International Journal on Engineering Technology (InJET)

www.kec.edu.np/kec-journal/
Volume 3, issue no. 1, Dec 2025, Page 146 - 155
Received Date: Jul 31, 2025
Accepted Date: Sep 15, 2025
ISSN: 3021-940X (Print)
DOI:https://doi.org/10.3126/injet.v3i1.87019

Comparing LaBSE with Contrastively and Soft-Label Fine-Tuned mBERT Models for Semantic Search over a Nepali Knowledge Base

Dipesh Baral[1]

[1]*Thapathali Campus, Tribhuvan University, Kathmandu, Nepal*, dipesh.078bei011@tcioe.edu.np

Abstract

In this paper, the performance of multilingual sentence embedding models in semantic search for the Nepali language has been compared through three approaches: LaBSE under a zero-shot setting, mBERT fine-tuned through contrastive learning, and mBERT fine-tuned through soft similarity scores obtained by knowledge distillation from LaBSE. A customized dataset of approximately 800 labeled sentence pairs was developed from the e-commerce and appointment booking domains. The dataset contains questions written in Devanagari Nepali, with some code-mixed English. Each sentence pair was labeled as either semantically similar or dissimilar. Hard binary labels and a margin-based contrastive loss were utilized to train the contrastively trained model, while the distilled model was trained using a regression loss to match similarity scores obtained using LaBSE embeddings. All models were evaluated on a semantic retrieval task in which 89 user queries were embedded and compared with a corpus of 130 candidate sentences using cosine similarity. The quality of retrieval was calculated in terms of Top-1, Top-5, and Top-10 accuracy, and Mean Reciprocal Rank (MRR). LaBSE, without any task-specific fine-tuning, topped the results with Top-1 accuracy of 41.57% and MRR of 0.5246. The contrastively fine-tuned mBERT model achieved a Top-1 accuracy of 21.35% and MRR of 0.3204. The soft-label distilled mBERT model ranked mid-range with Top-1 accuracy of 34.83% and MRR of 0.4488, which shows that knowledge distillation can effectively transfer semantic similarity knowledge from LaBSE to mBERT. These findings demonstrate that while zero-shot LaBSE is strong, multilingual models like mBERT can be repurposed for Nepali semantic search through targeted fine-tuning. This research establishes a baseline for semantic search in Nepali and suggests practical approaches to enhancing sentence embeddings in low-resource language environments.

Keywords: Semantic Search, Sentence Embeddings, Contrastive Learning, Knowledge Distillation, mBERT, LaBSE

1. Introduction

Recent advances in multilingual natural language processing have significantly enhanced machines' ability to understand and represent the meaning of sentences across different languages. While powerful sentence-embedding models like LaBSE (Language-agnostic BERT Sentence Embedding) and mBERT have proven effective in many settings (Dakwale et al., 2022; Timilsina et al., 2022), applying them to Nepali, a language with limited NLP resources and frequent code-mixing with English, remains challenging. This work compares zero-shot LaBSE embeddings with mBERT models fine-tuned through contrastive learning and knowledge distillation, using a new dataset of semantically labeled Nepali sentence pairs from e-commerce and appointment booking domains. The evaluation measures retrieval performance via Top-k accuracy and Mean Reciprocal Rank, highlighting the promise and limitations of adapting multilingual models to low-resource languages such as Nepali.

Nepali semantic search is affected by the non-availability of annotated data, linguistic variety, and domain-specific knowledge resources, which make traditional keyword-based retrieval ineffective in determining true semantic similarity. User queries are often diverse in wording and involve code-mixing, therefore prompting retrieval systems to ignore good results even though it expresses similar intent. While large pre-trained models exist, their zero-shot accuracy will not automatically be optimal for Nepali domains without fine-tuning, which is rendered challenging due to the limited amount of labeled data. This effort looks into how the gap will be bridged by comparing zero-shot and fine-tuned approaches towards achieving better accuracy and ranking of semantic search in actual Nepali applications.

2. Related Work

The progression of semantic representations in natural language processing has advanced swiftly in the past ten years, beginning with the launch of Word2Vec (Mikolov et al., 2013), which permitted compact vector encodings that reflect linear semantic connections among words. This advancement was succeeded by sequence modeling frameworks like LSTMs and, notably, the Transformer model (Vaswani et al., 2017), which transformed contextual modeling by enabling the parallelizable and scalable learning of long-range dependencies. The introduction of contextual embeddings like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) signified a major shift in paradigms showing that deep bidirectional pretraining on extensive unlabeled datasets could produce more complex, task-independent linguistic representations than static embeddings.

