Improving BGE-M3 Multilingual Dense Embeddings for Nigerian Low Resource Languages

Abdulmatin Omotoso[1]*, Habeeb Shopeju[1]*, Adejumobi Joshua[1]*, and Shiloh Oni[1]
[1]Machine Learning Collective

Abstract

Multilingual dense embedding models such as Multilingual E5, LaBSE, and BGE-M3 have shown promising results on diverse benchmarks for information retrieval in low-resource languages. But their result on low resource languages is not up to par with other high resource languages. This work improves the performance of BGE-M3 through contrastive fine-tuning; the model was selected because of its superior performance over other multilingual embedding models across MIRACL, MTEB, and SEB benchmarks. To fine-tune this model, we curated a comprehensive dataset comprising Yorùbá (32.9k rows), Igbo (18k rows) and Hausa (85k rows) from mainly news sources. We further augmented our multilingual dataset with English queries and mapped it to each of the Yoruba, Igbo, and Hausa documents, enabling cross-lingual semantic training. We evaluate on two settings: the Wura test set and the MIRACL benchmark. On Wura, the fine-tuned BGE-M3 raises mean reciprocal rank (MRR) to 0.9201 for Yorùbá, 0.8638 for Igbo, 0.9230 for Hausa, and 0.8617 for English queries matched to local documents, surpassing the BGE-M3 baselines of 0.7846, 0.7566, 0.8575, and 0.7377, respectively. On MIRACL (Yorùbá subset), the fine-tuned model attains 0.5996 MRR, slightly surpassing base BGE-M3 (0.5952) and outperforming ML-E5-large (0.5632) and LaBSE (0.4468).

1 Introduction

Nigeria is home to hundreds of languages, yet its three major tongues: Hausa, Yorùbá, and Igbo—are still considered low-resource for information retrieval (IR) tasks. These languages are morphologically rich and linguistically complex, featuring phenomena such as agglutinative affixes and, in the case of Yorùbá and Igbo, tonal diacritics that alter word meaning. A key challenge is that text in these languages often lacks standardized orthography (e.g., inconsistent use of Yorùbá tone marks), making it difficult for conventional IR systems to properly match queries with documents. Despite being spoken by tens of millions, Yorùbá, Igbo, and Hausa have relatively scarce digital corpora and limited NLP applications, which exacerbates the IR problem in these languages. The result is a significant vocabulary mismatch issue: users' queries may not lexically match relevant documents due to inflectional variations, compounding, or spelling inconsistencies, leading to poor recall in retrieval (Mitra and Craswell, 2017).

Traditional lexical retrieval methods (e.g., BM25 or tf–idf ranking) are insufficient for these low-resource, morphologically rich languages. Lexical IR relies on exact or near-exact token overlap between query and document, an assumption that breaks down when words have many surface forms or when spelling variations (such as omitted diacritics) are common. Consequently, purely lexical approaches struggle to retrieve semantically relevant content if there is no literal token match. This limitation is well-documented as the semantic and vocabulary mismatch problem. For example, a Yorùbá user might search for "àwòrán" (meaning "picture"), but a document containing the synonym "fotò" (a borrowed word for "photo") would be missed by lexical matching.

Recent advancements in neural IR show promising solutions by introducing dense multilingual embedding models, such as LaBSE (Feng et al., 2022), mE5 (Wang et al., 2024), and BGE-M3 (Chen et al., 2024). These models encode queries and documents into a shared vector space, enabling semantic matching beyond lexical similarity (Karpukhin et al., 2020; Feng and Pengcheng, 2020). Despite their effectiveness, general multilingual models do not obtain a very high performance for low-resource languages such as Yorùbá, Igbo, and Hausa, as opposed to English. (Alabi et al.,

*Equal contribution.

Proceedings of the 9th Widening NLP Workshop, pages 224–229

2020).

