Context-Aware Text Re-Ranking through Unsupervised Manifold Learning

Luis F. A. Venezian[1], **Daniel Carlos Guimarães Pedronette**[1]

[1] Departamento de Estatística, Matemática Aplicada e Computação (DEMAC)
Instituto de Geociências e Ciências Exatas (IGCE)
Universidade Estadual Paulista (UNESP)

{luis.venezian,daniel.pedronette}@unesp.br

Abstract. *Information Retrieval (IR) enables efficient knowledge discovery within large document collections. Despite the evolution from sparse lexical models such as BM25 to dense neural embeddings, retrieval ranking still relies predominantly on independent query–document scoring, and most re-ranking strategies require supervised relevance signals. In image retrieval, unsupervised methods that exploit manifold structure and contextual neighborhood relationships have proven highly effective, yet their transfer to text remains unexplored. This study investigates whether such rank-based contextual re-ranking techniques can enhance text retrieval without any modality-specific adaptation. We present a systematic pipeline integrating data collection, sparse and dense retrieval, contextual re-ranking, and effectiveness analysis, and conduct 972 experiments spanning three unsupervised methods across nine diverse public datasets and four retrieval models. Results measured by MAP, Precision, and Recall show consistent improvements over baseline retrievers, with relative Precision@20 gains reaching up to +19.8%. These findings support the hypothesis that unsupervised contextual re-ranking can significantly improve text retrieval effectiveness and suggest that the underlying techniques are modality-agnostic.*

1. Introduction

Information Retrieval (IR) studies how to build systems that satisfy an information need by retrieving, ranking, and presenting documents (or passages) from large collections. In simple terms, the main goal of an IR system is to retrieve all documents that are relevant to the user's information need while retrieving as few irrelevant documents as possible [Baeza-Yates et al. 1999].

Text retrieval has advanced substantially in recent years, moving from traditional lexical matching strategies such as Bag-of-Words, TF–IDF, and BM25 [Aizawa 2003, Robertson et al. 1994] to powerful dense retrieval models, particularly after the advent of Transformer-based architectures [Vaswani et al. 2017, Zhao et al. 2024]. Despite this progress, the dominant retrieval paradigm both lexical and dense remains based on independent query and document encoding, typically realized through bi-encoder architectures, in which relevance is estimated via similarity in a shared representation space. As a consequence, retrieval is commonly performed by scoring each document independently with respect to the query, without explicitly modeling fine-grained query and document

interactions or contextual relationships within the collection, such as topical structure or inter document dependencies [Xu et al. 2025].

While not always framed explicitly as manifold learning, several recent works exploit local neighborhood structure or inter-document relationships to refine ranking decisions, echoing the classical manifold-based approaches, which posit that relevance should be inferred by considering the local structure of the retrieval space, rather than treating documents as isolated points [Ai et al. 2018, Zerveas et al. 2022, Zerveas et al. 2023].

Building on this perspective, we go beyond independent document embeddings by incorporating structural signals derived from inter-document proximity in the retrieval space. Prior work in image retrieval has shown that contextual neighborhood-based re-ranking methods can effectively exploit dataset structure to refine rankings [Valem and Guimarães Pedronette 2016, Guimarães Pedronette et al. 2021, Pedronette et al. 2021]. To the best of our knowledge, these rank-based contextual re-ranking techniques have not been systematically investigated for text retrieval, remaining largely confined to visual domains in the existing literature.

The main contributions of this work are: (i) a novel application of rank-based manifold learning methods to text retrieval re-ranking, demonstrating their effectiveness in a modality for which they were not originally designed; (ii) a large-scale experimental comparison of three unsupervised contextual re-ranking methods evaluated across nine diverse datasets spanning five domains and four retrieval models (one sparse and three dense), totalling 972 re-ranking experiments; and (iii) empirical evidence that these methods yield consistent improvements over baseline retrievers across all evaluated settings, with relative Precision@20 gains of up to +19.8%, supporting the hypothesis that rank-based contextual re-ranking is modality-agnostic and complementary to both sparse and dense first-stage representations.

The remainder of this paper is organized as follows. Section 2 reviews classical and neural re-ranking strategies in the related literature. Section 3 describes the proposed approach, including the retrieval representations, the three contextual re-ranking methods, and implementation details. Section 4 presents the experimental evaluation, covering datasets, protocol, quantitative results, qualitative analysis, and discussion. Finally, Section 5 draws conclusions and outlines directions for future work.