Later models such as Sentence-BERT (Reimers & Gurevych, 2019) built upon this framework to produce semantically relevant sentence-level embeddings effectively, enabling large-scale semantic similarity and retrieval tasks to be computed feasibly. Leveraging multilingual pretraining, models like mBERT and XLM-R (Conneau et al., 2020) enabled cross-lingual transfer learning by training on datasets covering over 100 languages. LaBSE (Feng et al., 2022) enhanced this ability by implementing bilingual contrastive learning, creating a cohesive multilingual embedding space that exhibited impressive zero-shot retrieval results even for low-resource languages like Nepali.

Nonetheless, zero-shot models frequently exhibit lower performance when encountering language-specific syntactic peculiarities or code-mixed situations (Pires et al., 2019). To address these challenges, researchers have investigated domain-adaptive pretraining (Gururangan et al., 2020) and task-specific fine-tuning with contrastive and distillation-based goals. Artetxe et al. (2023) demonstrated that combining knowledge distillation from more powerful teacher models with contrastive objectives can greatly enhance multilingual sentence embeddings, especially in low-resource environments.

In the South Asian context, regional efforts have sought to enhance representation for Indic languages. IndicBERT (Kakwani et al., 2020) and MuRIL (Khanuja et al., 2021) are two large-scale multilingual models trained specifically on 17+ Indic languages, including Nepali, Hindi, and Bengali, capturing typological and script-level similarities across the region. For Nepali NLP, several studies have initiated localized model development and dataset creation. NepBERTa (Timilsina, Gautam, & Bhattarai, 2022) presented a monolingual Nepali transformer pretrained on Nepali news and Wikipedia corpora, demonstrating improved downstream performance on classification and NER tasks. Similarly, NepaliBERT (Pudasaini et al., 2023) and NepCovBERT (Khanal et al., 2021) showed the potential of domain-specific pretraining, particularly in COVID-related health text analysis and resource adaptation. Parallelly, community-driven efforts such as the NepaliNLU and NepaliGLUE benchmarks (Sharma et al., 2023) have begun providing standardized evaluation datasets for Nepali text understanding, though studies on semantic similarity and retrieval remain limited.

Although efforts have been made to broaden cross-lingual retrieval for Indic languages through multilingual embeddings (Kumar et al., 2022), Nepali is still less examined than its regional peers because of a scarcity of annotated corpora and code-mixed language varieties. This drives the necessity to explore if fine-tuned multilingual models can close this gap via transfer learning or distillation from high-resource teacher models like LaBSE.

3. Related Theory

Multilingual sentence embedding models have been used extensively for semantic search across low-resource languages. Among these, LaBSE has been used in this work as a zero-shot baseline. It was learned on parallel multilingual datasets using a dual-encoder strategy along with a translation ranking loss such that it can map semantically similar sentences of different languages to a shared vector space. This feature makes LaBSE usable for cross-lingual tasks even without fine-tuning.

To better handle domain-specific Nepali search terms, mBERT (Multilingual BERT) was trained with contrastive learning. It was given positive and negative pairs of sentences to a Siamese model, and a contrastive loss to pull close similar pairs and push distant dissimilar pairs in the embedding space was used.

This helped mBERT learn fine-grained semantic contrasts relevant to Nepali application domains like appointment booking and e-commerce.

In addition, a knowledge distillation approach was explored, where LaBSE served as a teacher to generate soft similarity scores (cosine similarities) among sentence pairs. These were used to supervise mBERT on a regression task with mean squared error loss. This allowed mBERT to learn a close approximation of LaBSE's similarity function without being too heavy and flexible to learn new Nepali domains.

4. Methodology

4.1. Dataset Preparation

To evaluate the semantic similarity capabilities of multilingual models in Nepali, two task-specific datasets were carefully constructed. These datasets were based on realistic queries from two practical domains: E-commerce and Appointment Booking. Each example consisted of a pair of Nepali or code-mixed (Nepali-English) user queries, accompanied by a similarity label indicating whether the two sentences conveyed the same underlying intent.

