Submitted By,

Muhammad Awaiz

Reg No: 221001 (BSAI-VI-A)

Assignment 2

Submitted to,

Dr. Humaira Waqas

Multilingual Medical Chatbot Assistant

1. Description of dataset

This dataset is a curated collection of publicly available MedlinePlus Languages health-related documents in multiple languages. It pairs English documents with their translated counterparts in multiple languages.
Data Sources

- CDC (Centers for Disease Control and Prevention)
- FDA (U.S. Food and Drug Administration)
- California Dental Association (CDA)
- American Cancer Society

The dataset currently includes aligned documents in **Urdu** and **Spanish**, covering essential health topics such as vaccination, chemotherapy, and diagnostic procedures. Each document pair contains identical or equivalent content in both the source and translated language, making it highly suitable for training or evaluating multilingual models.

2. Preprocessing and chunking

To prepare documents for downstream tasks, text was extracted from multilingual PDFs using different strategies based on document layout and language:

- **Standard Extraction**: Applied to most English and Spanish documents by reading pages sequentially.
- **Urdu Handling**: Involved word-order inversion to accommodate right-to-left script.
- **Two-Column Layouts**: Spanish documents with dual columns were split and read separately to maintain reading order.

Multiple chunking methods were used to segment text for retrieval and processing:

a) **Fixed-Size Chunking**: Splits text into uniform word-length chunks.
b) **Overlap-Based Chunking**: Similar to fixed-size but with overlapping words for better context retention.
c) **Semantic Chunking**: Groups sentence until a word limit is reached, preserving sentence boundaries.
d) **Dynamic Chunking**: Uses token counts to balance chunk size dynamically while respecting semantic flow.
e) **Delimiter-Based Chunking**: Splits text at natural breaks such as paragraph gaps.

3. Architecture of embedding models and vector store

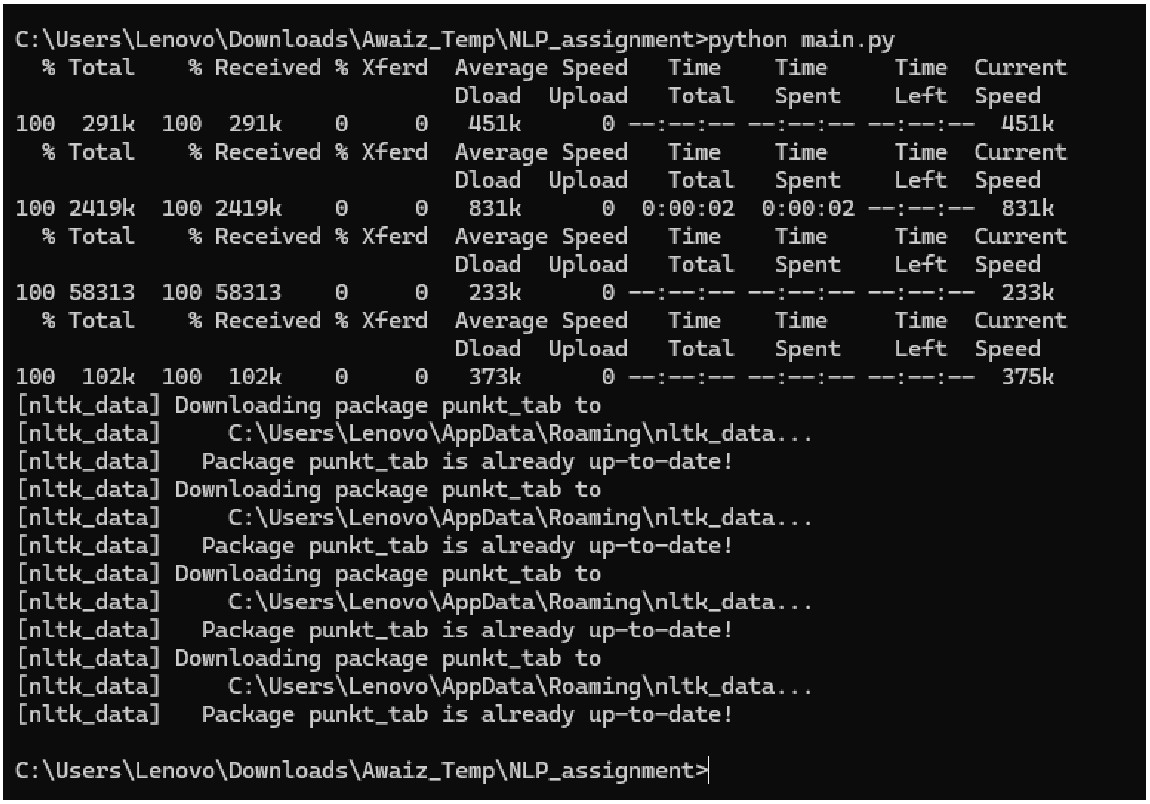
Model Name	Full Name	Base Architecture	Training Objective	Languages Supported	Embedding Dimension	Pooling Method	Input Format	Optimized For
LaBSE	Language-agnostic BERT Sentence Embedding	BERT	Translation ranking	109	768	Mean pooling	Plain sentences	Multilingual sentence similarity
XLM-R (STSB)	STSB-XLM-R-Multilingual	XLM-RoBERTa	STSB (Semantic Textual Similarity Benchmark)	100+	768	Mean pooling	Plain sentences	Semantic similarity
Multilingual-E5-Large	Multilingual-E5-Large	Transformer (contrastive learning)	Instruction tuning (query-passage contrast)	100+	1024	Mean pooling (token reps)	`query: <text>` or `passage:`	Dense retrieval, RAG

