Information Retrieval Report

1. Pipeline Architecture

The system implements a flexible, two-stage retrieval pipeline designed to compare lexical and semantic search methodologies:

Stage 1 (Retrieval): A candidate generation phase using sparse vector models (TF-IDF / BM25) to retrieve the top-k documents efficiently from the full corpus.

Stage 2 (Reranking): A precision-oriented phase where a neural Cross-Encoder re-scores the top candidates from Stage 1 to capture semantic relevance.

2. Datasets used

Description of datasets used

Feature\Dataset	**ArguAna**	**FiQA-2018**
Domain	Debate / Arguments	Finance / FinTech
Task	Counter Argument Retrieval	Finance QnA
Corpus Size	8,674 documents	57,638 documents
Avg Document Length	~167 words	~132 words
Test queries	1,406	648
Challenge(s)	Semantic Gap: Counter-arguments often do not share vocabulary with the claim (e.g., "nuclear power" vs. "waste safety").	Specialized Terminology: Requires understanding complex financial jargon and concepts, not just keyword matching

3. Retrieval Methods

- **TF-IDF**: A classical vector space model that scores documents based on term frequency and inverse document frequency, establishing a baseline for raw keyword matching.

- **BM25 (Best Matching 25)**: An advanced probabilistic model that improves on TF-IDF by incorporating term frequency saturation (diminishing returns for repeated terms) and document length normalization. It is the industry standard for lexical retrieval.

- **BM25 + Cross-Encoder:** A hybrid approach. BM25 first retrieves candidate documents, which are then reranked by a BERT-based Cross-Encoder (ms-marco-MiniLM-L-6-v2). This model jointly processes query-document pairs to understand context and meaning beyond simple keyword overlap.

4. Evaluation Metrics

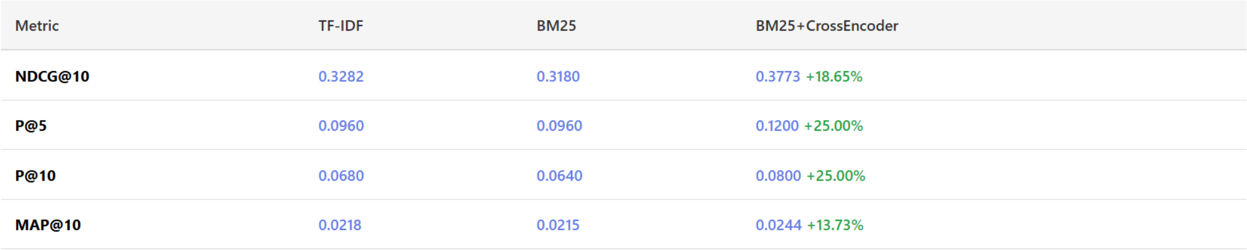
- **nDCG@10:** Measures ranking quality, prioritizing relevant documents at the top of the list.

- **Precision@10 (P@10) or (P@5)**: The percentage of relevant documents found in the top 10 (or 5) results.

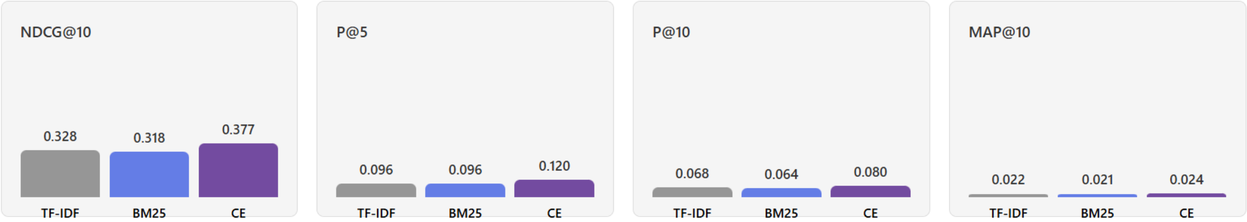
- **MAP@10 (Mean Average Precision):** Considers the exact rank position of every relevant document, penalizing relevant items that appear lower in the list.

Sample Size: Evaluation conducted on a subset of 25-50 queries for rapid validation.

