. We removed the Yes/No questions and subsampled the unanswerable questions to have the same frequency as the answerable ones.

Models	|Params.|	SC	NLI	NER	QA	BLUE Score
Zero-shot cross-lingual transfer learning						
mBERT	180M	–	62.22	39.27	59.44/65.00	–
XLM-R (base)	270M	–	72.18	45.37	57.66/63.47	–
XLM-R (large)	550M	–	78.16	57.74	70.35/77.22	–
Supervised fine-tuning						
mBERT	180M	67.59	75.13	68.97	75.51/80.12	73.46
XLM-R (base)	270M	69.54	78.46	73.32	75.81/79.98	75.42
XLM-R (large)	550M	70.97	**82.40**	**78.39**	**83.92/87.87**	80.71
sahajBERT	18M	71.12	76.92	70.94	76.40/81.44	75.36
BanglaBERT	110M	**72.89**	**82.80**	**77.78**	**82.89/87.60**	**80.79**

Table 2: Performance comparison of baselines and pretrained models on different downstream tasks. Scores in bold texts have statistically significant ($p < 0.05$) difference from others with bootstrap sampling (Koehn, 2004).

2020) pretrained Bangla model). We also show the zero-shot cross-lingual transfer results fine-tuned on the English counterpart of each dataset (except for SentNoB, which had no English equivalent).

All pretrained models were fine-tuned for 3-20 epochs with batch size 32, and the learning rate was tuned from {2e-5, 3e-5, 4e-5, 5e-5}. We performed fine-tuning with three random seeds and reported their average in Table 2. In all the tasks, BanglaBERT outperformed multilingual models and monolingual sahajBERT, achieving a BLUE score (the average score of all tasks) of 80.79, even coming head-to-head with XLM-R (large).

BanglaBERT is not only superior in performance, but it is also substantially compute- and memory-efficient. For instance, it may seem that sahajBERT is more efficient than BanglaBERT due to its smaller size, but it takes 2-3.5x time and 2.4-3.33x memory as BanglaBERT to fine-tune.[9]

Sample efficiency It is often challenging to annotate training samples in real-world scenarios, especially for low-resource languages like Bangla. So, in addition to compute- and memory-efficiency, sample-efficiency (Howard and Ruder, 2018) is another necessity of PLMs. To assess the sample efficiency of BanglaBERT, we limit the number of training samples and see how it fares against other models. We compare it with XLM-R (large) and plot their performances on the SC and NLI tasks[10] for different sample size in Figure 1.

Results show that when we have fewer number of samples ($\leq 1k$), BanglaBERT has substantially better performance (2-9% on SC, 6-10% on NLI) over XLM-R (large), making it more practically applicable for resource-scarce downstream tasks.

[9]Detailed comparison can be found in the Appendix.

[10]Results for the other tasks can be found in the Appendix.

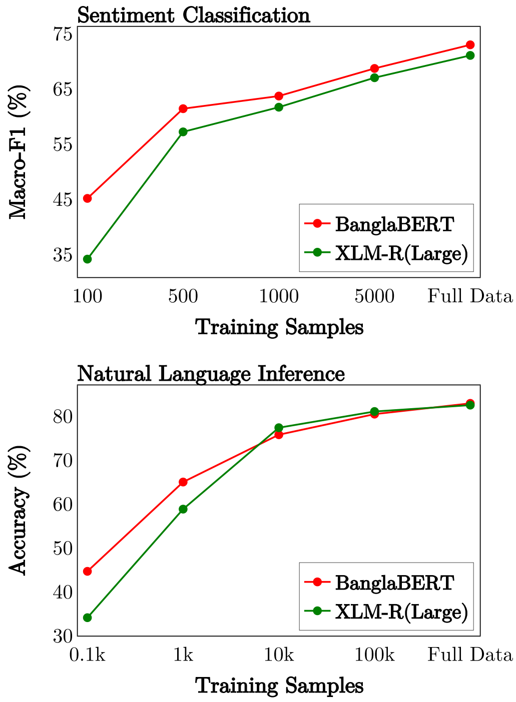

Figure 1: Sample-efficiency tests with SC and NLI.

