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

Keywords: Multilingual Dense Retrieval, BGE-M3, Fine-tuning, Yoruba, Igbo, Hausa, Low-resource Languages, Information Retrieval.

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. To further improve their performance, fine-tuning is encouraged.

In this work, we improve the performance of BGE-M3 through contrastive fine-tuning, we chose this model because it has demonstrated superior performance to other multilingual embedding models across diverse retrieval benchmarks such as the MIRACL information retrieval dataset, the Massive Text Embedding Benchmark (MTEB), and the Scandinavian Embedding Benchmark (SEB). To achieve this, we curated a comprehensive dataset comprising of Yorùbá (32.9k rows), Igbo (18k rows) and Hausa (85k rows), utilizing sources such as Alaroye, BBC Yoruba, VON, Premium Times Hausa, and the Wura corpus. We further augmented our multilingual dataset with English queries translated via the Gemma3:27b model, and mapped it to each of the Yoruba, Igbo and Hausa documents, enabling cross-lingual semantic training. This extensive multilingual fine-tuning resulted in significant performance improvement across all three languages.

Specifically, for same language query to documents pairs, the fine-tuned model achieved mean reciprocal rank (MRR) scores of 0.9201 for Yorùbá, 0.8638 for Igbo, 0.9230 for Hausa, while for the combined, English queries to one of Yoruba, Igbo or Hausa documents, the model obtained a score of 0.8617. These results substantially surpass the baseline performance of BGE-M3, for same language query to document pairs, it obtained MRR scores of 0.7846 for Yorùbá, 0.7566 for Igbo, 0.8575 for Hausa, while for English to one of Yoruba, Igbo or Hausa, a score of 0.7377. Potential applications of our fine-tuned embedding model include multilingual information retrieval systems, question-answering applications, and semantic search tools tailored to local communities in Nigeria. As the output of our work, we provide a repository of the final dataset, scripts for scraping and processing the data, and the fine-tuned embedding model weights.