UTILIZING DEEP LEARNING FOR HYBRID RETRIEVAL AND MULTI-STAGE TEXT RANKING

Group 10

Parag Sharma - 202103004

Parv Agarwal - 202103010

Darshak Zankat - 202103041

INTRODUCTION

This is an empirical study aiming to create an efficient ranking system, blending sparse and dense retrieval methods to optimize text retrieval and ranking in information retrieval systems

- Challenge: Deep learning methods demand vast datasets, hindering traditional information retrieval.
- Solution: The Deep Learning Track at TREC addresses this by supplying large-scale datasets for blind evaluation.
- Evolution: From BM25 to Dense Retrieval (DRES), and now leveraging pre-trained models like BERT, a transformative shift in retrieval methods has occurred.
- Optimization: Hybrid retrieval methodologies and diverse backbone networks promise a new era of comprehensive ranking performance.

DATASET DESCRIPTION AND EVALUATION

Graded Relevance

Normalized Cumulative Discount Gain (nDCG@k)} metric strikes an optimal balance for graded relevance assessments

The passage corpus is in jsonl format. Each passage has

pid

passage

spans

docid

MS MARCO dataset is union of the top ten passage lists for the one million queries, giving 8.8 million distinct passages

MODEL ARCHITECTURE

Retrieval

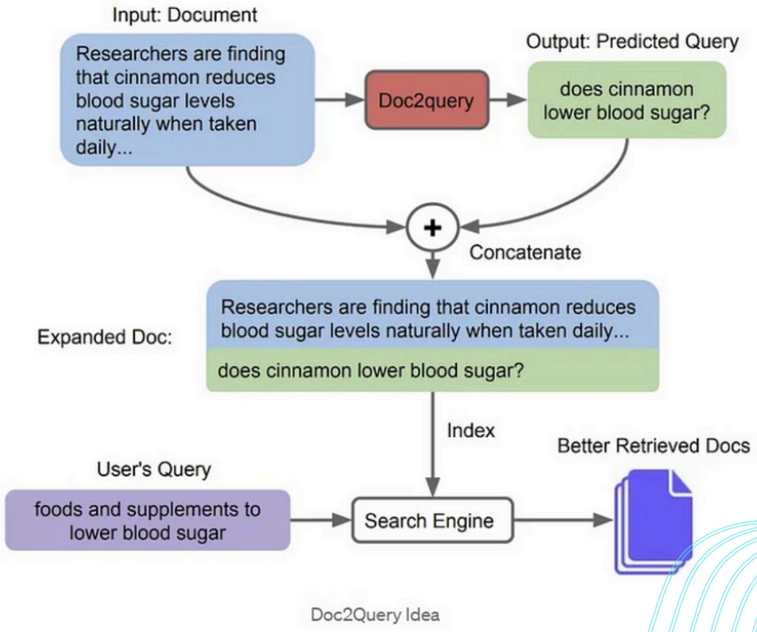
- BM25
- Elastic BM25
- Doc2Query
- SPARTA
- ACNE
- SBERT (trained with inbatch negatives)
- SBERT (pre-trained)
- SBERT with faiss indexing (HNSW)
- SBERT with faiss indexing(FLAT)

