BanglaEmbed: Efficient Sentence Embedding Models for a Low-Resource Language Using Cross-Lingual Distillation Techniques

Muhammad Rafsan Kabir
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
muhammad.kabir@northsouth.edu

Md. Mohibur Rahman Nabil
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mohibur.nabil@northsouth.edu

Mohammad Ashrafuzzaman Khan
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mohammad.khan02@northsouth.edu

Abstract—Sentence-level embedding is essential for various tasks that require understanding natural language. Many studies have explored such embeddings for high-resource languages like English. However, low-resource languages like Bengali (a language spoken by almost two hundred and thirty million people) are still under-explored. This work introduces two lightweight sentence transformers for the Bangla language, leveraging a novel cross-lingual knowledge distillation approach. This method distills knowledge from a pre-trained, high-performing English sentence transformer. Proposed models are evaluated across multiple downstream tasks, including paraphrase detection, semantic textual similarity (STS), and Bangla hate speech detection. The new method consistently outperformed existing Bangla sentence transformers. Moreover, the lightweight architecture and shorter inference time make the models highly suitable for deployment in resource-constrained environments, making them valuable for practical NLP applications in low-resource languages.

Index Terms—*Sentence Transformer, Knowledge Distillation, Paraphrase Detection, Semantic Textual Similarity*

I. INTRODUCTION

In recent years, there have been remarkable advancements in the field of natural language processing (NLP) [1]. One significant development is the emergence of high-quality sentence embedding models [2], which can effectively map sentences into an embedding space based on their meaning and context. However, low-resource languages like Bangla still lag behind high-resource languages like English in terms of the quality of sentence transformers. This disparity exists primarily due to the limited availability of high-quality data in languages like Bangla.

Low-resource languages lack high-quality, large text corpora necessary for developing effective sentence embedding models [3]. To address this gap, our work proposes an approach that utilizes a machine translation dataset [4] instead of a large text corpus to train lightweight sentence transformers for the Bangla language. In this study, we introduce two distinct embedding models [2] for Bangla, trained using an English-Bangla machine translation dataset [5] and two different loss
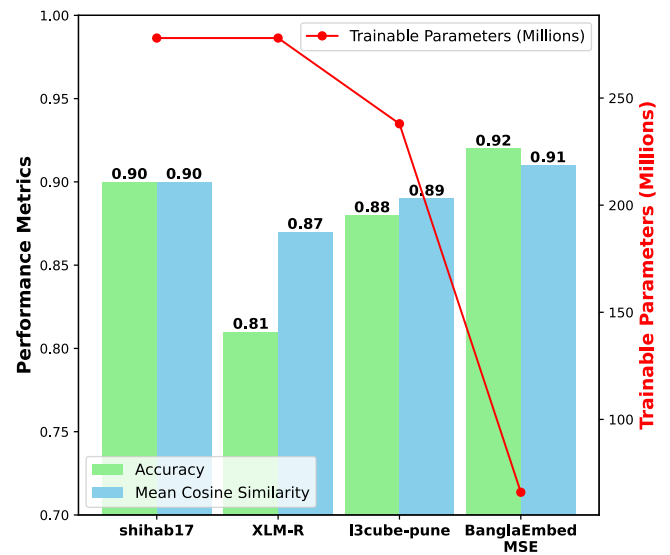

Fig. 1. Performance comparison of our proposed sentence transformer, `BanglaEmbed-MSE`, on the paraphrase detection task, evaluated using accuracy, mean cosine similarity, and the number of trainable parameters.

functions—mean squared error (MSE) and multiple negatives ranking loss [6].

