Retrieval Methods Evaluation Report: StarTech Dataset Analysis

Executive Summary

This report presents a comprehensive evaluation of six different information retrieval methods applied to the StarTech product dataset containing 8,462 product records. The analysis reveals significant performance differences between methods, with keyword-based search (BM25) substantially outperforming semantic search approaches across all evaluation metrics.

Methodology

Dataset Overview

- **Source**: StarTech product catalog
- **Size**: 8,462 product records
- **Fields**: Name, price, brand, category, subcategory, availability, image URL, product URL, model, rating
- **Document Preparation**: Created searchable text representations combining product attributes
- **Average Document Length**: 112 characters

Test Queries

25 diverse test queries were created covering:

- Brand-specific searches (ASUS laptops, Samsung monitors)
- Category-based queries (gaming laptops, 4K monitors)
- Feature-specific searches (RGB lighting, noise cancelling)
- Price-range queries (under 5000 taka, over 100000)
- Technical specifications (Intel i7, NVIDIA RTX, 16GB RAM)

Retrieval Methods Evaluated

1. **Semantic Search**: Using sentence-transformers (all-MiniLM-L6-v2) with FAISS indexing
2. **Keyword Search (BM25)**: Traditional term-based retrieval with Okapi BM25 scoring
3. **Hybrid Search (Weighted)**: Linear combination of semantic and keyword scores (α=0.7)
4. **Reciprocal Rank Fusion (RRF)**: Rank-based fusion of semantic and keyword results
5. **Semantic + Reranking**: Semantic search with cross-encoder reranking

6. **Hybrid + Reranking**: Hybrid search with cross-encoder reranking

Evaluation Metrics

- **Average Relevance**: Jaccard similarity between query and document terms
- **Top Relevance**: Relevance score of the highest-ranked document
- **Relevant Documents**: Count of documents exceeding relevance threshold (>0.1)
- **Diversity**: Complement of average pairwise similarity between results (1 - similarity)
- **Retrieval Score**: Method-specific scoring

Key Findings

Overall Performance Ranking

Rank	Method	Avg Relevance	Top Relevance	Relevant Docs	Diversity
1	Keyword Search (BM25)	0.1048	0.1195	4.80	0.4306
2	Reciprocal Rank Fusion	0.0891	0.0965	4.08	0.4039
3	Hybrid Search (Weighted)	0.0802	0.0895	3.76	0.3428
4	Hybrid + Reranking	0.0802	0.0922	3.76	0.3428
5	Semantic + Reranking	0.0768	0.0897	3.56	0.2724
6	Semantic Search	0.0697	0.0774	3.08	0.2971

Statistical Significance

- BM25 significantly outperformed semantic search ($p=0.0116$, $p<0.05$)
- No other method pairs showed statistically significant differences
- BM25 demonstrated the most consistent performance across all queries

Method-Specific Analysis

Keyword Search (BM25) - Best Performer

- **Dominance**: Won 100% of query comparisons (25/25 queries)
- **Performance**: 50.3% higher average relevance than semantic search
- **Strengths**: Superior term matching for product-specific queries
- **Diversity**: Highest diversity score (0.4306), indicating varied result sets

Semantic Search - Underperformed

- **Limitations**: Struggled with product-specific terminology and exact matches
- **Document Length Impact**: Short product descriptions (112 chars avg) may limit semantic understanding
- **Embedding Model**: General-purpose model may lack domain-specific knowledge

Hybrid Methods - Mixed Results

- **Unexpected Outcome**: 20.7% lower performance than pure BM25
- **Weight Distribution**: α=0.7 semantic weighting may be suboptimal
- **Fusion Challenges**: Simple linear combination may not effectively leverage both methods

Reranking - Limited Impact

- **Marginal Improvement**: Minimal gains over base retrieval methods
- **Cross-Encoder**: May be over-engineered for short product descriptions
- **Computational Overhead**: Added complexity without proportional benefit

Critical Analysis: Deviations from Expected Results

Expected vs. Actual Outcomes

Expected: Semantic Search Superiority

Theory: Semantic search should handle synonyms, context, and conceptual similarity better than keyword matching.

Reality: BM25 substantially outperformed semantic approaches.