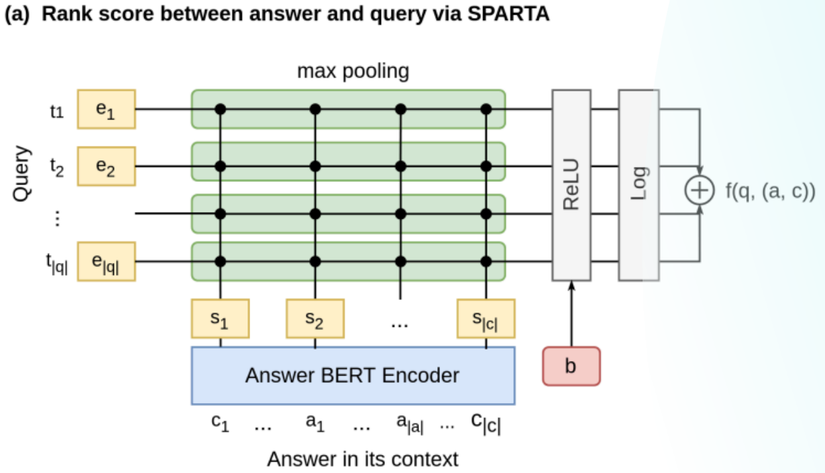
Retrieval + Reranking

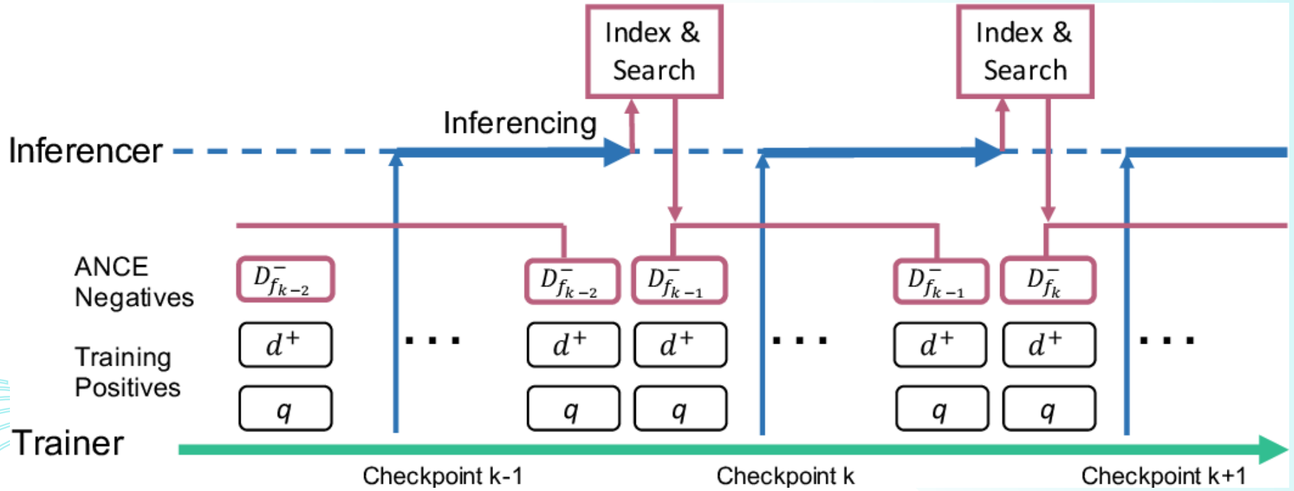
- Bi -encoders(Retrieval) + Cross-Encoders(Rerank)
- SBERT with faiss (HNSW) indexing(retrieve) + cross-encoders(rerank)
- SBERT with faiss indexing(retrieve) + cross-encoders(rerank)

SPARSE AND DENSE RETRIVAL

$$\sum_{t \in q} \log\left(\frac{N}{df_t}\right) \cdot \frac{(k_1+1)\cdot tf_{t,d}}{k_1 \cdot \left((1-b)+b\cdot\frac{L_d}{L_{ave}}\right)+tf_{t,d}}$$







1. BM25 Lexical Retrieval :

- Implemented BM25 model for word-level similarity in document and query matching.
- Tuned parameters k1 and b (0.9 and 0.4) for optimal document ranking.

2. Sparse Retrieval with Doc2Query :

- Used doc2query to enhance BM25 with sequence-to-sequence expansion.

3. Hybrid Sparse Retrieval (SPARTA) :

- Integrated SPARTA for two-stage sparse retrieval with transformer-based refinement.
- Utilized DistilBERT embeddings for similarity scoring.

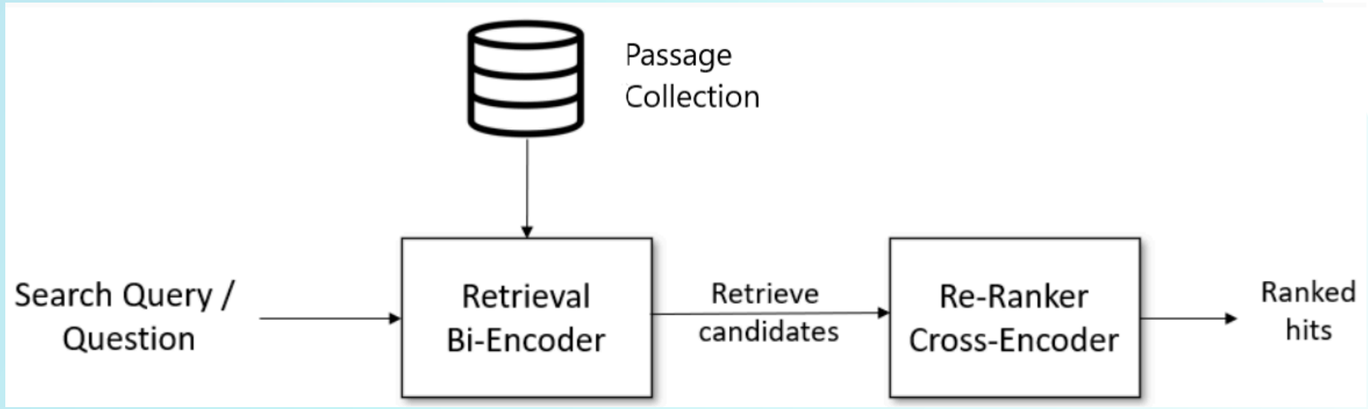
4. Dense Retrieval (ANCE Model) :