For the contrastive learning setup, each sentence pair was annotated with a binary label “1” for semantically similar (positive) and “0” for dissimilar (negative) and stored in the following JSON format:

```
{
  "sentence1": "User query 1",
  "sentence2": "User query 2",
  "label": 0 or 1
}
```

In addition to this, a soft-labeled variant of the dataset was created for knowledge distillation. In this version, cosine similarity scores between sentence embeddings were computed using LaBSE, and a floating-point soft_label field (ranging from 0 to 1) was added to each entry:

```
"soft_label": 0.84
```

Altogether, the dataset comprised approximately 800 sentence pairs, manually crafted by native Nepali speakers. Efforts were made to include diverse phrasings, informal expressions, and realistic variations in spelling and syntax reflecting how real users speak and type. A wide range of intents was covered across the two domains to simulate actual user interactions. This dataset served as the foundation for both contrastive fine-tuning and teacher-student distillation experiments conducted in this study.

4.2. Dataset Preprocessing

Before training, all sentence pairs were lowercased and normalized to remove unnecessary whitespace and non-informative punctuation. As the dataset included a mix of Nepali and English words written in Devanagari script, care was taken to preserve linguistic structure during cleaning. Sentence pairs were aligned with either binary labels (for contrastive learning) or continuous similarity scores (for soft-label distillation).

Tokenization was performed using the AutoTokenizer associated with the mBERT model. To ensure uniform input size and computational efficiency, all sequences were truncated or padded to a maximum length of 64 tokens. For the soft-label distillation setup, embeddings were precomputed using LaBSE, and cosine similarities were calculated to serve as supervision targets.

The dataset was then split into training and validation sets in an 80:20 ratio, ensuring intent and domain diversity in both splits. The training set was used for model fitting, and the validation set was used for hyperparameter tuning and early stopping. For final evaluation of semantic retrieval performance, including Top-k accuracy and Mean Reciprocal Rank (MRR), a separate held-out test set that was not seen during training or validation was used. All preprocessing steps were implemented using PyTorch and HuggingFace Transformers, with careful logging to ensure reproducibility.

4.3. Model Training

Two distinct training strategies were employed to adapt multilingual models for semantic similarity in Nepali: contrastive fine-tuning using hard binary labels and knowledge distillation using soft similarity labels derived from LaBSE.

4.3.1. Contrastive Fine-tuning with Sentence Transformers

In the contrastive fine-tuning setup, a Siamese network architecture was adopted with a shared mBERT encoder. Given a batch of sentence pairs $\{(x_i,y_i)\}^{N}_{i=1}$, where x_i and y_i represent semantically related sentences, the cosine similarity S (x_i, y_i) between their embeddings was computed:

$$S(x_i, y_i) = \frac{(E(x_i)\cdot E(y_i))}{(\|E(x_i)\| \times \|E(y_i)\|)} \qquad \text{(Equation 1)}$$

where E (·) denotes the embedding function produced by the encoder.

The Contrastive Loss function, as implemented in the Sentence Transformers library, encourages the similarity between positive pairs to be high and that of negative pairs to be low. Formally, the loss for a single pair is defined as:

$$L = \frac{1}{N}\sum_{i=1}^{N}\left[(1-l_i)\times S(x_i,y_i)^2 + l_i \times max\left(0, m - S(x_i,y_i)\right)^2\right] \qquad \text{(Equation 2)}$$

Where:

- $l_i \in \{0,1\}$ is the label indicating whether the pair is similar (0) or dissimilar (1),
- m is the margin hyperparameter that defines the minimum distance between dissimilar pairs,
- N is the batch size.

4.3.2. Knowledge Distillation Using Soft Labels

For knowledge distillation, soft similarity labels were generated by computing the cosine similarity between sentence embeddings obtained from the LaBSE teacher model**:**

$$s_i = \ cosinesim\left(E_{LaBSE(x_i)}, E_{LaBSE(y_i)}\right) \in \ [0,1] \qquad \text{(Equation 3)}$$

The mBERT student model was then trained to regress these soft labels using Mean Squared Error (MSE) loss between its predicted similarity and the LaBSE-generated soft labels:

$$L \ = \ \left(\frac{1}{N}\right)\sum_{i=1}^{N}(\hat{s}_i \ - \ s_i)^2 \qquad \text{(Equation 4)}$$

where:

- $\hat{s}_i$ is the predicted similarity score from the student,
- s_i is the soft label from LaBSE,
- N is the batch size.

