L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT

Samruddhi Deode[*], **Janhavi Gadre**[*]
Aditi Kajale[*], **Ananya Joshi**[*]
MKSSS' Cummins College of Engineering for Women, Pune, Maharashtra, India;
L3Cube Pune

Raviraj Joshi
Indian Institute of Technology Madras
Chennai, Tamil Nadu, India;
L3Cube Pune

Abstract

We propose a novel approach to learning cross-lingual sentence representations, eliminating the need for parallel corpora. We simply utilize synthetic monolingual corpora to align pre-trained multilingual BERT models into multi-lingual Sentence BERT (SBERT) models. The proposed approach involves a mixed training strategy that combines translated NLI or STS datasets from low-resource target languages and applies SBERT-like fine-tuning on the vanilla multilingual BERT model. This simple fine-tuning approach with mixed monolingual corpora yields outstanding cross-lingual properties without explicit cross-lingual training. Our approach is validated on 10 major Indic languages and non-Indic languages such as German and French. Using our approach, we introduce L3Cube-IndicSBERT, the first multilingual sentence representation model tailored specifically for Indian languages, including Hindi, Marathi, Kannada, Telugu, Malayalam, Tamil, Gujarati, Odia, Bengali, and Punjabi. IndicSBERT exhibits remarkable cross-lingual capabilities, outperforming alternatives like LaBSE, LASER, and paraphrase-multilingual-mpnet-base-v2 in Indic cross-lingual and monolingual sentence similarity tasks.

1 Introduction

Semantic Textual Similarity (STS) is a crucial task in Natural Language Processing (NLP), which measures the equivalence between the meaning of two or more text segments (Agirre et al., 2013; Cer et al., 2017). The aim of STS is to identify the semantic similarity between text inputs, taking into account their meaning rather than just surface features like word frequency and length (Adi et al.). The concept is widely used in various NLP applications, including question-answering (Huang et al., 2020), information retrieval (Li and Lu, 2016), text generation (Iqbal and Qureshi, 2022), etc.

[*]Authors contributed equally to this research.

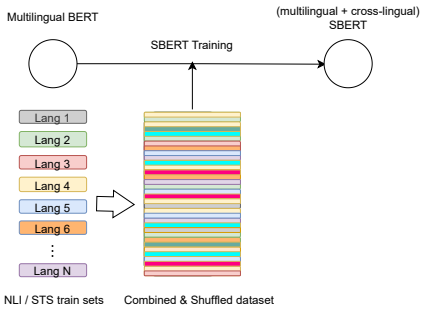

Figure 1: An embarrassingly simple approach for learning cross-lingual sentence representations using synthetic monolingual corpus.Monolingual datasets of different languages are combined together and used for SBERT training

Sentence-BERT (Reimers and Gurevych, 2019), a modified version of the BERT architecture was designed to generate sentence representations for the improved semantic similarity between sentences. The SBERT makes use of a siamese network (Koch et al., 2015) and is trained using specific datasets like STS, resulting in representations specifically geared for semantic similarity.

Recent works are focused on multilingual SBERT models capable of encoding sentences from different languages to the same representation space (Schwenk and Douze, 2017; Yang et al., 2020). With these models, it is possible to extend NLP tasks to different languages without training a language-specific model. These multilingual models often employ teacher-student training (Heffernan et al., 2022; Reimers and Gurevych, 2020) or are based on translation ranking tasks (Feng et al., 2022). These methods make use of parallel translation corpus in target languages for training a cross-lingual model (Tan and Koehn, 2022; Artetxe and Schwenk, 2019; Conneau et al., 2018). Even vanilla multilingual BERT models have been shown to have surprisingly good cross-lingual properties (Wu and Dredze, 2019; Pires et al., 2019; Wu and Dredze, 2020). However, their performance is

not good as the multilingual sentence BERT models.

In this work, we propose a simple approach to learning cross-lingual sentence representations without using any parallel corpus. We leverage pre-trained multilingual transformer models and fine-tune them using our mixed training strategy, as depicted in Figure 1. We mix the monolingual translated NLI / STSb corpus for target languages and fine-tune the multilingual BERT model in an SBERT setup. We show that this simple mixed data training is sufficient to train a multilingual SBERT model with strong cross-lingual properties. This strategy is capable of significantly amplifying the existing cross-lingual properties of the vanilla multilingual BERT model. Our approach is inspired by a recent work (Joshi et al., 2022) that shows that translated STSb and NLI can be used to train high-quality monolingual SBERT models.

We present L3Cube-IndicSBERT a multilingual SBERT model for 10 Indian regional languages Hindi, Marathi, Kannada, Telugu, Malayalam, Tamil, Gujarati, Odia, Bengali, Punjabi, and English. The IndicSBERT uses MuRIL (Khanuja et al., 2021) as the base model and performs better than existing multilingual/cross-lingual models like LASER, LaBSE, and paraphrase-multilingual-mpnet-base-v2. These models are compared on monolingual and cross-lingual sentence similarity tasks. We also evaluate these models on real text classification datasets to show that the synthetic data training generalizes well to real datasets. Further, we also release monolingual SBERT models for individual languages to show that IndicSBERT performs competitively with the monolingual variants.