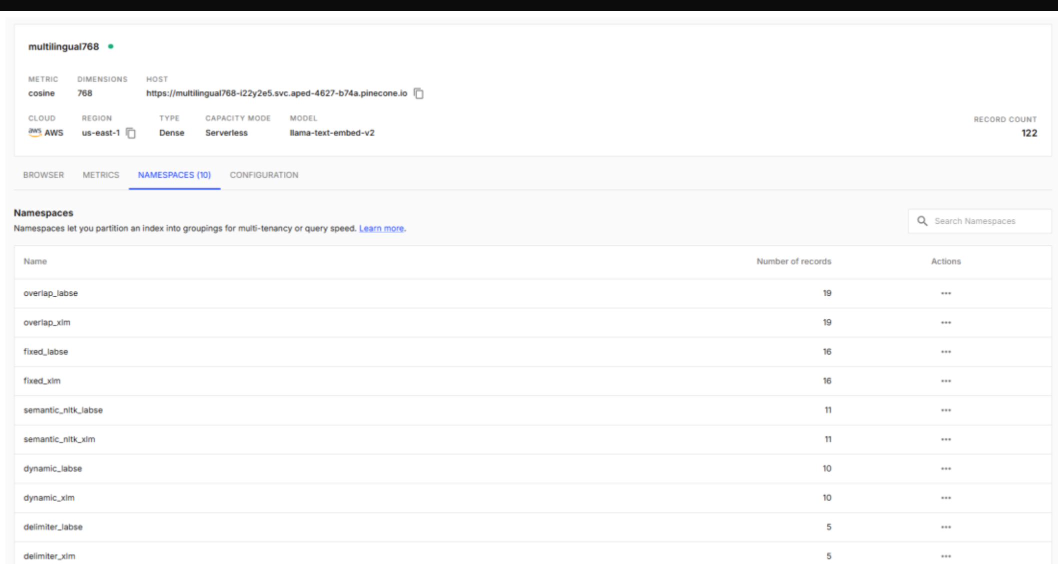
For every combination of document, chunking strategy, and embedding model, vectors were stored in a dictionary-like format with the following fields:

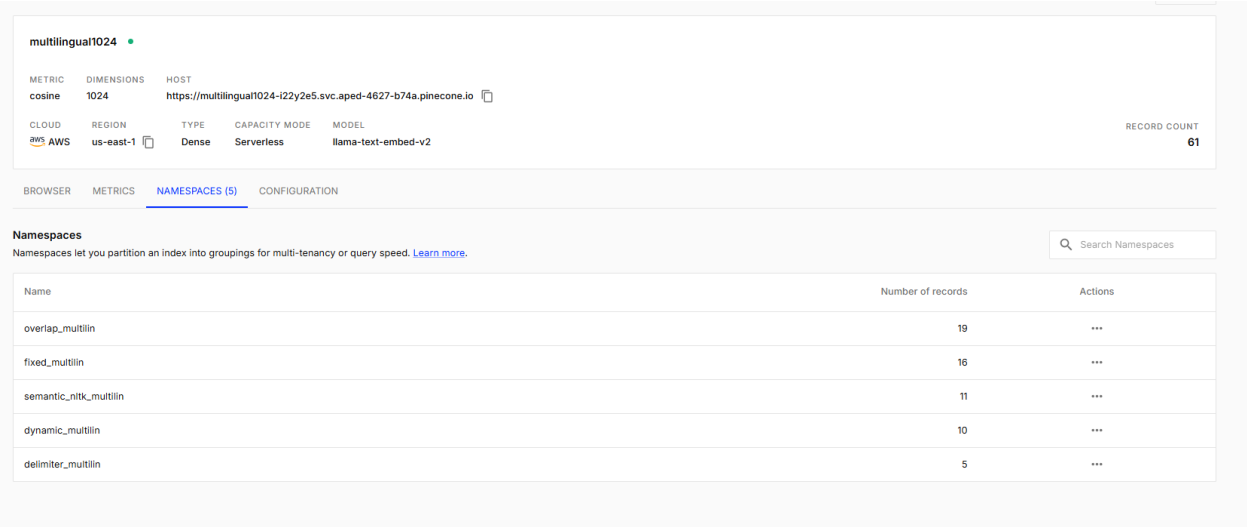
- id: A unique identifier for the chunk
- values: The generated embedding vector
- metadata: A dictionary containing