In our study, we employed a knowledge distillation pipeline [7] that utilizes a pre-trained teacher sentence embedding model alongside a custom student sentence embedding model. Since Bangla lacks a high-performing pre-trained sentence embedding model, we used an English pre-trained sentence transformer as the teacher model. As an English sentence and its translated Bangla counterpart convey the same contextual meaning, the embeddings for both sentences should map to the same embedding space. This approach allows us to supervise the training of lightweight Bangla sentence transformers (student models) using a pre-trained English sentence transformer (teacher model). Our proposed methodology facilitates the training of an embedding model for a low-resource language like Bangla by leveraging a pre-trained English sentence trans-

former and a machine translation dataset, rather than relying on a large text corpus. Specifically, we used the English-Bangla Machine Translation dataset introduced by [5] for training. Figure 1 presents a performance comparison of our introduced sentence embedding model, `BanglaEmbed-MSE`, with other existing Bangla sentence transformers on the paraphrase detection task. The significant contributions of this work are as follows:

- Introduction of two lightweight sentence transformers for a low-resource language, Bangla, utilizing mean squared error (MSE) and multiple negatives ranking loss.
- The novel training approach employs cross-lingual knowledge distillation and leverages a pre-trained model from a high-resource language to overcome the lack of large Bangla corpora, using a smaller translation dataset.
- The proposed models are evaluated against existing Bangla sentence transformers across multiple downstream tasks, demonstrating superior performance with reduced computational requirements.

II. RELATED WORKS

Sentence Embeddings. Sentence embedding plays a crucial role in modern language models, capturing semantic meaning that informs the reasoning abilities of these models. Numerous efforts have been made to improve sentence embeddings, particularly in high-resource languages like English [8], [9]. However, significant gaps remain for low-resource languages. To address this, various multilingual approaches [10], [11], have been proposed to extend sentence representation learning to a wider set of languages.

Multilingual Sentence Embeddings. The Multilingual Universal Sentence Encoder (mUSE) [12] employs a dual encoder architecture and was trained using a multi-task learning approach on the SNLI dataset [13]. The model's training included a translation ranking task, where it had to identify the correct translation from a set of candidates. To improve performance, it used hard negative samples—sentences similar to the correct translation but not entirely accurate. Trained on 16 languages, the model showed strong effectiveness in multilingual tasks. For low-resource language translation tasks, Gao et al. [14] employed the pre-trained SimCSE model to generate sentence embeddings, which were combined with representations using an embed-fusion module. This output was fed into the encoder-decoder attention layer of the model's decoder block. During training, only the parameters of the embed-fusion module and transformer model were updated, while the SimCSE parameters were frozen. During inference, the SimCSE and embed-fusion modules acted as plug-ins, ensuring compatibility with various model architectures.

Distillation for Sentence Transformers. Knowledge distillation (KD) has proven to be an effective approach for transferring knowledge from a large, high-capacity model (teacher) to a smaller, more efficient model (student) [7]. In [15], SimTDE

was introduced, employing knowledge distillation to train a student model with a shallower token embedding block and fewer transformer layers than the teacher model. Despite its simpler architecture, the student model achieved performance comparable to the teacher while offering twice the inference efficiency. Similarly, Reimers and Gurevych [10] used a KD-based approach, where a pre-trained English sentence transformer served as the teacher model, and a lightweight student model was trained to learn sentence representations for both English and additional languages. This process enabled the student model to capture cross-lingual representations under the guidance of the more robust teacher model.

III. METHODOLOGY

A. Problem Formulation

Let D be a machine-translation dataset, where each data point consists of a Bangla sentence BN_i and its corresponding English translation EN_i, i.e., $D = \{BN_i, EN_i\}_{i=1}^N$. When an English sentence EN_i is passed through the teacher model T, it generates an embedding $E_T(EN_i)$, which can be considered the *target embedding*, denoted as $\mathbf{y}_i$. Similarly, when the corresponding Bangla sentence BN_i is passed through the student model S, the model produces an embedding $E_S(BN_i)$, which can be considered the *predicted embedding*, denoted as $\hat{\mathbf{y}}_i$. The goal is to optimize the parameters of the student model S such that the distance between the *predicted embedding* $\hat{\mathbf{y}}_i = E_S(BN_i)$ and the *target embedding* $\mathbf{y}_i = E_T(EN_i)$ is minimized.

$$L = \frac{1}{N} \sum_{i=1}^N \text{distance}\,(\mathbf{y}_i, \hat{\mathbf{y}}_i) \tag{1}$$

Here, *distance* represents a metric that quantifies the difference between the teacher's and student's embeddings. By minimizing this distance, the student model S learns to align its output embeddings $\hat{\mathbf{y}}_i = E_S(BN_i)$ as closely as possible to the teacher's embeddings $\mathbf{y}_i = E_T(EN_i)$.