Fine-tuning multilingual models on targeted datasets has emerged as a promising strategy for improving retrieval performance on low-resource languages. More recently, the MIRACL dataset (18 languages)(Zhang et al., 2023) was used to fine-tune retrieval models, and a single model trained on all languages achieved robust performance, even outperforming some monolingual-tuned models on their own language (Chen et al., 2024). Recognized for its state-of-the-art performance across multilingual retrieval benchmarks such as MIRACL and SEB, we decided to fine-tune the BGE-M3 model (Chen et al., 2024), as it offers substantial potential for improvement through contrastive fine-tuning. A technique that encourages the embedding model to minimize distances between semantically similar document-query pairs and maximize distances for dissimilar pairs (Schroff et al., 2015; Ukarapol et al., 2024; Zhou et al., 2023). Our contributions are as follows:

i. We curated high-quality datasets for each target Nigerian language from trusted sources such as BBC Yoruba and Igbo, VON, Aláròyé, and other news sources.

ii. We fine-tuned BGE-M3 on the curated dataset using contrastive learning.

iii. We compared the fine-tuned model with the BGE-M3 baseline and other embedding models such as LaBSE, Multilingual E5, and OpenAI-text-embedding-3-large, utilizing a hold-out portion of the Wura test set.

iv. We release all data, code and weights used for our work.[1] [2]

2 Methodology

2.1 Dataset Extraction

We created a multilingual dataset of 115k query–document pairs in Yoruba, Igbo, and Hausa, plus synthetic English queries for cross-lingual training. The Yoruba set has about 32.9k pairs, mostly from **Aláròyé** (10k), **VON Yoruba** (6.5k), **BBC Yoruba** (1k), and the **Wura** dataset. The Igbo set has about 18k pairs from **Wura**, **VON Igbo**, and **BBC Igbo**. Hausa is the largest, with

Figure 1: Data Extraction

85k pairs from sources like **Premium Times Hausa**, **Fim Magazine**, **VOA Hausa, Katsina Post, Legit Hausa, Amaniya, and VON Hausa**. Queries were taken from headlines or sub-topics, with the matching article content as the positive document. We added English-translated queries using the Gemma3-27B model to support multilingual retrieval. About 15k Yoruba, 15k Igbo, and 15k Hausa queries (45k total) were translated and paired with their original-language documents, creating English–Yoruba, English–Igbo, and English–Hausa pairs for alignment.

2.2 Preprocessing and Cleaning

All datasets were preprocessed with **trafilatura** to strip boilerplate, ads, and navigation elements, then cleaned with **datatrove** for filtering and deduplication to ensure high quality and consistency (Chen et al., 2022). The **Wura dataset** needed extra cleaning to ensure consistency and avoid overlap. For entries from Wikipedia, we removed sentences where the query appeared at the start of a line to prevent leakage. We deleted duplicates by URL and excluded items whose source URLs overlapped with our scraped news datasets. We discarded all **jw.org** entries, which often contained duplicate pages, mismatched titles, or malformed text. To keep training and evaluation separate, we removed from training any Wura pairs that were already in its validation split. After this, we ran a general quality audit across all languages. Using **Gemma3-27B**, we flagged and removed passages that were not natural-language content (e.g., boilerplate, poorly formatted, or uninformative text). Finally, we applied length filters, discarding documents with fewer than five words, or fewer than 30 words when the query appeared at the start.

[1] https://github.com/HAKSOAT/wazobia-embed
[2] https://huggingface.co/abdulmatinomotoso/bge-finetuned

Embedding Model	Yoruba	Igbo	Hausa	English	Macro Avg.
ML-E5-large (Baseline)	0.6766	0.6795	0.6992	0.3526	0.6020
BGE-M3 (Baseline)	0.7846	0.7566	0.8575	0.7377	0.7841
LaBSE (Baseline)	0.3201	0.3001	0.3188	0.4349	0.3435
BGE-M3 (Fine-Tuned – Combined)	**0.9201**	**0.8638**	**0.9230**	**0.8617**	**0.8922**

Table 1: MRR and macro-average MRR of embedding models on Wura test sets.