Our main contributions are as follows:

- We propose a simple strategy to train cross-lingual sentence representations using a pre-trained multilingual BERT model and synthetic NLI/STS data. Unlike previous approaches, it does not use any cross-lingual data or any complex training strategy.
- We present **IndicSBERT**[1][2], the first multilingual SBERT model trained specifically for Indic languages. The model performs better than state-of-the-art LaBSE and paraphrase-multilingual-mpnet-base-v2 models.
- We also release monolingual SBERT models for 10 Indic languages. To the best of our knowledge, this work is first to introduce the majority of these models.

2 Related Work

BERT (Devlin et al., 2019) (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer network that is widely regarded as one of the best language models for natural language processing (NLP) tasks and can be used to extract sentence representations for a variety of tasks. For the Indian languages, the available multilingual BERT models include mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), IndicBERT (Kakwani et al., 2020), and MuRIL (Khanuja et al., 2021).

Specific sentence embedding models have been proposed over time (Cer et al., 2018; Conneau et al., 2017; Yang et al., 2020; Ni et al., 2022) and are superior to word embedding models (Pennington et al., 2014; Bojanowski et al., 2017; Peters et al., 2018; Ethayarajh, 2019) as they capture the meaning of the entire sentence rather than individual words. While BERT is trained to generate word embeddings, Sentence-BERT (Reimers and Gurevych, 2019) modifies the architecture and fine-tunes the pre-trained BERT model for generating sentence embeddings.

LaBSE (Feng et al., 2022), a sentence-BERT model is designed to generate language-agnostic sentence embeddings that can be used for cross-lingual NLP tasks, while LASER (Artetxe and Schwenk, 2019) is a multilingual sentence embedding model that generates high-quality sentence embeddings for multiple low-resource languages. These models have been explicitly trained using parallel translation corpus. Similarly, by aligning the embeddings of parallel sentences in many languages, Cross-Lingual Transfer (CT) (Artetxe and Schwenk, 2019) technique learns a shared space for sentence embeddings across multiple languages. Thus, in the multilingual category, several BERT, as well as Sentence-BERT models, have been developed to date.

However, monolingual models are typically found to be performing better than multilingual ones. In a previous study (Scheible et al., 2020), a German RoBERTa-based BERT model, with slight

[1] https://huggingface.co/l3cube-pune/indic-sentence-bert-nli

[2] https://huggingface.co/l3cube-pune/indic-sentence-similarity-sbert

adjustments to its hyperparameters, was found to yield superior results than all other German and multilingual BERT models. Similarly, in (Straka et al., 2021) a Czech RoBERTa language model has been shown to perform better than other Czech and multilingual models. In (Velankar et al., 2022) and (Joshi, 2022b), monolingual BERT models for the Marathi language were studied and found to perform better than their multilingual counterparts. Similarly, in this study, we propose monolingual SBERT models for the ten most prominent Indic languages. Additionally, we also propose a multilingual model tailored specifically to these languages. Considering that other multilingual models are trained to support a greater number of languages, our model is better suited for Indian languages, as it is specifically optimized for them.

3 Experimental Setup

3.1 Datasets

The results in (Joshi et al., 2022), indicate the efficacy of using synthetic datasets in creating MahaSBERT-STS and HindSBERT-STS. Thus, we utilize the machine-translated IndicXNLI and STSb datasets for training our models. Our models are evaluated on the synthetic STSb dataset, as well as on real-world classification datasets, as described below.

The **IndicXNLI**[3] dataset comprises of English XNLI data translated into eleven Indian languages including Hindi and Marathi. To train the monolingual Sentence-BERT models, we use the training samples of the corresponding language from IndicXNLI. To ensure balanced training data for the multilingual IndicSBERT, we combine and randomly shuffle the IndicXNLI training samples of ten languages.

The **STS benchmark (STSb)**[4] dataset is commonly utilized for evaluating supervised Semantic Textual Similarity (STS) systems. The dataset includes 8628 human-annotated sentence pairs from captions, news, and forums and is divided into 5749 for training, 1500 for development and 1379 for testing.

The lack of well-organized benchmark datasets like STS for Indic languages is a major issue, therefore to make the STSb dataset accessible for all ten Indian languages used in this study, we translate it using Google Translate and use the resulting train samples of the corresponding language for each monolingual model and a combined dataset of ten languages for the multilingual model. We use the testing samples from the corresponding translated STSb dataset to evaluate each model based on the embedding similarity metric. For evaluating the cross-lingual property, we construct a dataset of STSb sentence pairs with each pair comprising two sentences from different languages.

We also evaluate the models on a real text classification dataset to show that the sentence representations from the models trained using synthetic datasets also generalize well to real datasets. We choose the **IndicNLP news article classification datasets** (Kunchukuttan et al., 2020) for the purpose of evaluation as it is the only one currently present which supports 12 Indic languages.

3.2 Models

In our experiment, we use different BERT models, including both monolingual and multilingual ones which are described below. The training procedure is applied over some of these models which serve as a base for creating Sentence-BERT.