B. Training Pipeline for `BanglaEmbed`

Low-resource languages like Bangla still face a shortage of high-performing sentence embedding models compared to high-resource languages. In this study, we introduce two distinct Bangla sentence transformers: `BanglaEmbed-MNR` and `BanglaEmbed-MSE`. These sentence transformers are trained using a cross-lingual knowledge distillation approach on an English-Bangla machine translation dataset.

Training Data. For training purposes, we utilized the publicly available BanglaNMT machine translation dataset [5], which consists of parallel English and Bangla statements. The initial size of the corpus was 2.75 million sentence pairs. After pre-processing, the dataset was reduced to approximately 2.66 million sentence pairs, which is sufficient for training a sentence transformer. This dataset is the largest available machine translation resource for the Bangla language. The original and the pre-processed datasets are accessible through

English Sentence	Bangla Sentence
He eats nothing other than fruit.	উনি ফল ছাড়া আর কিছুই খান না।
Internet Gaming Worries Saudi Parents.	সৌদি অভিভাবকরা ইন্টারনেট গেমিং নিয়ে চিন্তিত।
Tunisians remembered the 10th anniversary of the death of their country's first president - Habib Bourguiba.	তিউনিশিয়া বাসী তাদের প্রথম প্রেসিডেন্ট হাবিব বুরগুইবার ১০ম মৃত্যুবার্ষিকী পালন করেছেন।

Fig. 2. Sample EN-BN sentence pairs from the machine translation dataset.

the official GitHub repository[1]. A few sample pairs from the training dataset are shown in Figure 2.

Knowledge Distillation. Instead of following the standard training approach, we adopted a unique strategy for training sentence transformers in a low-resource language by leveraging knowledge from a pre-trained sentence transformer in a high-resource language, as proposed by [10]. To implement this, we utilized a knowledge distillation pipeline [7], creating a teacher-student framework between a high-performing English sentence transformer (teacher) and a custom lightweight Bangla sentence transformer (student).

In this process, the pre-trained English sentence transformer (teacher) was provided with English statements from the machine translation dataset (discussed in Section III-B), while the custom lightweight embedding model (student) was fed the corresponding Bangla statements. Since the English and Bangla embeddings represent the same contextual meaning, they should ideally map to the same embedding space. The student model's output embeddings were supervised using the embeddings generated by the pre-trained teacher model. Consequently, knowledge was distilled from the pre-trained English sentence transformer into the lightweight Bangla sentence transformer, completing the distillation process. This cross-lingual distillation method allows for the training of a Bangla sentence embedding model without requiring an extremely large text corpus. Figure 3 illustrates the training pipeline used in this study.

Loss Functions. We employed two different loss functions during the training phase. Specifically, we utilized Mean Squared Error (MSE) and Multiple Negatives Ranking Loss for training `BanglaEmbed-MSE` and `BanglaEmbed-MNR`, respectively.

1) **Mean Squared Error:** To train the `BanglaEmbed-MSE` sentence transformer, the MSE loss function was used, which calculates the squared difference between the embeddings from the English transformer (teacher) and the Bangla transformer (student), as shown in equation 2. The goal is to minimize this loss by aligning the two embeddings in the embedding space.

$$L_{MSE} = \frac{1}{N} \sum_{n=1}^{N} (E_T^i - E_S^i)^2 \qquad (2)$$

Here, E_T and E_S refer to the embeddings generated by the teacher and student models, respectively, and N represents the total number of sentence pairs.