This loss encourages the student model to mimic the teacher’s semantic similarity output, enabling it to learn richer embeddings.

4.3.3. Model Training Details

A. Contrastively Fine-Tuned mBERT

- Optimizer: AdamW
- Batch Size: 16
- Learning rate: 2e-5
- Epochs: 5
- Loss Function: Contrastive Loss from the Sentence Transformers library was used to separate similar and dissimilar pairs in embedding space

- Evaluation: A Binary Classification Evaluator was used on a held-out test set to track performance during training

B. Knowledge-Distilled mBERT (Soft Labels from LaBSE)

- Optimizer: AdamW
- Batch Size: 16
- Learning rate: 2e-5
- Epochs: 4
- Loss Function: Mean Squared Error (MSE) Loss was used to regress cosine similarity values predicted by LaBSE
- Soft Labels: Sentence similarity scores (range: 0 to 1) were generated by computing cosine similarity between LaBSE embeddings

5. Result and Analysis

5.1. Results

The training behavior of the contrastively fine-tuned mBERT model is shown in Figures 1–6. Figure 1 and 2 illustrates the learning curves for cosine accuracy and F1 score on the validation dataset respectively. Both measurements rise consistently over the initial 120 steps and then level off, suggesting stable convergence and verifying that the model is not underfitting or overfitting. Figure 3 and 4 displays precision and recall metrics over training steps, emphasizing the anticipated trade-off throughout optimization. Although precision begins to drop as recall rises, both metrics reach stability after step 120, indicating a well-balance model. Lastly, Figure 5 and 6 illustrates the ROC and Precision-Recall curves respectively. The model attains an AUC of 0.92 and an average precision of 0.86, indicating robust discriminative effectiveness for retrieving semantic sentence similarity. Together, these curves confirm the training stability, convergence, and overall effectiveness of the model in low-resource, code-mixed Nepali-English contexts.

To further improve performance, mBERT was trained using a knowledge distillation approach. In the knowledge distillation stage, the student mBERT model was refined with Mean Squared Error (MSE) loss, utilizing cosine similarity scores generated by LaBSE as regression targets. Figure 7 illustrates that the training loss steadily declined over four epochs (from 0.0150 to 0.0036), reflecting reliable convergence and successful knowledge transfer.

Validation loss was not documented at every epoch because the distillation objective required teacher guidance instead of conventional validation-focused adjustments. Nonetheless, evaluation after training on a completely different test set (utilized for Top-K and MRR assessments) demonstrated consistent generalization, verifying that the model was neither overfitted nor underfitted. The training concluded after four epochs since the loss stabilized, indicating that additional epochs would likely not provide substantial enhancement.

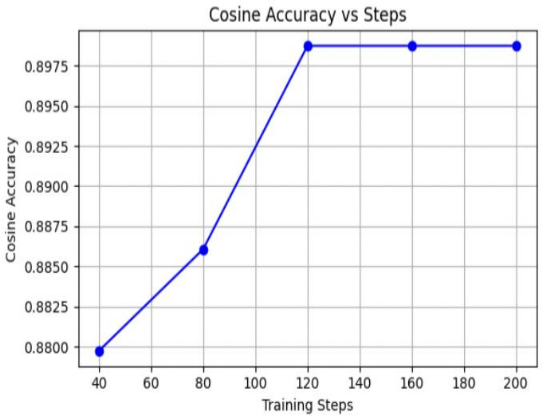

Figure 1. Cosine accuracy on the validation set across training steps for the contrastively fine-tuned mBERT model. Accuracy increases initially and plateaus after step 120, indicating stable convergence

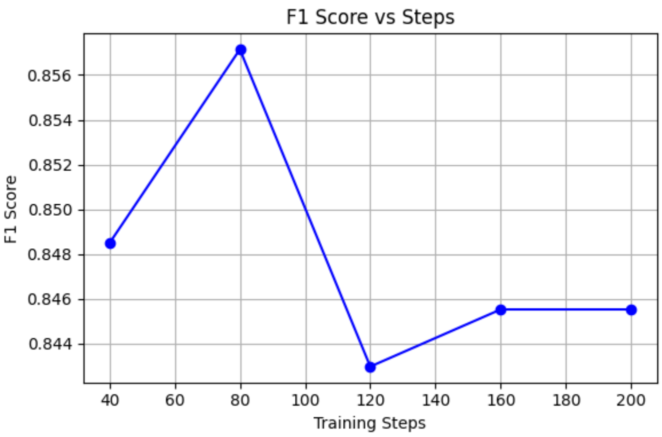

Figure 2. Precision-Recall curve of the model, achieving an Average Precision (AP) of 0.86.

