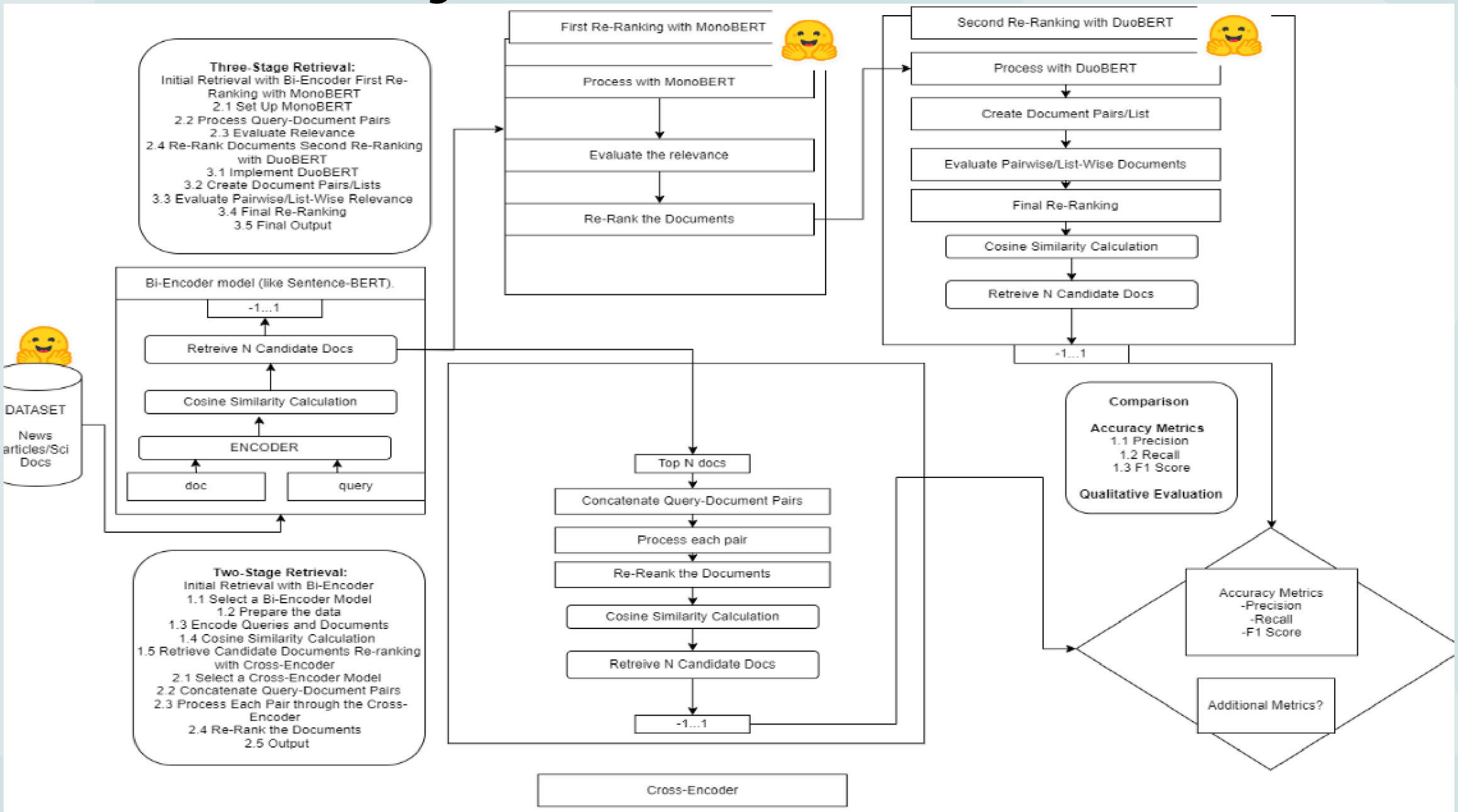
Comparative Analysis

Two-Stage vs. Three-Stage Retrieval Models in Information Retrieval

Public Repository Link:
https://github.com/skilh/AIR23/tree/main

Ilhan Skender
Tarik Barucija
Namik Barucija

System Arhitecture



MS MARCO Dataset Overview

- Purpose: Enhancing machine reading comprehension and information retrieval.
- Content: Real-world questions and search queries from Bing.
- Components:
  - Passage Ranking Dataset: Passages ranked for query relevance.
- Scale: Large, diverse dataset with hundreds of thousands of entries.
- Real-World Application: Questions based on actual user queries.
- Research Utility: Ideal for NLP models in question answering and information retrieval.
- For our project we used a subset of 10000 samples