2) **Multiple Negatives Ranking Loss:** This loss function is employed to train `BanglaEmbed-MNR`. The objective of this ranking loss is to minimize the distance between the teacher embedding and the positive sample (correct translation), while maximizing the distance between the teacher embedding and the negative samples (incorrect translations) in the embedding space. During training, one sample from the batch is considered the positive sample, while the remaining samples are treated as negative examples. Equation 3 illustrates the multiple negatives ranking loss.

$$L_{MNR} = -\log \frac{e^{(\cos(E_T, E_S^+))}}{e^{(\cos(E_T, E_S^+))} + \sum_{n=1}^{K} e^{(\cos(E_T, E_S^{-(n)}))}} \qquad (3)$$

Here, E_S^+ refers to the positive sample, and $E_S^{-(n)}$ refers to the n-th negative sample, and K represents the total number of negative samples.

C. Model Overview

For our training pipeline, we utilize two models—a teacher and a student. The teacher model is a pre-trained sentence transformer capable of generating high-quality embeddings for English sentences. Specifically, we use the pre-trained **multi-qa-distilbert-cos-v1** from SBERT, which contains 66.4 million parameters. This model has been trained on 215 million diverse question-answer pairs, making it highly effective for generating robust sentence embeddings.

The student models, `BanglaEmbed-MSE` and `BanglaEmbed-MNR`, are custom embedding models based on the **distilbert-base-uncased** architecture [16], a lightweight variant of BERT [17]. Like the teacher model, the student model also has 66.4 million parameters. The student model was adapted for sentence embedding tasks by adding a pooling layer that averages token embeddings, resulting in compact and efficient sentence representations. The lightweight student model, along with its pooling mechanism, makes it ideal for downstream tasks in low-resource languages like Bangla.

IV. EXPERIMENTS

A. Setup

Dataset. To evaluate the performance of the introduced sentence transformers, we assessed them on three distinct tasks: paraphrase detection, semantic textual similarity (STS), and hate-speech classification. For the paraphrase detection task, we employed the test set from the public *BanglaParaphrase* dataset by [18], which contains 23,300
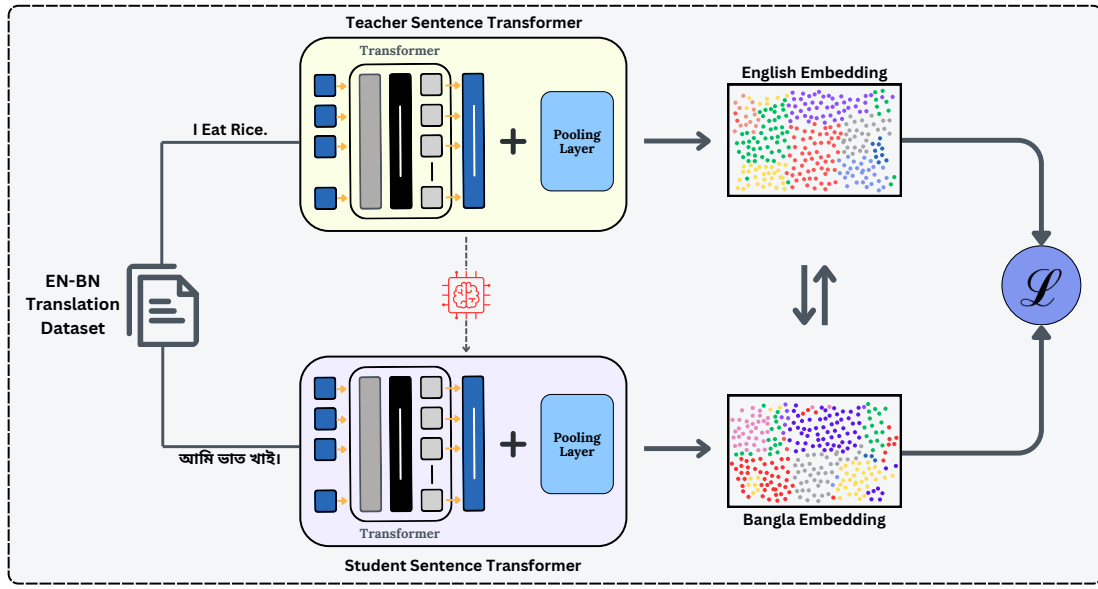
Fig. 3. Proposed cross-lingual knowledge distillation methodology for training the Bangla sentence transformer, leveraging an English-Bangla machine translation dataset. The bidirectional arrow (↓↑) indicates that both English and Bangla embeddings are aligned to map into the same embedding space.