Reasons for Deviation:

1. **Product Domain Characteristics**: E-commerce queries often contain specific brand names, model numbers, and technical specifications that require exact matching
2. **Document Structure**: Product descriptions are feature-rich but semantically sparse, favoring term-based matching
3. **Query Nature**: Test queries contained many specific terms (brand names, technical specs) that BM25 handles more effectively

Expected: Hybrid Method Advantage

Theory: Combining semantic and keyword approaches should leverage strengths of both methods.

Reality: Hybrid methods underperformed pure BM25.

Reasons for Deviation:

1. **Weak Semantic Component**: Poor semantic search performance dragged down hybrid results
2. **Suboptimal Weighting**: Linear combination may not be ideal for this domain
3. **Score Normalization**: Simple min-max normalization may have introduced artifacts

Expected: Reranking Benefits

Theory: Cross-encoder reranking should refine initial retrieval results.

Reality: Minimal improvements, sometimes worse than base methods.

Reasons for Deviation:

1. **Document Length Mismatch**: Cross-encoder trained on longer passages, not product descriptions
2. **Limited Context**: Short product descriptions provide insufficient context for sophisticated reranking
3. **Training Domain**: General-purpose reranker may lack e-commerce domain knowledge

Dataset-Specific Factors

Product Catalog Characteristics

- **Structured Data**: Products have well-defined attributes (brand, model, category)
- **Terminology Precision**: Technical specifications require exact matching
- **Limited Narrative**: Minimal descriptive text beyond factual attributes
- **Naming Conventions**: Product names follow patterns that keyword search handles well

Query Characteristics

- **Specificity**: Many queries target exact brands or technical specifications
- **Commercial Intent**: Users searching for specific products rather than conceptual exploration
- **Term Overlap**: High overlap between query terms and product names/attributes

Implications and Recommendations

Immediate Recommendations

1. **Deploy BM25 as Primary Method**: Given clear performance superiority across all metrics
2. **Optimize BM25 Parameters**: Fine-tune k1 and b parameters for product domain
3. **Enhanced Preprocessing**: Improve text preprocessing to handle product-specific terminology
4. **Query Expansion**: Consider synonym dictionaries for brand names and technical terms

Long-term Improvements

1. **Domain-Specific Embeddings**: Train embeddings on product catalogs and e-commerce data
2. **Hybrid Architecture Redesign**: Explore learned fusion approaches rather than linear combination
3. **Product-Aware Features**: Incorporate structured product attributes directly into retrieval
4. **Custom Evaluation Metrics**: Develop metrics that better reflect e-commerce search quality

Alternative Approaches to Consider

1. **Elasticsearch with Custom Scoring**: Leverage product structure with boosted fields
2. **Learning-to-Rank**: Train ranking models on click-through or purchase data
3. **Neural Information Retrieval**: Dense passage retrieval with domain-specific training
4. **Multi-Modal Retrieval**: Incorporate product images alongside textual descriptions

Limitations of Current Study

Evaluation Constraints

- **Subjective Relevance**: Jaccard similarity may not reflect user satisfaction
- **Limited Query Diversity**: 25 queries may not cover full search space
- **Missing User Feedback**: No actual user interaction data for validation
- **Single Domain**: Results may not generalize beyond e-commerce

Technical Limitations

- **Embedding Model**: General-purpose model not optimized for product data
- **Preprocessing**: Basic text preprocessing may miss important product features
- **Parameter Tuning**: Limited exploration of optimal hyperparameters
- **Hardware Constraints**: Evaluation conducted on limited computational resources

Conclusions

The evaluation reveals that traditional keyword-based retrieval (BM25) significantly outperforms modern semantic approaches for product search in the StarTech dataset. This outcome challenges conventional wisdom about semantic search superiority but aligns with the specific characteristics of e-commerce product catalogs.

The dominance of BM25 can be attributed to the structured nature of product data, the prevalence of exact term matching in product queries, and the limitations of general-purpose semantic models in specialized domains. Hybrid and reranking approaches failed to improve

performance, suggesting that simple combination strategies are insufficient when one component significantly underperforms.

These findings emphasize the importance of domain-specific evaluation and the need to match retrieval methods to data characteristics rather than assuming universal superiority of any single approach. For the StarTech dataset and similar product catalogs, traditional IR methods remain highly effective and should be the foundation for production systems.

Future work should focus on domain adaptation of semantic methods, sophisticated hybrid architectures, and incorporation of product-specific features to bridge the performance gap while maintaining the conceptual advantages of semantic search.