3.2.1 Multilingual BERT models:

In this work, we utilize multilingual BERT models like mBERT (Devlin et al., 2019), MuRIL (Khanuja et al., 2021) and multilingual sentence representation models like LaBSE (Feng et al., 2022), paraphrase-multilingual-mpnet-base-v2[5], and LASER (Artetxe and Schwenk, 2019). Out of these MuRIL is used as a base model for the cross-lingual setting whereas all models have be fine-tuned in the monolingual setting.

3.2.2 Monolingual BERT models:

We also use the monolingual BERT models for the 10 Indic languages, released by L3cube-Pune[6] as the base models. These models are termed as HindBERT, MahaBERT (Joshi, 2022b), KannadaBERT, TeluguBERT, MalayalamBERT, TamilBERT, GujaratiBERT, OdiaBERT, BengaliBERT, and PunjabiBERT. Further details about these models can be found in (Joshi, 2022a).

[3] https://github.com/divyanshuaggarwal/IndicXNLI

[4] https://huggingface.co/datasets/stsb_multi_mt

[5] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[6] https://huggingface.co/l3cube-pune

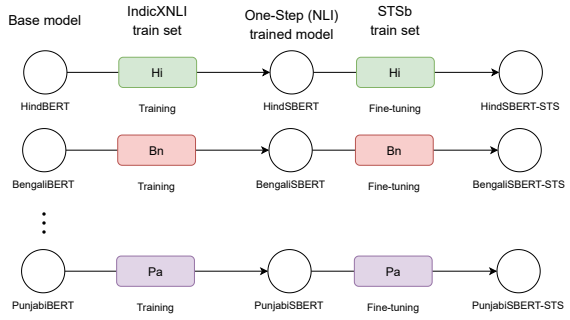

Figure 2: Two-step (NLI + STS) training process of the monolingual SBERT models

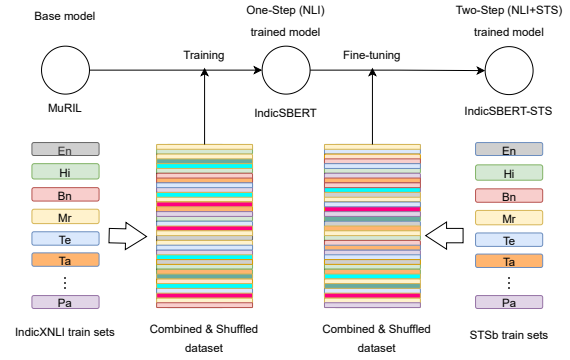

Figure 3: Two-step (NLI + STS) training process of the multilingual IndicSBERT models

3.3 SBERT Training

In order to achieve competitive performance, sentence embedding models typically require significant amounts of training data and fine-tuning over the target task. Unfortunately, in many scenarios, only limited amounts of training data are available. Several unsupervised and semi-supervised approaches have been proposed to overcome the lack of a large training dataset. However, the models trained using unsupervised techniques give inferior performance than those trained using supervised learning.

In this work, we, therefore, use a supervised training approach, wherein we address the scarcity of specialized datasets, such as NLI and STS, in Indian languages by machine translating the English versions of these datasets into the respective Indian languages. We follow a two-step procedure to train the monolingual SBERT models and the multilingual IndicSBERT model. The monolingual BERT model serves as the base for monolingual SBERT while MuRIL serves as the base model for IndicSBERT.

During the **first training step**, the model focuses on natural language inference, which involves assessing the logical relationship between a premise and hypothesis. The goal is to classify the relationship into three categories: entailment (hypothesis can be inferred from the premise), contradiction (negation of the hypothesis can be inferred from the premise), or neutral (no clear relationship between the two). The base model is trained on the IndicXNLI dataset, containing 392702 labeled sentence pairs for this purpose.

To enhance the model's effectiveness, we replace the Softmax-Classification-Loss with the Multiple Negatives Ranking Loss function used in (Reimers and Gurevych, 2019). This change is beneficial for similarity-based tasks as it enables the model to learn similarities and dissimilarities between examples using multiple negative samples simultaneously. This results in a more complex decision boundary and improves the model's robustness to outliers and variations in data.

The training data consists of triplets: [(a1, b1, c1), ..., (an, bn, cn)], where (ai, bi) represents similar sentences and (ai, ci) represents dissimilar sentences. To create these triplets, (bi) is randomly selected from sentences labeled as 'entailment' for (ai), and (ci) is chosen from sentences labeled as 'contradiction' for (ai), referred to as hard-negatives. Despite lexical similarities, (bi) and (ci) have different meanings on a semantic level. The model is trained with 1 epoch, a batch size of 4, using the AdamW optimizer with a learning rate of 2e-05. AdamW incorporates weight decay regularization to prevent overfitting and enhance the model's generalization.

The models obtained after applying the first step (NLI only) of training are named as **MahaSBERT, HindSBERT, KannadaSBERT, TeluguSBERT, MalayalamSBERT, TamilSBERT, GujaratiSBERT, OdiaSBERT, BengaliSBERT, and PunjabiSBERT** that are made publicly available[6].