Error! No text of specified style in document.

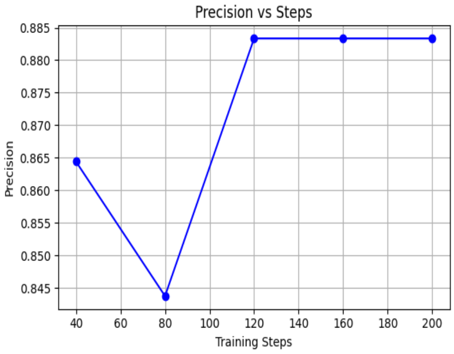

Figure 3. Precision on the validation set across training steps. Initial fluctuations reflect the trade-off with recall, stabilizing after step 120 as the model converges.

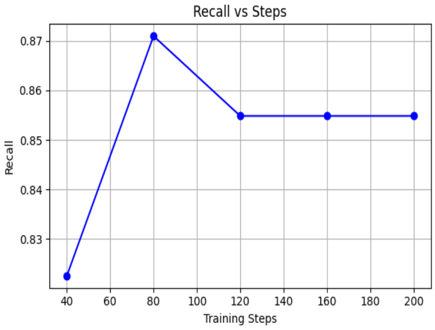

Figure 4. Recall on the validation set across training steps. The metric initially increases while precision fluctuates, demonstrating the expected trade-off, and plateaus after step 120

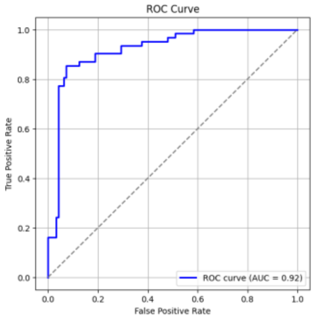

Figure 5. ROC curve evaluated on the validation set for the contrastively fine-tuned mBERT model. The curve demonstrates strong discriminative ability with an AUC of 0.92.

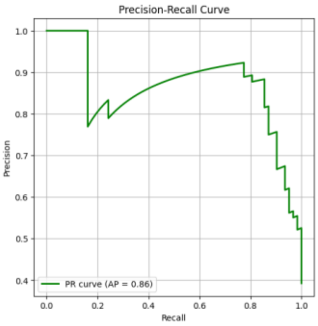

Figure 6. Precision-Recall curve evaluated on the validation set. The model achieves an average precision of 0.86, indicating effective classification of semantically similar and dissimilar sentences

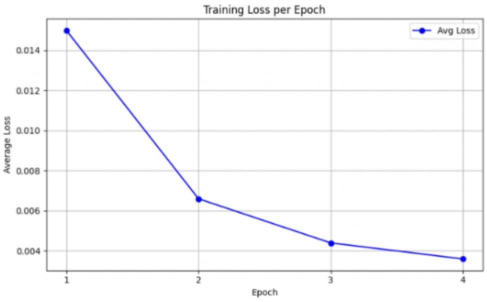

Figure 7. MSE loss curve observed during the training of the mBERT model using knowledge distillation from LaBSE. A consistent reduction in loss per epoch reflects stable convergence.

A comparative evaluation of the models was conducted based on Top-1, Top-5, and Top-10 accuracy, as well as Mean Reciprocal Rank (MRR), to assess their effectiveness in the semantic retrieval task. The results are summarized in table below.

Table 1. Comparison of model performance in the semantic retrieval task using Top-K Accuracy and Mean Reciprocal Rank (MRR).

Models	**Top-1 Accuracy**	**Top-5 Accuracy**	**Top-10 accuracy**	**MRR**
LaBSE (Zero-Shot)	0.4157	0.6067	0.7303	0.5246
mBERT (Knowledge Distilled)	0.3483	0.5618	0.6292	0.4488
mBERT (Contrastive Learning)	0.2135	0.4270	0.4719	0.3204

5.2. Analysis

The experimental findings indicate that LaBSE (zero-shot) attained the best retrieval performance, recording a Top-1 Accuracy of 0.4157 and an MRR of 0.5246. In contrast, the mBERT model fine-tuned via knowledge distillation achieved a Top-1 Accuracy of 0.3483 and an MRR of 0.4488, whereas the contrastively fine-tuned mBERT exhibited the poorest retrieval results (Top-1 = 0.2135, MRR = 0.3204). While the contrastive model showed robust pairwise discrimination on validation data, it did not generalize well to real-world retrieval scenarios. The performance ranking noted (LaBSE > Distilled mBERT > Contrastive mBERT) signifies essential distinctions in pretraining goals, supervision cues, and data attributes.