2. Related Work

Re-ranking refers to the process of refining an initial retrieval ranking in order to improve effectiveness. In most retrieval systems, documents are first selected using a relatively simple scoring function, such as lexical similarity or independent embedding-based matching [Liu 2009]. However, this first-stage retrieval often relies on independent query–document scoring and may fail to capture richer relevance signals. Re-ranking methods address this limitation by applying a second-stage refinement step over the top-ranked documents.

Historically, this problem has been approached through relevance feedback and probabilistic models, and later through supervised Learning-to-Rank techniques. More recently, neural architectures and contextual modeling strategies have further advanced re-ranking effectiveness. In the following sections, we review classical and modern re-ranking approaches.

2.1. Classical re-ranking strategies

Re-ranking is not a recent challenge in Information Retrieval. Early approaches such as Relevance Feedback [Rocchio Jr 1971, Salton and Buckley 1990] and probabilistic relevance models [Lavrenko and Croft 2003] already pursued to refine initial retrieval results through post-retrieval adjustments. Classical relevance feedback methods iteratively updated the query representation using information from relevant documents, often requiring explicit user judgments. In contrast, probabilistic relevance models formalized pseudo-relevance feedback performing query expansion by estimating a relevance language model from top-ranked documents, eliminating the need for manual supervision.

While relevance feedback methods focus on refining query representations, Learning-to-Rank (LTR) approaches reframed ranking as a supervised optimization problem. Rather than modifying the query, LTR methods learn a ranking function directly from labeled relevance judgments. Early pointwise methods treated ranking as regression or classification, while later pairwise and listwise methods such as RankSVM [Joachims 2002] and LambdaMART [Burges 2010] explicitly modeled relative ordering and optimized ranking metrics. These approaches marked a significant shift from heuristic or probabilistic query refinement toward data-driven ranking models trained on large-scale relevance annotations.

2.2. Neural retrieval and contextual multimedia re-ranking

The introduction of deep neural architectures reshaped both retrieval and re-ranking paradigms by shifting from sparse lexical matching to learned dense semantic representations. Instead of relying solely on term overlap, neural methods encode queries and documents into continuous vector spaces, enabling semantic matching beyond exact lexical correspondence.

Two main architectural strategies became prominent in neural retrieval pipelines. The first is the bi-encoder architecture, in which queries and documents are encoded independently into dense embeddings and ranked via similarity measures such as dot product or cosine similarity. Dense Passage Retrieval (DPR) [Karpukhin et al. 2020] demonstrated that supervised dense representations substantially improve semantic recall, while subsequent work such as Contriever [Izacard et al. 2021] showed that competitive retrievers can be learned through unsupervised contrastive pre-training.

The second strategy is the cross-encoder architecture, typically employed as a post-retrieval re-ranking component. In two-stage pipelines, an initial retriever generates a candidate set, which is then refined by jointly encoding each query–document pair. Transformer-based models such as BERT [Devlin et al. 2019] significantly strengthened this approach, as bidirectional self-attention enables deep contextual interaction between query and document tokens. Passage re-ranking with BERT [Nogueira and Cho 2019] established cross-encoders as strong supervised re-ranking baselines.

Although cross-encoder models mitigate the limitations of independent encoding by jointly modeling query–document interactions, they are typically applied in supervised settings and operate over restricted candidate sets. Moreover, relevance is still estimated independently for each document, without explicitly modeling structural relationships among retrieved items.

In parallel, a distinct line of research in multimedia retrieval has pursued unsupervised re-ranking strategies that exploit the manifold structure of ranked lists rather than learning from relevance labels. Methods such as CPRR [Valem and Guimarães Pedronette 2016], RDPAC [Guimarães Pedronette et al. 2021], and BFS-Tree [Pedronette et al. 2021] refine initial rankings by propagating affinity through K-nearest-neighbor graphs constructed solely from ordinal positions. These techniques have demonstrated substantial effectiveness gains in image retrieval benchmarks, yet their application to text retrieval remains, to the best of our knowledge, unexplored.

Because these methods operate exclusively on rank positions, without accessing feature vectors, scores, or modality-specific representations, they are, in principle, directly transferable to any retrieval setting. This observation motivates the present work, which investigates whether the same contextual re-ranking methods can improve text retrieval effectiveness across diverse datasets and retrieval models.