sentence pairs (source and target). For the STS evaluation, we utilized the *SemEval Task 1: Semantic Textual Similarity (STS)* dataset, consisting of 250 pairs of statements with labels ranging from 0 to 5, a widely recognized gold-standard dataset [19]. Given the dataset's high quality, we translated it into Bangla rather than using a lower-quality Bangla STS dataset. The English sentence pairs were translated into Bangla using GPT-4o, followed by manual human validation, while preserving the original labels. Finally, we conducted a qualitative evaluation on the Bangla hate-speech classification task using the *Bengali Hate Speech Dataset* introduced by [20]. This dataset contains 3,420 instances and categorizes hate speech into five distinct types: personal, political, geopolitical, gender abuse, and religious hate speech.

Implementation Details. We introduced two versions of a sentence transformer for Bangla, a low-resource language. These two versions were trained on the BanglaNMT machine translation dataset [5], using mean squared error and multiple negatives ranking loss, respectively. Both sentence transformers were trained for 10 epochs with a batch size of 4, utilizing the *SentenceTransformerTrainer* from Sentence Transformers (SBERT) for efficient model training. The experiments were implemented using the *PyTorch* framework. All training and evaluation were conducted on an NVIDIA RTX 3080 GPU with 12 GB of VRAM. Table I presents the hyperparameters employed during our experiments.

Evaluation. We used three strategies to evaluate the performance of our sentence transformers, alongside other available Bangla sentence transformers. First, we evaluated all models on the Bangla paraphrase detection task by calculating the Mean Cosine Similarity (MCS) between source and target

TABLE I
HYPERPARAMETERS USED DURING OUR EXPERIMENTS.

Hyperparameter	Value
Epochs	10
Batch Size	4
Warmup Ratio	0.1
Learning Rate	5e-5
Optimizer	AdamW

embeddings. Classification accuracy was also measured using a cosine similarity threshold of 0.8; pairs with cosine similarity ≥ 0.8 were classified as paraphrases, while those below were classified as non-paraphrases. Second, we assessed the sentence transformers' performance on the Semantic Textual Similarity (STS) task, where the objective was to assign a similarity score (ranging from 0 to 5) to pairs of Bangla sentences. We used two evaluation metrics for this task: Pearson correlation (r) and Spearman correlation (ρ). Lastly, we conducted a qualitative evaluation by generating t-SNE clustering plots [21] for a Bangla hate-speech classification task consisting of five distinct classes. By examining the separation and clustering of the generated t-SNE plots, we compared the quality of the embeddings given by each of the sentence transformers.

B. Results and Analysis

We compare the performance of our proposed Bangla embedding models across three downstream tasks: Paraphrase Detection, Semantic Textual Similarity (STS), and Hate Speech Classification. Table II presents the results for the paraphrase detection and STS tasks, achieved by both the introduced sentence embedding models and existing embeddings for the Bangla language. Additionally, the table includes the

TABLE II

COMPARISON OF EVALUATION RESULTS FOR TWO DOWNSTREAM TASKS: PARAPHRASE DETECTION AND SEMANTIC TEXTUAL SIMILARITY (STS). HERE, *MCS* DENOTES MEAN COSINE SIMILARITY.