In the **second step**, the model trained in the previous step undergoes fine-tuning using the translated STSb dataset. The STS benchmark is a widely used dataset for evaluating NLP models' performance in determining text similarity. The fine-tuning process utilizes the Cosine Similarity Loss as the loss function, which measures similarity based on the angle between vectors rather than their magnitudes. The training includes 4 epochs, adopting the AdamW optimizer with a learning rate of 2e-05.

The final models obtained after applying the two-step procedure (NLI + STS) are named as **MahaSBERT-STS, HindSBERT-STS, KannadaSBERT-STS, TeluguSBERT-STS, MalayalamSBERT-STS, TamilSBERT-STS, GujaratiSBERT-STS, OdiaSBERT-STS, BengaliSBERT-STS, and PunjabiSBERT-STS** and are made publicly available[6]. In addition to the models mentioned above, we also release the multilingual SBERT models named **IndicSBERT**[1] and **IndicSBERT-STS**[2]. These multilingual models support the 11 languages of English, Hindi, Marathi, Kannada, Telugu, Malayalam, Tamil, Gujarati, Odia, Bengali, and Punjabi.

4 Evaluation

4.1 Evaluation Methodology

We evaluate the SBERT models on the basis of the Embedding Similarity score as well as classification accuracy. The Embedding Similarity evaluation is performed by calculating the Pearson and Spearman rank correlation of the embeddings for different similarity metrics with the gold-standard scores. A high score in embedding similarity indicates that the embeddings being compared are of high quality in relation to the benchmark embeddings.

In our experiment, we use the cosine similarity metric and the value of Spearman correlation to evaluate sentence similarity models. The choice of cosine similarity is based on its superiority compared to traditional distance metrics such as Euclidean or Manhattan distance. Unlike these distance metrics, cosine similarity measures the cosine of the angle between the vectors representing the sentences and considers only their direction, making it less affected by scaling and more computationally efficient. Additionally, cosine similarity takes into account shared terms and contexts, providing a more accurate representation of semantic relationships between sentences. Spearman correlation, on the other hand, is used in preference to Pearson correlation because it is more robust to non-linear relationships and handles ties in data. Unlike Pearson correlation, which assumes a linear relationship, Spearman correlation measures the rank relationship between two variables, making it better equipped to accurately assess a model's performance in cases where the relationship is non-linear.

In this study, the text classification datasets were used to evaluate the performance of BERT and SBERT models in generating sentence embeddings. The K Nearest Neighbors (KNN) algorithm was applied to classify the texts based on proximity. The Minkowski distance, a generalized form of both the Euclidean and Manhattan distances, is employed. The optimal value of k was determined using a validation dataset and then used to calculate the accuracy of the test dataset, with results reported in Tables 2, 3.

Table 1: Embedding Similarity scores of vanilla, one-step and two-step trained variants of the mBert, MuRIL, LaBSE and monolingual BERT across 10 Indian languages

	Multilingual base									Monolingual base		
Base model:	**mBERT**			**MuRIL**			**LaBSE**			**BERT**		
Training steps[††]	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
Hindi (hi)	0.49	0.64	**0.75**	0.52	0.74	**0.83**	0.72	0.75	**0.84**	0.5	0.77	**0.85**
Bengali (bn)	0.5	0.65	**0.75**	0.55	0.74	**0.82**	0.71	0.75	**0.81**	0.5	0.72	**0.81**
Marathi (mr)	0.47	0.65	**0.72**	0.56	0.74	**0.81**	0.7	0.75	**0.82**	0.54	0.77	**0.83**
Telugu (te)	0.53	0.62	**0.73**	0.6	0.71	**0.8**	0.73	0.73	**0.81**	0.58	0.72	**0.8**
Tamil (ta)	0.49	0.65	**0.75**	0.6	0.72	**0.8**	0.72	0.74	**0.82**	0.59	0.72	**0.8**
Gujarati (gu)	0.47	0.65	**0.74**	0.58	0.72	**0.8**	0.73	0.73	**0.82**	0.55	0.74	**0.82**
Kannada (kn)	0.52	0.68	**0.75**	0.6	0.75	**0.82**	0.72	0.76	**0.82**	0.57	0.74	**0.82**
Odia (or)[†]	-	-	-	0.45	0.58	**0.69**	0.6	0.6	**0.73**	0.45	0.59	**0.71**
Malayalam (ml)	0.46	0.57	**0.67**	0.53	0.66	**0.74**	0.66	0.66	**0.74**	0.5	0.69	**0.76**
Punjabi (pa)	0.43	0.59	**0.68**	0.45	0.65	**0.74**	0.64	0.67	**0.75**	0.5	0.68	**0.75**

Table 2: KNN classification scores of vanilla, one-step and two-step trained variants of the mBert, MuRIL, LaBSE and monolingual BERT across 10 Indian languages