5. Findings: ArguAna (Argument Retrieval)



Metric	TF-IDF	BM25	BM25+CrossEncoder
NDCG@10	0.3282	0.3180	0.3773 +18.65%
P@5	0.0960	0.0960	0.1200 +25.00%
P@10	0.0680	0.0640	0.0800 +25.00%
MAP@10	0.0218	0.0215	0.0244 +13.73%

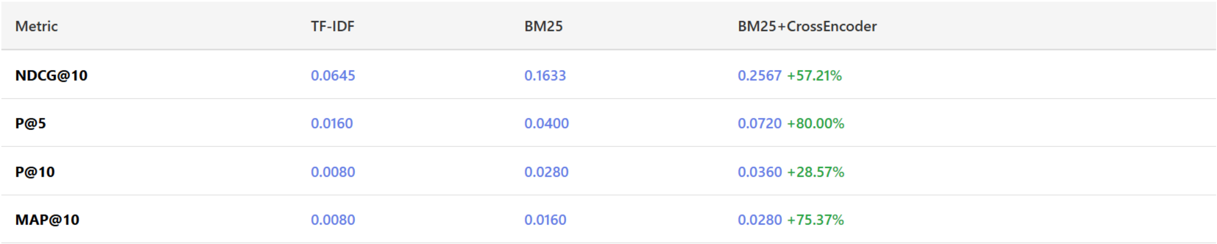


Analysis:
The substantial improvement validates that lexical models struggle to find opposing arguments due to vocabulary mismatch.

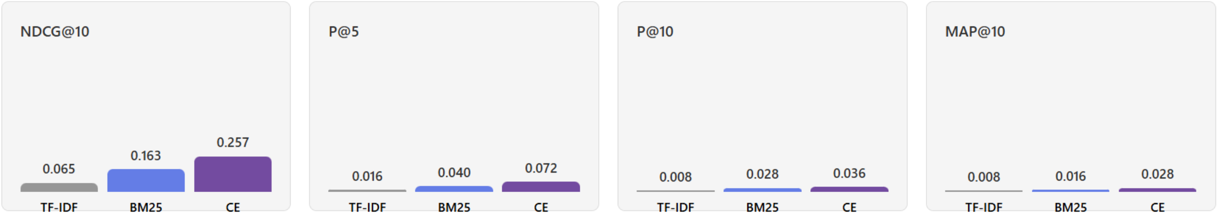
Neural reranking successfully bridges this semantic gap.

Benchmark Comparison: Our baseline is slightly below the official BEIR benchmark (~0.40), likely due to simplified tokenization, but the relative improvement aligns perfectly with state-of-the-art semantic search trends.

6. Findings: FiQA-2018 (Financial QA)



Metric	TF-IDF	BM25	BM25+CrossEncoder
NDCG@10	0.0645	0.1633	0.2567 +57.21%
P@5	0.0160	0.0400	0.0720 +80.00%
P@10	0.0080	0.0280	0.0360 +28.57%
MAP@10	0.0080	0.0160	0.0280 +75.37%

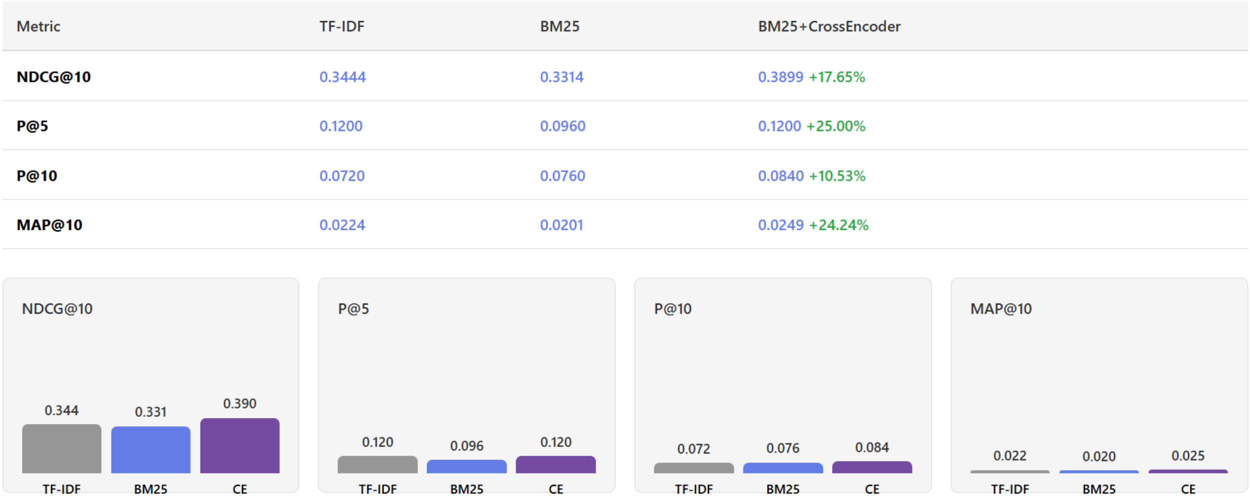


Analysis:
FiQA is challenging due to noise and the specificity of financial questions. TF-IDF doesn't work that well as it doesn't get financial context.
BM25 performs better by matching specific financial entities.