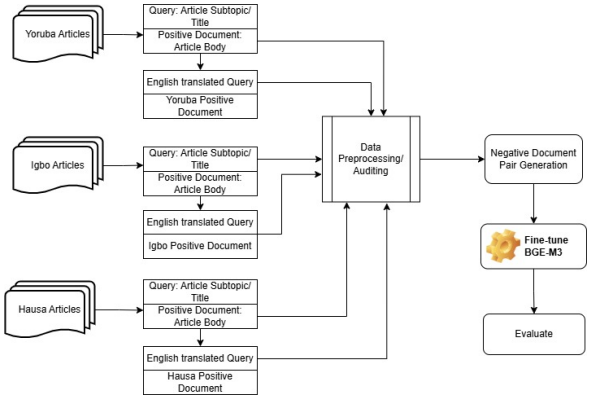

Figure 2: Methodology

2.3 Negative Pair Generation

For effective contrastive learning in retrieval, each training query is paired not only with its relevant document (positive example) but also with one or more irrelevant documents (negative examples). We built a pool of up to 7 unrelated passages per query, obtained through random sampling from other queries' documents.

2.4 Fine-Tuning Procedure

We fine-tuned BGE-M3 on the combined languages data (Yoruba, Igbo, Hausa, and the English-translated queries) using contrastive fine-tuning method (query-positive-negative triplet) (1) with self-distillation disabled and without unifying dense, sparse, and multi-vector retrieval, as we were focused on fine-tuning only the dense embedding aspect of BGE-M3. The configuration used is shown in Table 4

$$
\begin{aligned}
s_{qp} &= \exp\big(\mathrm{sim}(q,p)/\tau\big), \\
s_{qn_i} &= \exp\big(\mathrm{sim}(q,n_i)/\tau\big), \\
L\big(q,p,\{n_i\}_{i=1}^{N}\big) &= -\log \frac{s_{qp}}{s_{qp}+\sum_{i=1}^{N} s_{qn_i}}.
\end{aligned}
\tag{1}
$$

Also turning on this parameter, at our early experiment phase did not improve the performance of the model. Each training step sampled a query from any of the languages, along with its corresponding positive document and one negative document (randomly drawn from that query's negative pool as described). We found that using a single positive and a single negative per query in each step was sufficient to learn effectively. This simple one-to-one (positive-to-negative) ratio, combined with the rotation of negatives across epochs, yielded the best validation performance.

We also explored alternative fine-tuning strategies, but these proved less effective. In one of such strategies we experimented with increasing the number of negatives per query, using one positive paired with two simultaneous negatives. This approach led to a significantly worse retrieval accuracy, potentially due to overly challenging or noisy training signals when multiple negatives were introduced at once. We did not pursue the cause of this further, neither increasing the negatives nor training for longer. On increasing the number of negatives, training became time intensive, where the use of two negatives took 15 hours compared to 6 hours for one negative. These specific experiments were done on a Google Colab A100 machine. Second, we attempted sequential fine-tuning across languages—for example, starting with a model fine-tuned on Yoruba data, then further fine-tuning that model on Igbo or Hausa data. This sequential transfer approach resulted in a degradation of performance on the initially trained language Table 3; a behaviour explained by catastrophic forgetting (van de Ven et al., 2024). Thus, switching a model's focus to a new language corpus tended to undermine the representations learned for the original language. In contrast, the combined multilingual training from a common initialization preserved balanced performance across languages, so we adopted that as our primary fine-tuning method.

3 Results

For evaluation, we utilized the held-out portions of the Wura dataset as our primary benchmark for all three languages. The Wura dataset contains annotated query–document pairs in Yoruba, Igbo,

Embedding Model	MRR (Yorùbá)
ML-E5-large (Baseline)	0.5632
BGE-M3 (Baseline)	0.5952
LaBSE (Baseline)	0.4468
BGE-M3 (Fine-Tuned – Combined)	**0.5996**

Table 2: MIRACL benchmark (Yorùbá subset).