	Multilingual base									Monolingual base		
Base model:	**mBERT**			**MuRIL**			**LaBSE**			**BERT**		
Training steps[††]	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
Hindi (hi)	0.62	0.6	0.62	0.67	0.7	0.69	0.68	0.64	0.65	0.7	0.69	0.68
Bengali (bn)	0.97	0.96	0.97	0.97	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.98
Marathi (mr)	0.98	0.97	0.97	0.97	0.98	0.98	0.98	0.99	0.99	0.98	0.98	0.99
Telugu (te)	0.98	0.97	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.95	0.99
Tamil (ta)	0.96	0.96	0.95	0.96	0.97	0.97	0.96	0.97	0.96	0.96	0.97	0.97
Gujarati (gu)	0.95	0.94	0.93	0.97	0.98	0.99	0.99	0.96	0.99	0.98	0.99	0.99
Kannada (kn)	0.96	0.95	0.94	0.97	0.97	0.97	0.96	0.96	0.97	0.97	0.97	0.97
Odia (or)[†]	-	-	-	0.97	0.97	0.97	0.98	0.97	0.97	0.98	0.97	0.97
Malayalam (ml)	0.85	0.86	0.84	0.9	0.92	0.91	0.92	0.9	0.9	0.92	0.92	0.92
Punjabi (pa)	0.94	0.96	0.92	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.95	0.96

Table 3: Comparison of Embedding Similarity scores and KNN classification scores of the IndicSBERT model with it's 2 step trained variant: IndicSBERT-STS

	Embedding Similarity		**Classification Accuracy**	
	IndicSBERT	**IndicSBERT-STS**	**IndicSBERT**	**IndicSBERT-STS**
Hindi (hi)	0.76	**0.8**	0.68	0.65
Bengali (bn)	0.76	**0.81**	0.98	0.97
Marathi (mr)	0.75	**0.8**	0.98	0.98
Telugu (te)	0.74	**0.8**	0.99	0.98
Tamil (ta)	0.74	**0.8**	0.96	0.95
Gujarati (gu)	0.76	**0.81**	0.99	0.99
Kannada (kn)	0.76	**0.81**	0.96	0.95
Odia (or)	0.66	**0.73**	0.97	0.95
Malayalam (ml)	0.7	**0.76**	0.91	0.89
Punjabi (pa)	0.7	**0.76**	0.95	0.96

4.2 Evaluation Results & Discussion

Our observations are discussed below.

[†]Odia language is not supported by mBERT

[††]Training steps= 0 indicates the vanilla base model, 1 denotes single-step NLI training over the base model, while 2 denotes the two-step trained model

Table 4: Zero-shot performance of multilingual models, in terms of Embedding similarity score

	mBERT	MuRIL	LASER	mpnet-base	LaBSE	IndicSBERT	IndicSBERT-STS
Hindi	0.49	0.52	0.64	0.79	0.72	0.75	0.82
Bengali	0.5	0.55	0.68	0.66	0.71	0.76	0.82
Marathi	0.47	0.56	0.6	0.75	0.7	0.76	0.81
Telugu	0.53	0.6	0.59	0.64	0.73	0.74	0.81
Tamil	0.49	0.6	0.49	0.65	0.72	0.73	0.82
Gujarati	0.47	0.58	0.14	0.73	0.73	0.74	0.82
Kannada	0.52	0.6	0.17	0.65	0.72	0.76	0.83
Odia[†]	-	0.45	0.29	0.48	0.6	0.62	0.75
Malayalam	0.46	0.53	0.6	0.6	0.66	0.68	0.78
Punjabi	0.43	0.45	0.12	0.56	0.64	0.68	0.77

1. **AVG pooling shows better performance than CLS**

We find that monolingual SBERT models generate superior embedding similarity scores when AVG pooling is utilized instead of CLS, across all 10 Indic languages. The same trend is observed for IndicSBERT, where AVG pooling produces better results than CLS for embedding similarity. Hence, the AVG pooling values are reported in this work.

2. **NLI + STS training works better**

Fine-tuning the pre-trained models using NLI followed by STSb gives an upper hand over single-step training using the NLI dataset alone. Figure 4 compares the embedding similarities for the Vanilla, One-step trained, and Two-step trained monolingual models. We observe that the Two-step trained models surpass the one-step and Vanilla models in terms of performance across all 10 Indic languages. Fine-tuning with the STSb dataset results in a significant increase in embedding similarity for the monolingual SBERT models as well as for IndicSBERT, as demonstrated by Tables 1, 3, and Figures 4,5. Figure 5 demonstrates that the two-step training on IndicSBERT, which employs MuRIL as its base model, increases the embedding similarity scores nearly two-fold in comparison to the vanilla MuRIL model. While we mainly focus on cross-lingual performance in this work, similar observations in the context of monolingual SBERT have been thoroughly documented in (Joshi et al., 2022).

3. **SBERT models trained on synthetic corpus work well with real-world classification datasets**

We evaluate our sentence-BERT models on real-world news classification datasets to ensure that the models do not overfit the noise in the synthetic corpus. The results presented in Tables 2 and 3 indicate that SBERT models trained on translated corpora perform competitively compared to their original base models on classification datasets. The classification accuracy is neither improved nor deteriorated due to the two-step training.