5 Conclusion & Future Works

Creating language-specific models is often infeasible for low-resource languages lacking ample data. Hence, researchers are compelled to use multilingual models for languages that do not have strong pretrained models. To this end, we introduced BanglaBERT, an NLU model in Bangla, a widely spoken yet low-resource language. We presented a new downstream dataset on NLI and established the BLUE Benchmark, setting new state-of-the-art results with BanglaBERT. In future, we will include other Bangla NLU benchmarks (e.g., dependency parsing (de Marneffe et al., 2021)) in BLUE and investigate the benefits of initializing Bangla NLG models from BanglaBERT.

References

Firoj Alam, Shammur Absar Chowdhury, and Sheak Rashed Haider Noori. 2016. Bidirectional lstms—crfs networks for bangla pos tagging. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 377–382. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

I. Ashrafi, M. Mohammad, A. S. Mauree, G. M. A. Nijhum, R. Karim, N. Mohammed, and S. Momen. 2020. Banner: A cost-sensitive contextualized model for bangla named entity recognition. *IEEE Access*, 8:58206–58226.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Proceedings of the Practical ML for Developing Countries Workshop at ICLR 2020, PML4DC*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020b. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Amitava Das and Sivaji Bandyopadhyay. 2010. Phrase-level polarity identification for bangla. *Int. J. Comput. Linguist. Appl.(IJCLA)*, 1(1-2):169–182.

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

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitry Popov, Dmitry 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

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.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020a. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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 *International Conference on Learning Representations*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

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. *arXiv preprint arXiv:1907.11692*.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, CEUR Workshop Proceedings.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Md Shajalal and Masaki Aono. 2018. Semantic textual similarity in bengali text. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE.

Abdullah Aziz Sharfuddin, Md Nafis Tihami, and Md Saiful Islam. 2018. A deep recurrent neural network with bilstm model for sentiment classification. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.

Appendix

Pretraining Data Sources

We used the following sites for data collection. We categorize the sites into six types:

Encyclopedia:

- bn.banglapedia.org
- bn.wikipedia.org
- songramernotebook.com

News:

- anandabazar.com
- arthoniteerkagoj.com
- bangla.24livenewspaper.com
- bangla.bdnews24.com
- bangla.dhakatribune.com
- bangla.hindustantimes.com
- bangladesherkhela.com
- banglanews24.com
- banglatribune.com
- bbc.com
- bd-journal.com
- bd-pratidin.com
- bd24live.com
- bengali.indianexpress.com
- bigganprojukti.com
- bonikbarta.net
- chakarianews.com
- channelionline.com
- ctgtimes.com
- ctn24.com
- daily-bangladesh.com
- dailyagnishikha.com
- dainikazadi.net
- dainikdinkal.net
- dailyfulki.com
- dailyinqilab.com
- dailynayadiganta.com
- dailysangram.com
- dailysylhet.com
- dainikamadershomoy.com
- dainikshiksha.com
- dhakardak-bd.com
- dmpnews.org
- dw.com
- eisamay.indiatimes.com
- ittefaq.com.bd
- jagonews24.com
- jugantor.com
- kalerkantho.com
- manobkantha.com.bd
- mzamin.com
- ntvbd.com
- onnodristy.com
- pavilion.com.bd
- prothomalo.com
- protidinersangbad.com
- risingbd.com
- rtvonline.com
- samakal.com
- sangbadpratidin.in
- somoyerkonthosor.com
- somoynews.tv
- tbsnews.net
- teknafnews.com
- thedailystar.net
- voabangla.com
- zeenews.india.com
- zoombangla.com

Blogs:

- amrabondhu.com
- banglablog.in
- bigganblog.org
- biggani.org
- bigyan.org.in
- bishorgo.com
- cadetcollegeblog.com
- choturmatrik.com
- horoppa.wordpress.com
- muktangon.blog
- roar.media/bangla
- sachalayatan.com
- shodalap.org
- shopnobaz.net
- somewhereinblog.net
- subeen.com
- tunerpage.com
- tutobd.com

E-books/Stories:

- banglaepub.github.io
- bengali.pratilipi.com
- bn.wikisource.org
- ebanglalibrary.com
- eboipotro.github.io
- golpokobita.com
- kaliokalam.com
- shirisherdalpala.net
- tagoreweb.in

Social Media/Forums:

- banglacricket.com
- bn.globalvoices.org
- helpfulhub.com
- nirbik.com
- pchelplinebd.com
- techtunes.io

Miscellaneous:

- banglasonglyric.com
- bdlaws.minlaw.gov.bd
- bdup24.com
- bengalisongslyrics.com
- dakghar.org
- gdn8.com
- gunijan.org.bd
- hrw.org
- jakir.me
- jhankarmahbub.com
- jw.org
- lyricsbangla.com
- neonaloy.com
- porjotonlipi.com
- sasthabangla.com
- tanzil.net

We wrote custom crawlers for each site above (except the Wikipedia dumps).

Quality Control in Human Translation:

Translations were done by expert translators who provide translation services for renowned Bangla newspapers. Each translated sentence was further assessed for quality by another expert. If found to be of low quality, it was again translated by the original translator. The sample was then discarded altogether if found to be of low quality again. Fewer than 100 samples were discarded in this process.

Compute and Memory Efficiency Tests

To validate that BanglaBERT is more efficient in terms of memory and compute, we measured each model's training time and memory usage during the fine-tuning of each task. All tests were done on a desktop machine with an 8-core Intel Core-i7 11700k CPU and NVIDIA RTX 3090 GPU. We used the same batch size, gradient accumulation steps, and sequence length for all models and tasks for a fair comparison. We use relative time and memory (GPU VRAM) usage considering those of BanglaBERT as units. The results are shown in Table 3. (We mention the upper and lower values of the different tasks for each model)

Additional Sample Efficiency Tests

Due to space restrictions, we move the sample efficiency results of the NER and QA tasks to the appendix. We plot the results in Figure 2.

Similar results are also observed here for the NER task, where BanglaBERT is more sample-efficient when we have $\leq 1k$ training samples. In the QA task however, both models have identical performance for all sample counts.

Model	Time	Memory Usage
mBERT	1.14x-1.92x	1.12x-2.04x
XLM-R (base)	1.29-1.81x	1.04-1.63x
XLM-R (large)	3.81-4.49x	4.44-5.55x
SahajBERT	2.40-3.33x	2.07-3.54x
BanglaBERT	**1.00x**	**1.00x**

Table 3: Compute and memory efficiency tests

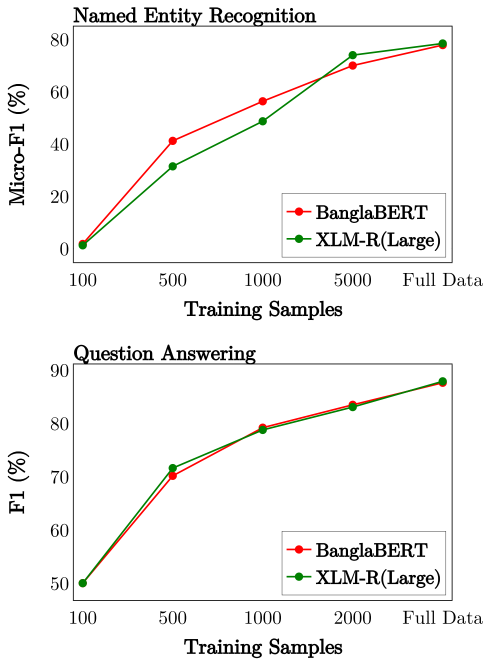

Figure 2: Sample-efficiency tests with NER and QA.

Ethical Considerations

Dataset and Model Release: The **Copy Right Act, 2000**[11] of Bangladesh allows reproduction and public release of copy-right materials for non-commercial research purposes. As a transformative research work, we will release BanglaBERT under a non-commercial license. Furthermore, we will release only the pretraining data for which we know the distribution will not cause any copyright infringement. The downstream task datasets can all be made publicly available under a similar non-commercial license.
Human Translation: Human translators were paid as per standard rates in local currencies.
Text Content: We tried to minimize offensive texts by explicitly crawling the sites where such contents would be nominal. However, we can guarantee absolutely no objectionable content and recommend using the model carefully, especially for text gen-

[11] http://bdlaws.minlaw.gov.bd/act-details-846.html

eration purposes. Furthermore, we removed the personal information of the content writers by not considering the author fields while collecting the data.