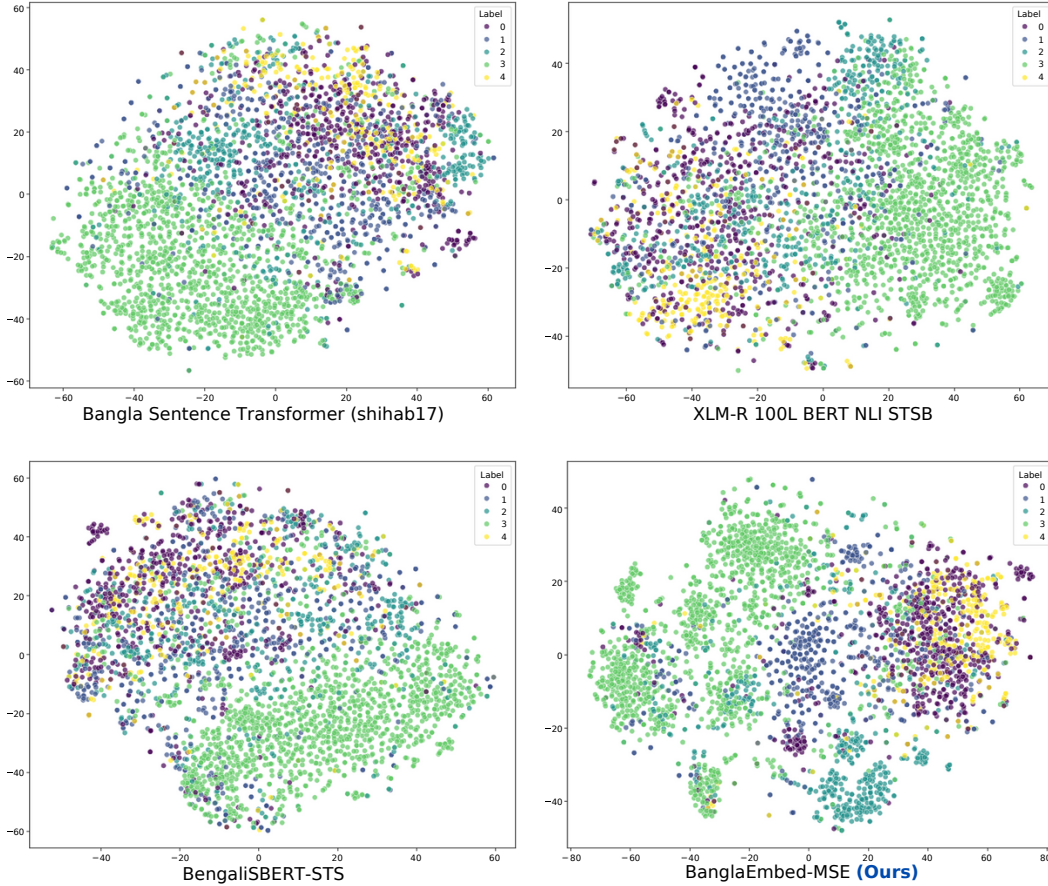| Models	Parameters	Inference Time	Paraphrase Detection		STS	
			MCS	Accuracy	ρ	r
Bangla Sentence Transformer [2]	278 M	13.4 s	0.90	0.90	0.65	0.65
XLM-R 100L BERT NLI STSB [2]	278 M	13.9 s	0.87	0.81	0.65	0.64
BengaliSBERT-STS [22]	238 M	12.7 s	0.89	0.88	0.72	**0.72**
BanglaEmbed-MNR (Ours)	66 M	12.3 s	0.87	0.82	0.63	0.62
BanglaEmbed-MSE (Ours)	66 M	**12.0 s**	**0.91**	**0.92**	**0.73**	0.70

Fig. 4. t-SNE visualizations of four distinct sentence transformers. The `BanglaEmbed-MSE` model shows superior performance in separating clusters, indicating higher quality sentence embeddings compared to the other models.

inference time required to compute embeddings for the source sentences in the BanglaParaphrase dataset [18].

From Table II, we observe the following: *(1)* Both of the introduced embedding models, `BanglaEmbed-MNR` and `BanglaEmbed-MSE`, are lightweight, each with approximately 66 million parameters, making them computationally efficient. This reduced parameter count also results in shorter inference times compared to other existing models.