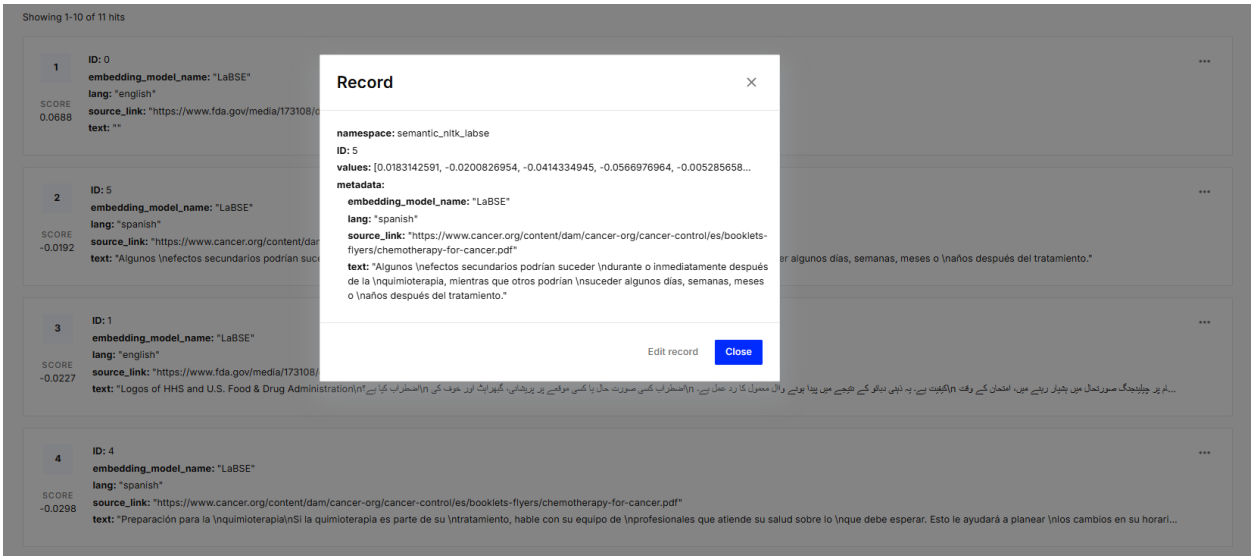
- text: The original chunk text
- lang: Language of the document
- source_link: URL to the source PDF
- embedding_model_name: Name of the embedding model used

4. Working implementation









5. Comparative table for model chunking combination

Model	Chunking Strategy	Average Query Time (ms)	Average Chunk Size (chars)	Total No.Chunks	Query Time Quality	Chunk Size Quality	Chunk Count Quality	Highest Spanish Score	Highest Urdu Score
labse	fixed	898.05	598.9	10	Poor	Medium	Low	0.404207706	0.471276402
labse	overlap	395.88	571.31	13	Good	Medium	High	0.419589311	0.471276402
labse	semantic_nltk	427.75	553.27	11	Good	Medium	Medium	0.542837	0.422487408
labse	dynamic	331.28	608.7	10	Excellent	Medium	Low	0.557167649	0.422487408
labse	delimiter	317.97	1523.5	4	Excellent	Large	Low	0.366119742	0.422487408
xlm	fixed	363.8	598.9	10	Good	Medium	Low	0.372436374	0.624856412
xlm	overlap	319.14	571.31	13	Excellent	Medium	High	0.342016429	0.624856412
xlm	semantic_nltk	315.49	553.27	11	Excellent	Medium	Medium	0.544519901	0.354225129
xlm	dynamic	309.77	608.7	10	Excellent	Medium	Low	0.655139923	0.343804121
xlm	delimiter	301.99	1523.5	4	Excellent	Large	Low	0.268244863	0.343804121
multilin	fixed	972	598.9	10	Poor	Medium	Low	0.76240325	0.901147723
multilin	overlap	435.83	571.31	13	Good	Medium	High	0.764075816	0.901147723
multilin	semantic_nltk	580	553.27	11	Good	Medium	Medium	0.85089618	0.863780558
multilin	dynamic	423.26	608.7	10	Good	Medium	Low	0.862845242	0.863780558
multilin	delimiter	441.63	1523.5	4	Good	Large	Low	0.765482962	0.863780558