5.2.1. Sentence-level pretraining versus token-level pretraining

LaBSE's enhanced retrieval capability primarily arises from its pretraining objective at the sentence level, which aligns semantically similar sentences across more than 100 languages through a translation ranking task (Feng et al., 2019). This method directly enhances the structure of the embedding space for cross-lingual similarity and dense retrieval, allowing LaBSE to accurately represent Nepali and code-mixed inputs even in the absence of task-specific fine-tuning.

In contrast, mBERT (Devlin et al., 2018) underwent pretraining via masked language modeling (MLM) and next sentence prediction (NSP) on multilingual Wikipedia. These goals function mainly at the token or local context level, instead of enhancing global sentence embeddings. As a result, although mBERT offers multilingual contextual representations, it is not specifically intended to generate consistent distances or rankings at the sentence level. Previous research has also determined that mBERT embeddings exhibit reduced isotropy and diminished alignment between languages compared to embeddings generated by models specifically designed for sentence similarity (Reimers & Gurevych, 2019).

5.2.2. Effect of fine-tuning objectives

The mBERT that underwent knowledge distillation gained from “soft” supervision, utilizing LaBSE cosine similarities as continuous regression objectives. These soft labels reflect subtle levels of semantic similarity, offering more nuanced gradients than simple binary contrastive labels. This method enabled mBERT to mimic the smooth configuration of LaBSE’s embedding space, resulting in retrieval-optimized representations.

Conversely, the contrastively fine-tuned mBERT was developed to distinguish positive from negative pairs by means of relative distance learning. Although contrastive learning is theoretically strong, its effectiveness relies significantly on the negative sampling approach, batch size, and scale of data (Karpukhin et al., 2020). In this research, the small dataset size (~800 sentence pairs) and the lack of systematically generated hard negatives probably led the model to overfit to pairwise separability, not establishing a well-organized global ranking space. This clarifies why the contrastive model obtained high ROC and PR metrics yet displayed low retrieval accuracy.

5.2.3. Data scale, negative diversity and domain coverage

Contrastive frameworks excel with extensive, varied datasets containing many difficult negatives in each batch. In the absence of these, the model is likely to create disjointed clusters instead of cohesive, semantically structured manifolds. Moreover, the Nepali dataset utilized in this context, while of high quality, is somewhat limited in size and specific to a particular domain. In contrast, LaBSE was pretrained using billions of monolingual and parallel sentences from diverse domains, which provided it with resilience against linguistic noise, code-mixing, and regional differences.

This disparity in data intensifies the performance difference. The lack of domain-adaptive pretraining for mBERT further restricts its ability to process mixed Devanagari and Romanized tokens, which is a frequent characteristic of Nepali digital text. Numerous studies have demonstrated that ongoing masked language modeling, referred to as domain-adaptive pretraining, on in-domain datasets greatly enhances downstream performance for low-resource languages (Gururangan et al., 2020).

5.2.4. Evaluation perspective: classification versus ranking

An additional source of discrepancy arises from the evaluation paradigm. Pairwise similarity classification solely determines if two sentences are alike or not alike. Dense retrieval, nevertheless, demands consistent global ranking among numerous candidates. A model may exhibit strong binary discrimination (AUC, PR) yet assign inconsistent relative distances, resulting in poor Top-k and MRR metrics. This difference emphasizes that contrastive goals refined for pairwise distinction do not consistently lead to optimal ranking performance.

5.2.5. Practical factors and embedding calibration

Various factors related to implementation also affect retrieval results. The selection of pooling strategy, normalization of embeddings, and scaling of temperature can significantly influence cosine similarities. LaBSE employs precisely adjusted normalization and margin-based softmax to guarantee isotropic distributions of embeddings. Replicating this geometry using a smaller fine-tuning dataset is complex, which accounts for why the distilled mBERT despite being trained on LaBSE scores still falls short of the teacher model.