Embedding Model	Yoruba	Igbo	Hausa	Macro Avg.
ML-E5-large(Baseline)	0.663341	0.760283	0.752902	0.725508
BGE-M3(Baseline)	0.823499	0.850487	0.881689	0.851892
LaBSE(Baseline)	0.346926	0.489230	0.323469	0.386542
BGE-M3-yoruba-alldata-Epochs-3	**0.937361**	0.904532	0.912439	0.918111
BGE-M3-yoruba-igbo-alldata-Epochs-3	**0.930475**	0.932700	0.913530	0.925568
BGE-M3-yoruba-igbo-hausa-alldata-Epochs-3	**0.911866**	0.904386	0.930070	0.915441

Table 3: MRR and macro-average MRR of embedding models on Wura test set using the sequential transfer approach. Only the Yorùbá column is bolded for fine-tuned variants.

and Hausa, making it well-suited for evaluating our multilingual retriever in each language. We partitioned Wura's data into validation and test splits to tune the model and assess final performance. Approximately 60% of the Wura queries (up to a maximum of 2,000 per language) were set aside as a validation set for development and hyperparameter tuning. The remaining 40% of the queries (again up to 2,000 per language) was reserved as the final test set on which we report results. Importantly, these evaluation queries were never seen during training (as ensured by the preprocessing step that removed Wura validation examples from the training data). In addition to Wura, we also evaluated on the yoruba subset of MIRACL (Zhang et al., 2023) a widely used multilingual retrieval benchmark that provides monolingual ad-hoc retrieval tasks over Wikipedia across 18 languages with hundreds of thousands of high-quality relevance judgments—following its standard development/test protocol to cross-check robustness. We evaluate retrieval performance primarily using Mean Reciprocal Rank (MRR) (2), which measure the model's ability to successfully retrieve the correct document for each query in the test set.

$$\text{MRR} = \frac{1}{|Q|} \sum\_i = 1^{|Q|} \frac{1}{\text{rank}_i} \qquad (2)$$

On the Wura test set, Table 1, the fine-tuned BGE-M3 model consistently achieved superior results across all languages evaluated. Specifically, for same-language query–document pairs, the fine-tuned model achieved mean reciprocal rank (MRR) scores of 0.9201 for Yorùbá, 0.8638 for Igbo, and 0.9230 for Hausa; for English-to-(Yorùbá/Igbo/Hausa) cross-lingual queries, the model obtained 0.8617, clearly surpassing all baseline embedding models. In addition, on the MIRACL benchmark (Zhang et al., 2023) Table 2, our fine-tuned BGE-M3 achieved **0.5996** MRR on the Yorùbá subset, slightly outperforming base BGE-M3 (0.5952) and substantially exceeding LaBSE (0.4468).

4 Conclusion

This study has shown that fine-tuning multilingual embedding models, particularly BGE-M3, can significantly improve information retrieval performance for low-resource Nigerian languages such as Yoruba, Igbo, and Hausa. Through contrastive learning and cross-lingual alignment using English translated queries mapped to one of Yoruba, Igbo and Hausa documents, the fine-tuned models achieved a results and outperformed established baselines. Our findings emphasize that low-resource languages can benefit greatly from recent advances in large-scale multilingual embeddings when appropriately adapted. The outcomes also reinforce the potential for building inclusive, language aware IR systems that serve diverse linguistic communities.

5 Limitations

While the fine-tuned model shows strong MRR across all languages, we conducted a brief manual review of retrieval errors. Common failure cases included queries with ambiguous meaning or requiring contextual inference beyond sentence-level similarity. For instance, some Yoruba queries containing idiomatic expressions were mismatched

with overly literal documents. These findings suggest room for improvement via domain-specific tuning or the inclusion of richer context during training. The English queries are synthetic data as they were generated using the **Gemma3-27B** model. Efforts were made to manually review a handful of those queries, but this does not scale to 45k queries. Hence, the queries may be of lesser quality than human-written queries and therefore the model may not generalize properly.