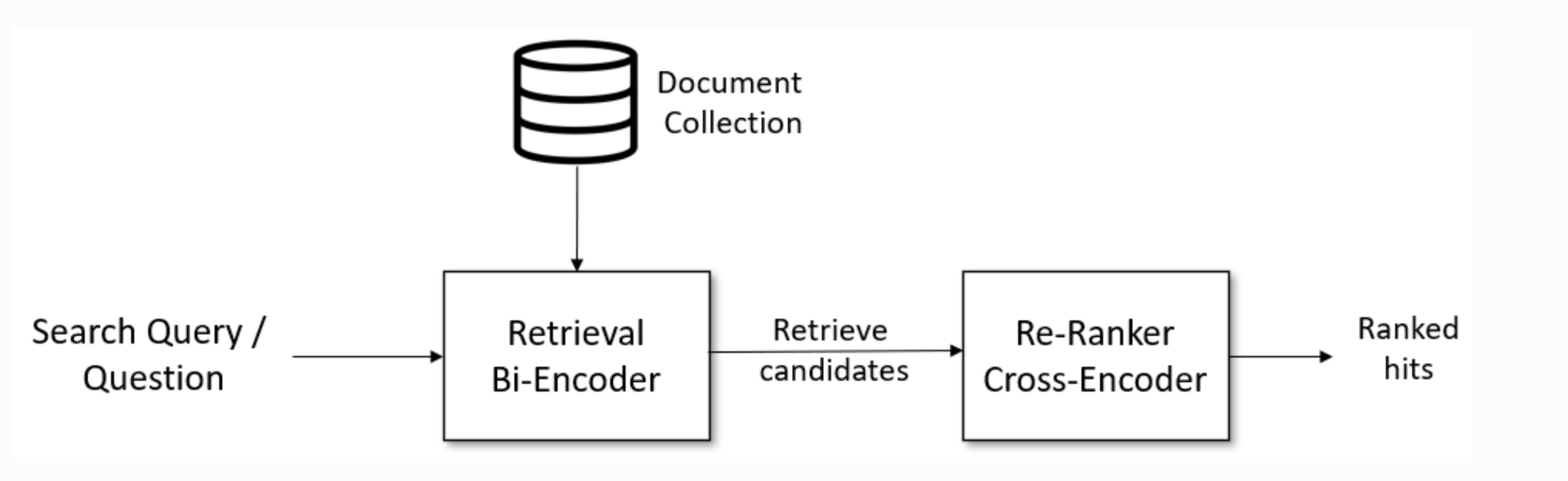
https://microsoft.github.io/msmarco/

Overview of the Retrieval Models

- **Two-Stage Retrieval Model**
- First Stage: Utilizes SBERT for generating embeddings and cosine similarity for initial ranking.
- Second Stage: Employs a Cross-Encoder for re-ranking, focusing on relevance.
- **Three-Stage Retrieval System:**
- MonoBERT and DuoBERT as additional stages.
  https://github.com/castorini/duobert

sbert

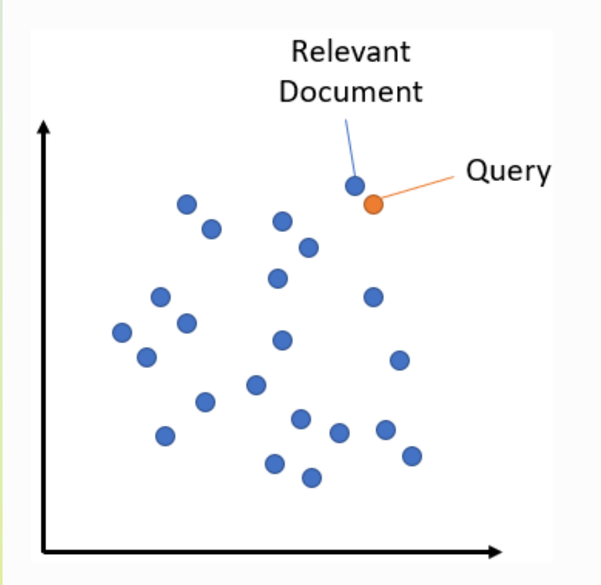
First Stage Retrieval - SBERT and Cosine Similarity



SBERT

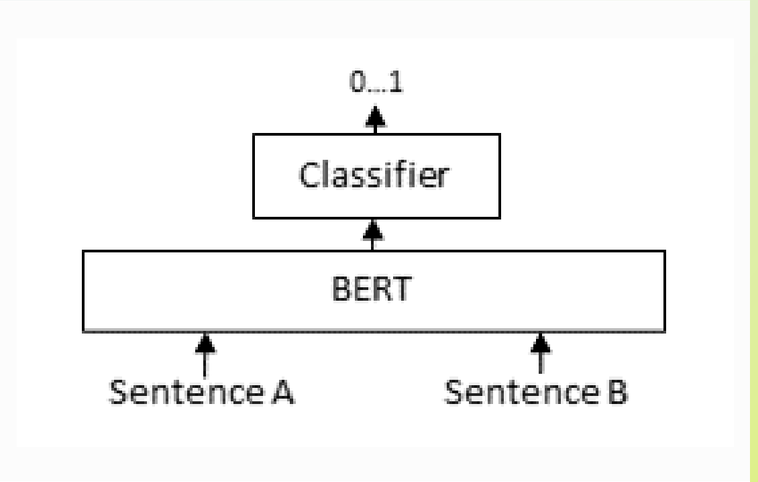
https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