6. Conclusion

This research investigated semantic search within a Nepali knowledge base utilizing multilingual transformer-based models. Two refined versions of mBERT were created: one that employed contrastive learning with handpicked sentence pairs, and another that was distilled from LaBSE utilizing soft similarity scores as regression targets. The zero-shot LaBSE model was assessed as a baseline to gauge cross-lingual generalization without the need for task-specific guidance.

Experimental findings showed that although the contrastively trained mBERT delivered encouraging differentiation in similarity classification tasks, clear from the ROC and Precision-Recall curves, it lagged in practical retrieval scenarios. In comparison, the knowledge-distilled mBERT showed enhanced retrieval metrics, indicating that soft supervision from LaBSE embeddings offered a more robust and seamless learning signal compared to binary contrastive labels. The best retrieval performance, however, was attained by LaBSE in its zero-shot mode, highlighting the potency of extensive multilingual sentence embedding models for grasping semantic connections, even in low-resource languages such as Nepali.

This study emphasizes the possibilities and constraints of transformer-based semantic search for languages that are less represented. Future efforts can focus on improving performance by increasing the dataset size and linguistic variety, employing effective hard-negative mining techniques, and testing hybrid training goals that merge contrastive and regression losses. Additional enhancements could emerge from ongoing pretraining of multilingual models on expanded Nepali datasets, along with domain-specific adjustments for uses like conversational agents, information retrieval, or knowledge-based dialogue systems.

References

Dakwale, P., Chaudhary, V., Sitaram, S., Gupta, A., & Sarawagi, S. (2022). Empirical evaluation of language agnostic filtering of parallel data for low resource languages. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation (PACLIC 36)*. pp. 417–426. Association for Computational Linguistics Available at: https://aclanthology.org/2022.paclic-1.38 [Accessed 20 Nov. 2025].

Timilsina, S., Gautam, M., & Bhattarai, B. (2022). NepBERTa: Nepali language model trained in a large corpus. 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 (AACL-IJCNLP 2022)*, pp. 273-284. Available at: https://aclanthology.org/2022.aacl-short.34 [Accessed 20 Nov. 2025].

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. Available at: https://arxiv.org/abs/1301.3781

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, pp. 5998–6008. https://arxiv.org/abs/1706.03762

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*. https://arxiv.org/abs/1802.05365

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*. https://arxiv.org/abs/1908.10084

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Feng, F., Jiang, P., Kiela, D., Latini, A., Nishida, K., Bose, A., Guo, D., Yang, D., & Zhou, J. (2022). Language-agnostic BERT sentence embedding (LaBSE). *arXiv preprint arXiv:2209.08459*. https://arxiv.org/abs/2209.08459

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 4996–5001. https://doi.org/10.18653/v1/P19-1493

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8342–8360.

Artetxe, M., Schwenk, H., Sagot, B., & Agirre, E. (2023). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics, 11*, pp. 253–270. https://aclanthology.org/2023.tacl-1.21.pdf

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicBERT: A pre-trained multilingual transformer for Indian languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 6769–6781. https://aclanthology.org/2020.emnlp-main.445

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., Gupta, S., Gali, S. C. B., & Subramanian, V. (2021). MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*. https://arxiv.org/abs/2103.10730

Pudasaini, S., Shakya, S., Tamang, A., Adhikari, S., Thapa, S., & Lamichhane, S. (2023). NepaliBERT: Pre-training of masked language model in Nepali corpus. *Proceedings of the 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) 2023*, pp. 325–330. https://doi.org/10.1109/I-SMAC58438.2023.10290690

Khanal, N. P., & Sharma, R. (2021). NepCovBERT: A domain-adapted transformer model for Nepali COVID-19 text. *Proceedings of the 2021 Conference on Asian Language Processing (AACL 2021)*. https://aclanthology.org/2021.aacl-short.56

Sharma, B. R., & Thapa, P. (2023). NepaliGLUE: A benchmark for Nepali natural language understanding. *Proceedings of the 2023 Conference on Asian Language Processing (AACL 2023)*. https://aclanthology.org/2023.aacl-short.67

Kumar, P., Kunchukuttan, A., & Khapra, M. M. (2022). IndicNLG Suite: Datasets and evaluation benchmarks for natural language generation in Indian languages. *Proceedings of the 2022 Conference on Natural Language Generation (INLG 2022)*. https://aclanthology.org/2022.inlg-1.2

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781