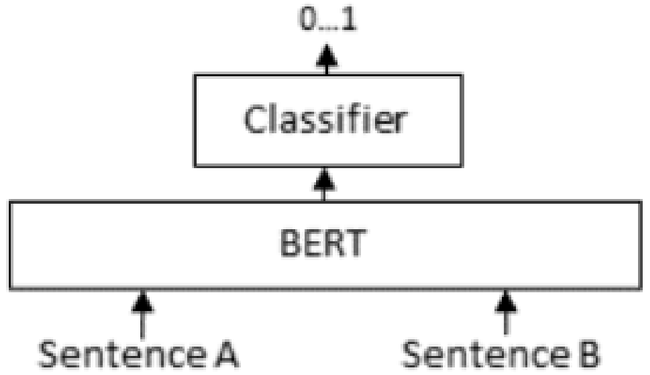
- Executed dense retrieval using ANCE model with optimal pre-training.
- Leveraged the effectiveness of pre-trained models for efficient document retrieval.

5. SBERT for Dense Retrieval :

- Utilized SBERT (msmarco-distilbert-base-v3) for dense retrieval.
- - Found superior performance in pre-trained models compared to in-batch negatives in SBERT (distilbert-base-uncased) fine-tuning.

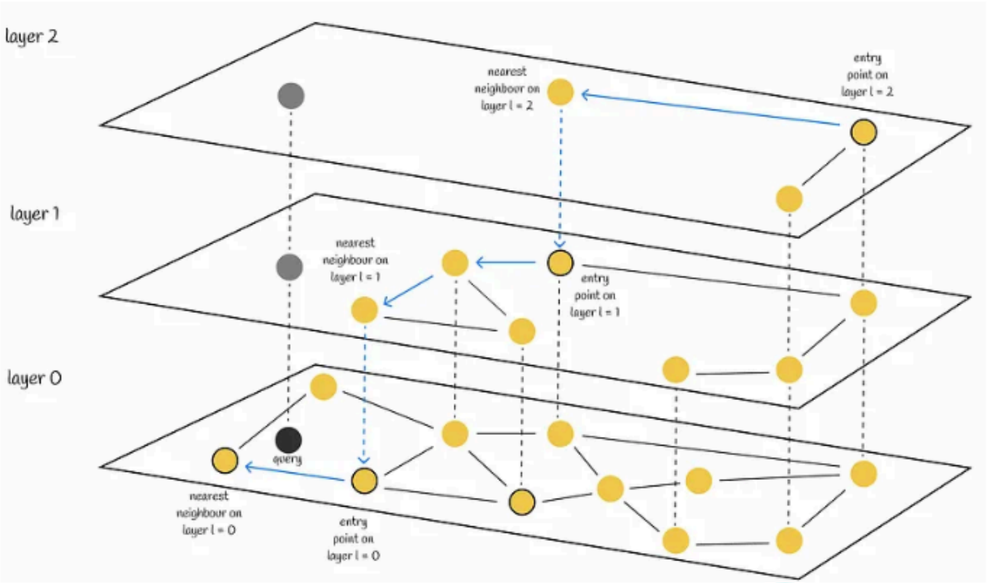
BIENCODER(RETRIEVE) + CROSS ENCODER(RERANK)

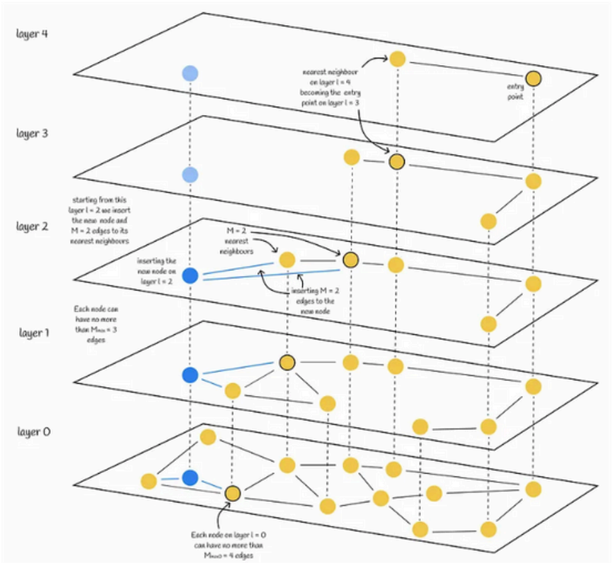




SBERT WITH FAISS INDEXING(RETRIEVE) + CROSS-ENCODERS(RERANK)