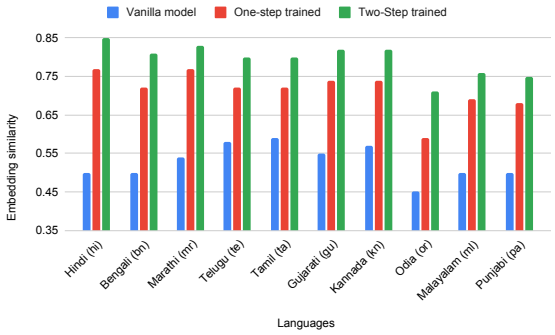

Figure 4: Embedding similarity score comparison of our SBERT models with monolingual BERT base models (E.g. For Bengali language, vanilla model= bengali-bert, One-step model= BengaliSBERT, Two-step model= BengaliSBERT-STS are compared)

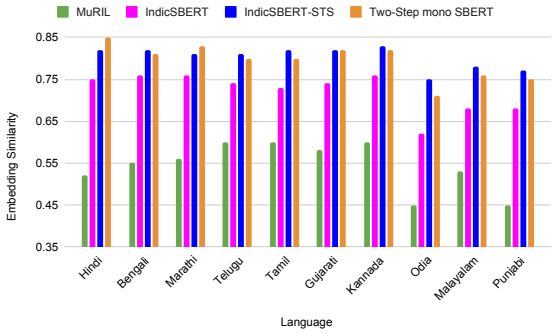

Figure 5: Embedding similarity score comparison of multilingual models: MuRIL, IndicSBERT, IndicSBERT-STS with that of monolingual SBERT models (E.g. for Bengali language, the embedding similarity scores of MuRIL, IndicSBERT, IndicSBERT-STS and BengaliSBERT-STS are compared)

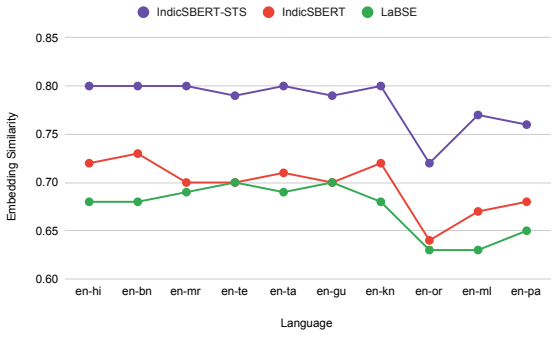

Figure 6: Cross-lingual performance shown by IndicSBERT-STS, IndicSBERT and LaBSE models. The performance is determined in terms of embedding similarity between English and Indian languages. (E.g. en-hi column shows the embedding similarity between English and Hindi texts)

4. **Multilingual Indic-SBERT is competitive with monolingual SBERT models**
Our experiments indicate that the multilingual IndicSBERT model demonstrates equivalent or better performance compared to monolingual SBERT models in terms of embedding similarity scores, as evidenced by Tables 1, 3. In Figure 5, we observe that both the IndicSBERT-STS and two-step monolingual SBERT models perform comparably, with slight performance differences for certain languages. Except for Hindi, Marathi and Gujarati languages, the IndicSBERT-STS outperforms the SBERT models of the respective languages. This shows that the languages have assisting capabilities and the gains are higher for extremely low-resource languages like Odia and Punjabi.

5. **IndicSBERT works significantly better than existing multilingual models**
Figure 7, as well as Table 4, compare the zero-shot embedding similarities of mBERT, MuRIL, LASER, multilingual-mpnet-base, LaBSE, and IndicSBERT models on STSb for all 10 Indic languages, with the IndicSBERT based models clearly outperforming the others. Both IndicSBERT and IndicSBERT-STS produce richer embeddings than the publicly available LaBSE, which is shown in Table 4. Thus, the IndicSBERT is the best-performing model among all the other publicly available multilingual models despite having the lowest number of trainable parameters.

6. **IndicSBERT shows exceptional cross-lingual properties, outperforming the LaBSE**
The results presented in the Table 6 and Figure 6 demonstrate IndicSBERT's robust cross-lingual performance across all language pairs, surpassing the performance of LaBSE by a significant margin. Overall, the multilingual IndicSBERT model demonstrates proficiency in processing both monolingual and multilingual datasets. In addition, IndicSBERT has the potential to enhance the precision and effectiveness of cross-lingual information retrieval systems and semantic search engines as it can handle queries and documents in multiple Indian languages. This characteristic holds particular importance for countries such as India, where multilingual communication is common, and organizations face the challenge of accommodating diverse language requirements.

7. **Multilingual models are indeed cross-lingual learners, the enhancement of cross-lingual properties is generalizable to non-Indic languages**
The performance of mBERT with mixed language NLI training on diverse languages like English, Hindi, German, and French is presented in Table 5. The results demonstrate a considerable improvement in the cross-lingual performance of the one-step trained model as compared to the vanilla mBERT. These findings support the effectiveness of the proposed mixed-language training technique in producing models with enhanced cross-lingual properties not only for Indic languages but also for other languages.