(2) `BanglaEmbed-MSE` achieves the best performance in paraphrase detection, with a mean cosine similarity of 0.91 and an accuracy of 0.92, outperforming all other models. *(3)* In the semantic textual similarity (STS) task, `BanglaEmbed-MSE` shows strong performance with a Spearman correlation of 0.73 and Pearson correlation of 0.70, while remaining efficient with fewer parameters. *(4)* Although `BanglaEmbed-MNR` does not surpass the highest-performing models, it remains competitive, offering computational efficiency with fewer parameters and a shorter inference time.

[2]https://huggingface.co/shihab17/bangla-sentence-transformer

For the hate speech classification task, we conducted a qualitative evaluation using t-SNE visualizations, as shown in Figure 4. This visualization compares the clustering produced by four different sentence embeddings. As illustrated, `BanglaEmbed-MSE` shows superior clustering and clearer separation of hate speech classes, indicating the generation of higher-quality embeddings compared to the other models.

V. CONCLUSION

In this paper, we introduced two novel Bangla sentence transformers, `BanglaEmbed-MSE` and `BanglaEmbed-MNR`, which outperform existing models with fewer parameters. By employing cross-lingual knowledge distillation, we successfully trained these models for the Bangla language, making a significant contribution to advancing natural language understanding in this low-resource language. While our models demonstrate superior performance on multiple downstream tasks, they also establish a strong baseline for future work in Bangla NLP. However, there remains room for improvement. Future efforts could focus on exploring more efficient architectures and training on more diverse and larger datasets to further enhance the models' performance.

REFERENCES

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410

[3] B. P. King, "Practical natural language processing for low-resource languages." Ph.D. dissertation, 2015.

[4] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering*, vol. 18, pp. 143–153, 2022.

[5] T. Hasan, A. Bhattacharjee, K. Samin, M. Hasan, M. Basak, M. S. Rahman, and R. Shahriyar, "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)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2612–2623. [Online]. Available: https://aclanthology.org/2020.emnlp-main.207

[6] M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil, "Efficient natural language response suggestion for smart reply," *arXiv preprint arXiv:1705.00652*, 2017.

[7] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 670–680. [Online]. Available: https://aclanthology.org/D17-1070

[9] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder for english," in *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, 2018, pp. 169–174.

[10] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. [Online]. Available: https://aclanthology.org/2020.emnlp-main.365

[11] A. Saraswat, K. Abhishek, and S. Kumar, "Text classification using multilingual sentence embeddings," in *Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1*. Springer, 2021, pp. 527–536.

[12] M. Chidambaram, Y. Yang, D. Cer, S. Yuan, Y.-H. Sung, B. Strope, and R. Kurzweil, "Learning cross-lingual sentence representations via a multi-task dual-encoder model," *arXiv preprint arXiv:1810.12836*, 2018.

[13] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, L. Màrquez, C. Callison-Burch, and J. Su, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. [Online]. Available: https://aclanthology.org/D15-1075

[14] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552

[15] J. Xie, C. He, J. Wang, C. Qiu, A. Kebarighotbi, and F. Ghassemi, "Simtde: Simple transformer distillation for sentence embeddings," in *SIGIR 2023*, 2023. [Online]. Available: https://www.amazon.science/publications/simtde-simple-transformer-distillation-for-sentence-embeddings

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[18] A. Akil, N. Sultana, A. Bhattacharjee, and R. Shahriyar, "BanglaParaphrase: A high-quality Bangla paraphrase dataset," in *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)*, Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, Eds. Online only: Association for Computational Linguistics, Nov. 2022, pp. 261–272. [Online]. Available: https://aclanthology.org/2022.aacl-short.33

[19] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, Eds. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–14. [Online]. Available: https://aclanthology.org/S17-2001

[20] M. R. Karim, B. R. Chakravarti, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," in *7th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA,2020)*. IEEE, 2020.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[22] S. Deode, J. Gadre, A. Kajale, A. Joshi, and R. Joshi, "L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert," *arXiv preprint arXiv:2304.11434*, 2023.