FAISS
Scalable Search With Facebook AI





Insertion of a node (in blue) in HNSW. The maximum layer for a new node was randomly chosen as l = 2. Therefore, the node will be inserted on layers 2, 1 and 0. On each of these layers, the node will be connected to its M = 2 nearest neighbours.

1. Bi-Encoders & Cross-Encoders :

- Implemented a two-step approach, employing bi-encoders for initial retrieval and cross-encoders for reranking, ensuring superior performance by considering both query and passage.

2. FAISS Inner Product Indexing :

- Used FAISS with Flat Inner Product indexing for efficient cosine similarity search, employing SBERT for vector generation and cross-encoder reranking.

3. HNSW for Nearest Neighbors :

- Implemented HNSW indexing for advanced nearest neighbor search, utilizing a multi-layered graph structure for efficient query results.

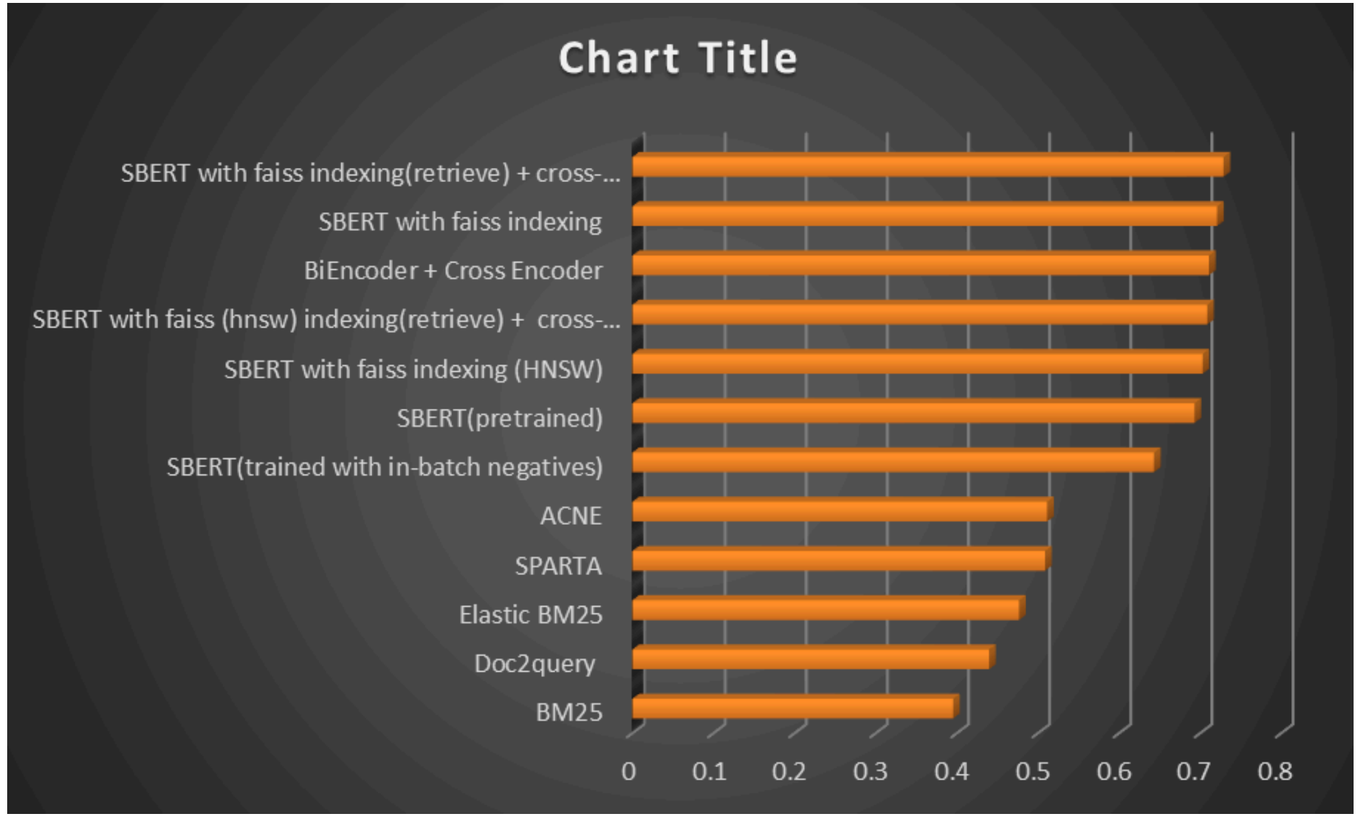
4. Efficient Top-20 Reranking :