Table 5: Cross-lingual performance of mBERT, single-step trained for 4 languages: Hindi, English, German and French. For every language-pair, the values reported from top to bottom correspond to the Embedding Similarity scores of one-step mBERT, and vanilla mBERT respectively

	Hindi	English	German	French
Hindi	**0.68** 0.48	**0.5** 0.3	**0.5** 0.3	**0.48** 0.32
English	**0.51** 0.31	**0.77** 0.5	**0.6** 0.4	**0.63** 0.41
German	**0.49** 0.3	**0.6** 0.4	**0.7** 0.48	**0.56** 0.39
French	**0.49** 0.29	**0.63** 0.39	**0.57** 0.37	**0.72** 0.49

5 Conclusion

Our research addresses the lack of high-quality language models for low-resource Indian languages by introducing SBERT models trained on synthetic corpora for ten popular Indian languages. Evaluations based on embedding similarity and text classification datasets show that our monolingual SBERT models outperform vanilla BERT models. Additionally, we have developed IndicSBERT, a strong performing multilingual model that surpasses existing models like LaBSE and paraphrase-multilingual-mpnet-base-v2. Our proposed approach simplifies cross-lingual sentence BERT training using translated monolingual datasets and vanilla multilingual BERT.

Indian languages pose a unique challenge, being diverse and having low-resource corpora. Our study highlights the effectiveness of the two-step training method in developing both monolingual SBERT models and the multilingual IndicSBERT. Its robust cross-lingual capability makes IndicSBERT a superior choice for applications that require accurate and efficient multilingual NLP. Thus,

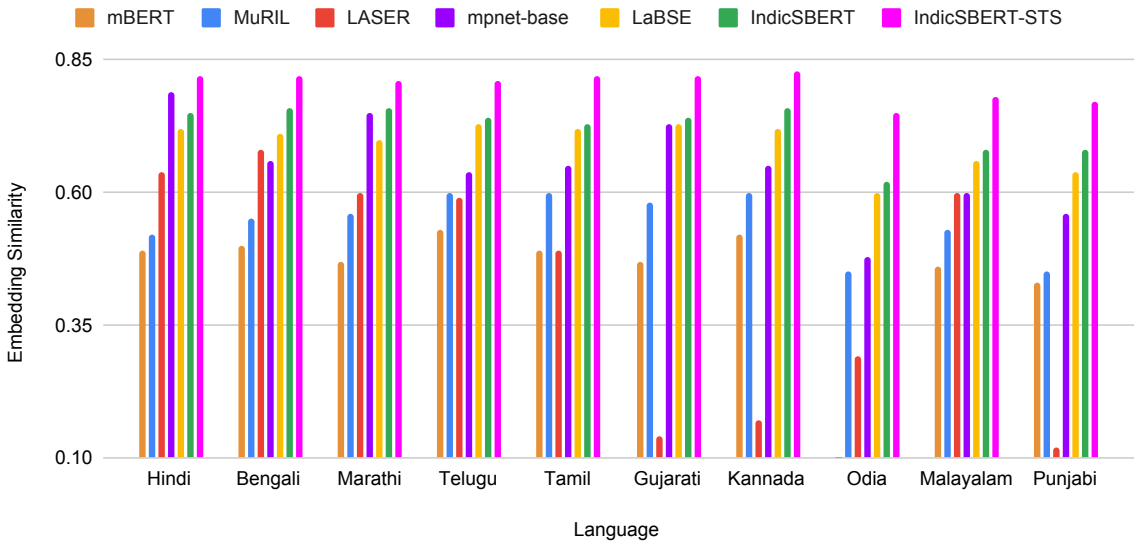

Figure 7: Embedding Similarity score comparison of multilingual models

Table 6: Cross-lingual performance of IndicSBERT-STS, IndicSBERT and LaBSE. For every language-pair, the values reported from top to bottom correspond to the Embedding Similarity scores of IndicSBERT-STS, IndicSBERT and LaBSE respectively