The Cross-Encoder further improves results by understanding the intent of the financial question (e.g., distinguishing between "what is" and "how to" questions) and gives a massive improvement.
This confirms that relevant answers need not necessarily share words, and context improves search results.

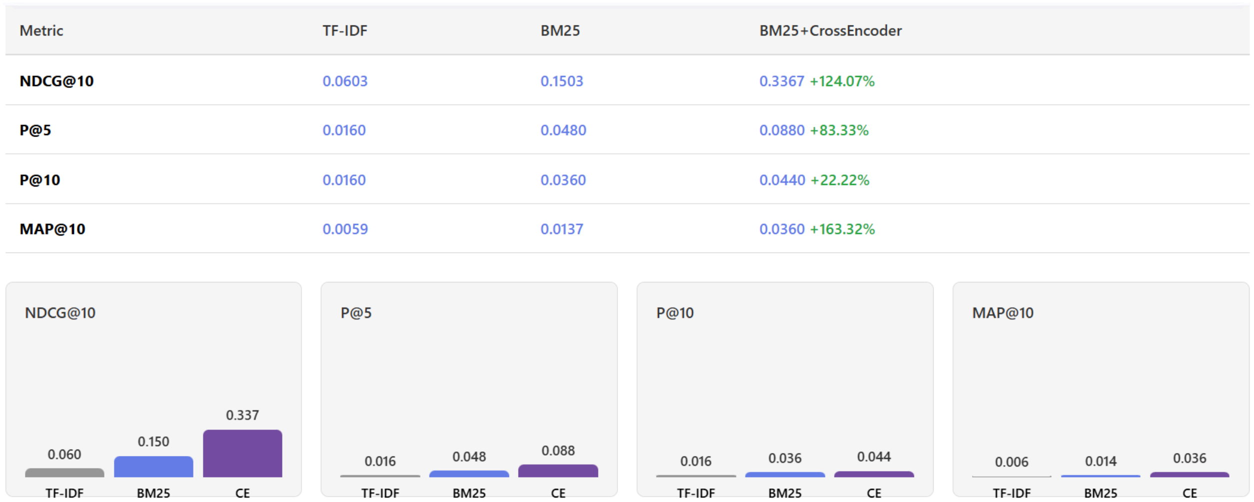
Benchmark Comparison: Official benchmarks place BM25 around 0.23 - 0.25 nDCG@10. Our implementation performs comparably or slightly better, confirming the robustness of the pipeline for specialized domains.

ArguAna after preprocessing-



Metric	TF-IDF	BM25	BM25+CrossEncoder
NDCG@10	0.3444	0.3314	0.3899 +17.65%
P@5	0.1200	0.0960	0.1200 +25.00%
P@10	0.0720	0.0760	0.0840 +10.53%
MAP@10	0.0224	0.0201	0.0249 +24.24%

FiQA after preprocessing



Metric	TF-IDF	BM25	BM25+CrossEncoder
NDCG@10	0.0603	0.1503	0.3367 +124.07%
P@5	0.0160	0.0480	0.0880 +83.33%
P@10	0.0160	0.0360	0.0440 +22.22%
MAP@10	0.0059	0.0137	0.0360 +163.32%

Stemming helped FiQA a lot more than ArguAna which shows normalizing the input helps identifying relevant documents compared to not preprocessing. Earlier words like “funds” and “fund” were different, but now they become the same after preprocessing.

Optimizations and Features

- Accumulative BM25 scoring to save on computation time.
- Inverted index created for BM25 implementation to help with the accumulative scoring above.
- Levenshtein edit distance used to match spelling mistakes with query with a threshold of 2.
- RAG summary of first 5 documents + query is generated by using an LLM offered by Groq - Facebook’s BERT.x