Steps:
Embedding Generation using SBERT
Compute similarity between query embeddings and document embeddings



Second Stage Retrieval - Cross Encoders

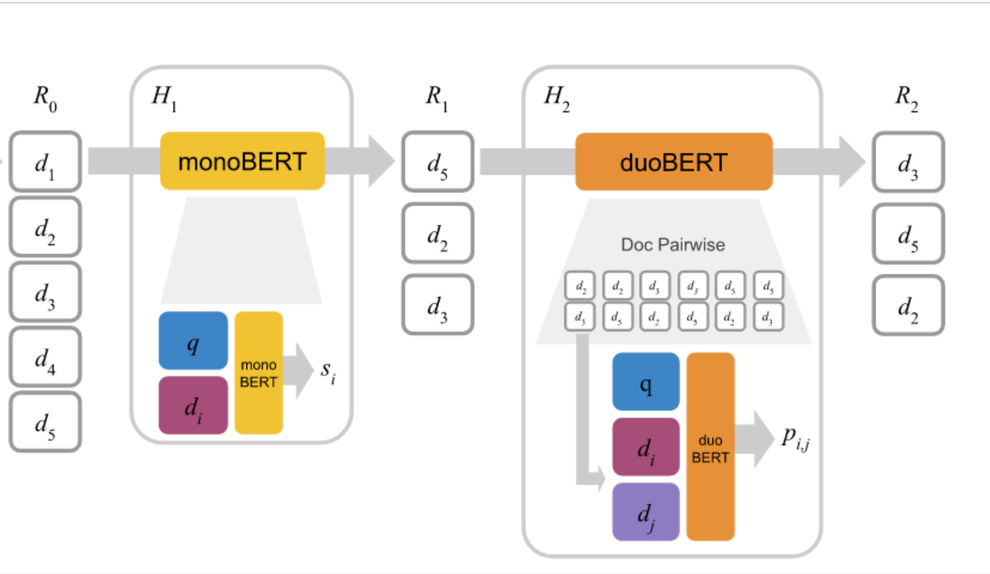
- Used for re-ranking the initial set of documents retrieved by the first stage.
- It assesses each query-document pair and assigns a relevance score
- Provides a more nuanced and context-aware ranking compared to the initial cosine similarity-based retrieval.
- used a pre-trained Cross-Encoder model (nboost/pt-bert-base-uncased-msmarco), specifically fine-tuned on the MS MARCO dataset



cross_encoder_model_name = 'nboost/pt-bert-base-uncased-msmarco'

MonoBERT and DuoT5

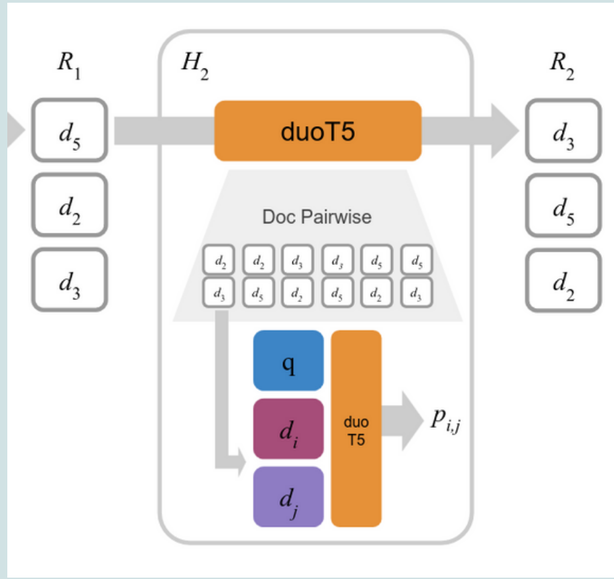
Inital Idea for pipeline:



Detailed examination and refinement of initial ranking
MonoBERT (BERT model) fine-tuned for MS MARCO dataset

Source: https://github.com/castorini/duobert

DuoT5 instead of DuoBERT



DuoT5 - a variant of the T5 model adapted for pairwise comparison tasks, fine-tuned for MS MARCO dataset

Compares pairs of documents from the refined set (output of MonoBERT) to each other in the context of the query

Source: https://arxiv.org/pdf/2101.05667.pdf

Final results

Two-Stage

Initial Retrieval:

F1 Score: 0.1568627450...

MRR: 0.21192810...

Re-Rank:

F1 Score: 0.29901960...

MRR: 0.30392156862...

Three-Stage

Final:

F1 Score: 0.034313...

MRR: 0.1241830...

We have a mistake here obviously :(

THANK YOU