	English	Hindi	Bengali	Marathi	Telugu	Tamil	Gujarati	Kannada	Oriya	Malayalam	Punjabi
English	**0.85** 0.8 0.72	**0.8** 0.72 0.68	**0.8** 0.73 0.68	**0.8** 0.7 0.69	**0.79** 0.7 0.7	**0.8** 0.71 0.69	**0.79** 0.7 0.7	**0.8** 0.72 0.68	**0.72** 0.64 0.63	**0.77** 0.67 0.63	**0.76** 0.68 0.65
Hindi	**0.82** 0.72 0.7	**0.82** 0.75 0.72	**0.79** 0.71 0.69	**0.79** 0.7 0.7	**0.77** 0.68 0.7	**0.78** 0.68 0.69	**0.79** 0.69 0.71	**0.79** 0.69 0.68	**0.72** 0.62 0.62	**0.76** 0.65 0.62	**0.76** 0.68 0.64
Bengali	**0.82** 0.73 0.69	**0.79** 0.7 0.69	**0.82** 0.76 0.71	**0.79** 0.7 0.69	**0.77** 0.68 0.7	**0.77** 0.68 0.69	**0.79** 0.7 0.71	**0.79** 0.7 0.69	**0.73** 0.63 0.64	**0.76** 0.65 0.64	**0.76** 0.67 0.66
Marathi	**0.8** 0.7 0.68	**0.78** 0.7 0.68	**0.78** 0.7 0.69	**0.81** 0.76 0.7	**0.76** 0.67 0.69	**0.77** 0.66 0.68	**0.78** 0.69 0.7	**0.78** 0.69 0.68	**0.72** 0.62 0.63	**0.75** 0.65 0.64	**0.75** 0.67 0.65
Telugu	**0.79** 0.72 0.7	**0.77** 0.68 0.7	**0.77** 0.68 0.7	**0.76** 0.68 0.7	**0.81** 0.74 0.73	**0.77** 0.68 0.7	**0.76** 0.67 0.71	**0.78** 0.69 0.69	**0.71** 0.6 0.63	**0.74** 0.64 0.64	**0.73** 0.65 0.66
Tamil	**0.8** 0.71 0.69	**0.77** 0.67 0.7	**0.77** 0.67 0.69	**0.76** 0.67 0.69	**0.76** 0.67 0.7	**0.82** 0.73 0.72	**0.76** 0.65 0.7	**0.77** 0.68 0.68	**0.7** 0.58 0.62	**0.75** 0.64 0.62	**0.73** 0.64 0.64
Gujarati	**0.8** 0.7 0.7	**0.79** 0.69 0.7	**0.78** 0.69 0.7	**0.79** 0.69 0.69	**0.76** 0.67 0.7	**0.76** 0.66 0.69	**0.82** 0.74 0.73	**0.77** 0.68 0.68	**0.73** 0.6 0.63	**0.74** 0.63 0.63	**0.76** 0.67 0.66
Kannada	**0.8** 0.71 0.68	**0.77** 0.68 0.67	**0.77** 0.69 0.68	**0.77** 0.68 0.68	**0.77** 0.68 0.69	**0.77** 0.67 0.67	**0.76** 0.66 0.69	**0.83** 0.76 0.72	**0.7** 0.59 0.62	**0.75** 0.65 0.62	**0.73** 0.64 0.64
Oriya	**0.72** 0.62 0.6	**0.71** 0.61 0.59	**0.72** 0.61 0.6	**0.7** 0.6 0.6	**0.7** 0.6 0.6	**0.7** 0.58 0.6	**0.72** 0.6 0.61	**0.7** 0.6 0.6	**0.75** 0.62 0.6	**0.68** 0.56 0.58	**0.7** 0.6 0.59
Malayalam	**0.77** 0.68 0.64	**0.74** 0.65 0.62	**0.75** 0.66 0.64	**0.74** 0.66 0.64	**0.74** 0.65 0.65	**0.75** 0.65 0.64	**0.73** 0.63 0.65	**0.74** 0.65 0.64	**0.69** 0.57 0.6	**0.78** 0.68 0.66	**0.7** 0.62 0.6
Punjabi	**0.76** 0.68 0.65	**0.76** 0.67 0.63	**0.76** 0.67 0.66	**0.75** 0.66 0.65	**0.73** 0.65 0.66	**0.74** 0.64 0.65	**0.76** 0.66 0.66	**0.74** 0.66 0.64	**0.7** 0.6 0.62	**0.71** 0.61 0.6	**0.77** 0.68 0.64

we make a significant contribution to IndicNLP, particularly in the context of the world becoming more globalized, and the need for accurate and efficient multilingual NLP models.

As part of this publication, we release the monolingual SBERTs and the multilingual IndicSBERT, opening up new possibilities for NLP research and applications in low-resource Indian languages. Our work emphasizes the significance of combining sentence-level embeddings and multilingual capabilities for optimal results in multilingual NLP tasks, contributing to the development of high-quality language models for Indian languages.

Acknowledgements

This work was done under the L3Cube Pune mentorship program. We would like to express our gratitude towards our mentors at L3Cube for their continuous support and encouragement. This work is a part of the L3Cube-MahaNLP project (Joshi, 2022c).

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Daniel Cer, Mona Diab, Eneko E Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. In *The 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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 55–65.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages.

Zhen Huang, Shiyi Xu, Minghao Hu, Xinyi Wang, Jinyan Qiu, Yongquan Fu, Yuncai Zhao, Yuxing Peng, and Changjian Wang. 2020. Recent trends in deep learning based open-domain textual question answering systems. *IEEE Access*, 8:94341–94356.

Touseef Iqbal and Shaima Qureshi. 2022. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2515–2528.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2022. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. *arXiv preprint arXiv:2211.11187*.

Raviraj Joshi. 2022a. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Raviraj Joshi. 2022b. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 97.

Raviraj Joshi. 2022c. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages.

Hang Li and Zhengdong Lu. 2016. Deep learning for information retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1203–1206.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. Gottbert: a pure german language model.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *ACL 2017*, page 157.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, pages 197–209. Springer.

Weiting Tan and Philipp Koehn. 2022. Bitext mining for low-resource languages via contrastive learning.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. In *Artificial Neural Networks in Pattern Recognition: 10th IAPR TC3 Workshop, ANNPR 2022, Dubai, United Arab Emirates, November 24–26, 2022, Proceedings*, pages 121–128. Springer.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. 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 833–844.

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.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.