6 Ethical Considerations

We manually inspected all news source websites for terms of use, paywalls, or copyright notices and found none; only Legit.ng Hausa published a robots.txt file, which we fully respected. Our dataset included only newsroom content and contained names of public figures as part of standard reporting, but no user comments or private data. All data was used strictly for research purposes, with copyright remaining with the original publishers. We released only short text snippets and article metadata under a research-only license, in accordance with the rights of the original publishers.

References

Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire, David I. Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

R. Chen, Y. Zhao, D. Wang, and 1 others. 2022. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2203.00505.

Feng and Pengcheng. 2020. Labse: Language-agnostic bert sentence embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Bhaskar Mitra and Nick Craswell. 2017. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Trapoom Ukarapol, Zhicheng Lee, and Amy Xin. 2024. Improving text embeddings for smaller language models using contrastive fine-tuning. *arXiv preprint arXiv:2408.00690*.

Gido M van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. 2024. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Wenxuan Zhou, Sheng Zhang, Tristan Naumann, Muhao Chen, and Hoifung Poon. 2023. Continual contrastive finetuning improves low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13249–13263, Toronto, Canada. Association for Computational Linguistics.

A Ablation: Effect of Dense-Only Fine-Tuning on BGE-M3 Sparse and Multi-vector Layers

Fine-tuning BGE-M3's dense layer significantly improves multi-vector retrieval performance across all languages (6-9% MRR gains for Yoruba/Igbo) but severely degrades sparse retrieval (85-94% MRR drops) Table 5. We evaluated at max-lengths 2048 and 8192 tokens; the dense results at 8192 tokens are used in the main paper.

Table 4: Training command and key parameters.

```
torchrun --standalone --nproc_per_node 8 \
-m FlagEmbedding.finetune.embedder.encoder_only.m3 \
--model_name_or_path BAAI/bge-m3 \
--output_dir ./bge-m3 \
--cache_dir ./cache/model \
--cache_path ./cache/data \
--train_data ./filtered_combine_train_dataset.jsonl \
--trust_remote_code True \
--train_group_size 2 \
--query_max_len 512 \
--passage_max_len 2048 \
--overwrite_output_dir \
--learning_rate 1e-5 \
--fp16 \
--dataloader_num_workers 12 \
--gradient_checkpointing \
--deepspeed ds_stage0.json \
--num_train_epochs 3 \
--per_device_train_batch_size 16 \
--dataloader_drop_last False \
--warmup_ratio 0.1 \
--report_to none \
--logging_steps 100 \
--save_steps 500 \
--temperature 0.01 \
--sentence_pooling_method cls \
--normalize_embeddings True \
--knowledge_distillation False \
--kd_loss_type m3_kd_loss \
--unified_finetuning False \
--use_self_distill False \
--fix_encoder False
```

Embedding Type	**Model**	**Yoruba**	**Igbo**	**Hausa**	**English**
Max Length: 2048 tokens					
Sparse	Baseline	**0.697**	**0.751**	**0.233**	**0.044**
Sparse	Fine-Tuned	0.048	0.080	0.022	0.004
Multi-vector (FP16)	Baseline	0.835	0.814	0.254	0.051
Multi-vector (FP16)	Fine-Tuned	**0.906**	**0.832**	**0.259**	**0.061**
Max Length: 8192 tokens (used in main results)					
Sparse	Baseline	**0.671**	**0.727**	**0.229**	**0.043**
Sparse	Fine-Tuned	0.046	0.076	0.020	0.004
Multi-vector (FP16)	Baseline	0.831	0.813	0.255	0.050
Multi-vector (FP16)	Fine-Tuned	**0.908**	**0.830**	**0.260**	**0.061**

Table 5: Impact of dense-only fine-tuning on BGE-M3 retrieval layers. MRR scores across embedding types, max-length settings, and Nigerian languages. Bold indicates best performance per language within each configuration.