- Balanced efficiency and performance by bi-encoder retrieval of top 20 candidates, applying computationally intensive cross-encoder reranking solely on the top candidates.

RESULTS AND ABLATION STUDY

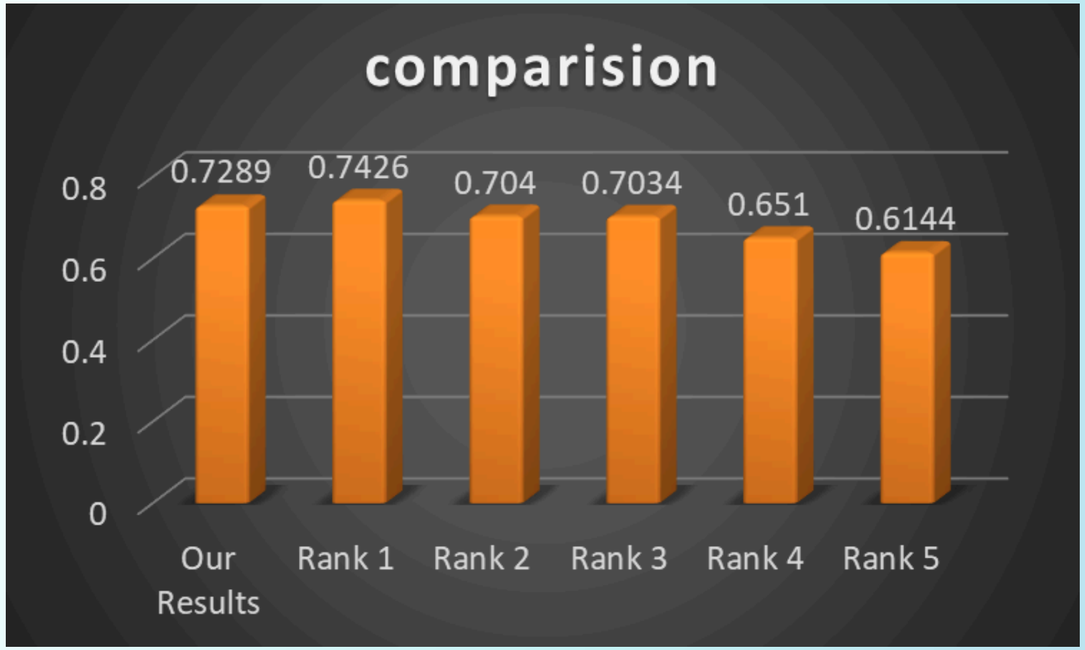
Model	nDCG@10
BM25	0.3954
Doc2query	0.4398
Elastic BM25	0.4768
SPARTA	0.5087
ACNE	0.5114
SBERT (trained with in-batch negatives)	0.6431
SBERT (pretrained)	*0.6930*
SBERT with faiss indexing (IINSW)	*0.7030*
SBERT with faiss (hnsw) indexing(retrieve) + cross-encoders(rerank)	*0.7087*
BiEncoder + Cross Encoder	*0.7109*
SBERT with faiss indexing	*0.7209*
SBERT with faiss indexing(IP)(retrieve) + cross-encoders(rerank)	***0.7289***

TABLE I: Ablation experiments result on MS MARCO passage dataset.



COMPARISION

Model	nDCG@10
Our Results	**0.7289**
Rank 1	0.7426
Rank 2	0.704
Rank 3	0.7034
Rank 4	0.651
Rank 5	0.6144



OVERVIEW PAPER AND TOP PAPERS

- https://trec.nist.gov/pubs/trec31/papers/Overview_deep.pdf
- https://trec.nist.gov/pubs/trec31/papers/Ali.D.pdf
- https://arxiv.org/pdf/2208.07670.pdf
- https://trec.nist.gov/pubs/trec31/papers/CIP.D.pdf
- https://trec.nist.gov/pubs/trec31/papers/UoGTr.D.pdf
- https://trec.nist.gov/pubs/trec31/papers/yorku22.D.pdf