3. Proposed Approach

In this study, the central idea is that ranking quality can be improved by considering not only the pairwise similarity between a query and each candidate document, but also the broader neighborhood relationships among all retrieved items, that is, by leveraging contextual information encoded in the structure of multiple ranked lists simultaneously. Figure 1 illustrates the overall pipeline, which is structured into two main stages: initial retrieval and contextual re-ranking.

In the first stage, documents are encoded using either dense or sparse representations and ranked according to their similarity to the query. The second stage introduces a contextual re-ranking mechanism that operates over the ranked candidate set. Rather than relying solely on independent query–document similarity scores, this step leverages structural relationships among retrieved documents to refine their ordering. By incorporating neighborhood and manifold information, the framework aims to improve ranking effectiveness through contextual coherence. The following sections detail the representation models employed, the manifold-based re-ranking strategies, and the implementation specifics of the proposed approach.

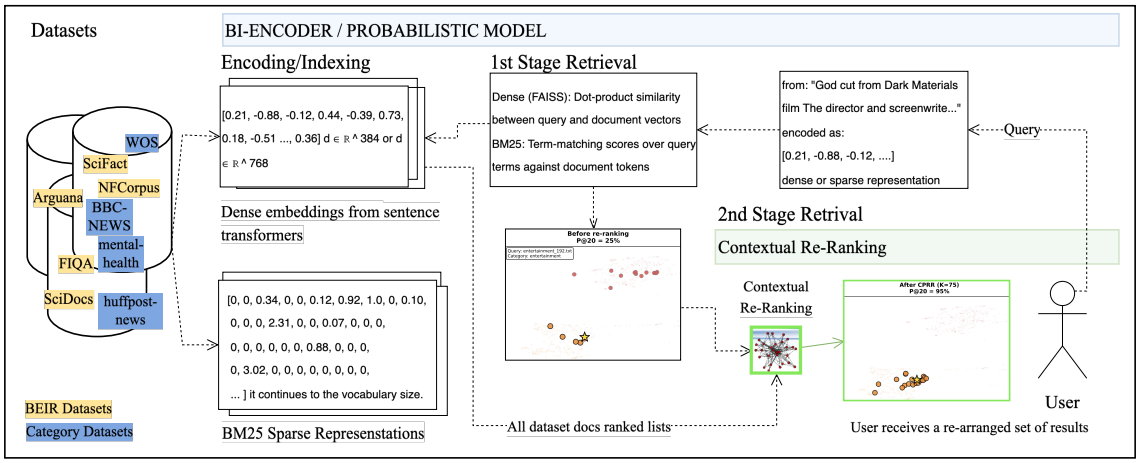

Figure 1. Pipeline overview: retrieval followed by contextual re-ranking.

3.1. Dense and sparse representations

To evaluate the proposed contextual re-ranking framework across different retrieval paradigms, we employ both sparse and dense representations in the first-stage retrieval.

As a lexical baseline, we use BM25, where documents and queries are represented as sparse term-weighted vectors. Relevance is computed through probabilistic term matching between query and document tokens, producing an initial ranked list.

For dense retrieval, documents and queries are encoded into fixed-dimensional embeddings using pretrained sentence-transformer models (MiniLM-L6-v2, BGE-small-en-v1.5, and Contriever-MSMARCO). Retrieval is performed by computing similarity in the embedding space, using dot-product similarity within a FAISS index.

In all cases, the first-stage retriever outputs a ranked list of candidate documents, which serves as input to the contextual re-ranking stage described in the following section.

3.2. Manifold learning for re-ranking

The three contextual re-ranking methods used in this work CPRR, BFS-Tree, and RDPAC were selected to represent distinct algorithmic families (Cartesian-product accumulation, breadth-first expansion, and diffusion-based random walk, respectively), while sharing a common principle: they refine an initial ranked list by exploiting neighborhood relationships among items in the collection, rather than relying on independent query–document scores. All three operate exclusively on ranked lists (ordinal positions), making them agnostic to the underlying retrieval representation.

Each method receives as input (i) a *ranked list file*, containing one line per query with space-separated document identifiers ordered by decreasing relevance, truncated to a window of L positions; and (ii) an *identifier list* with every unique element in the collection. The output is a re-ordered ranked list in the same format. Because only ordinal positions are consumed (not scores or embeddings) the same re-ranking methods apply identically to BM25 and dense retrieval outputs.

The three contextual methods differ in how they propagate neighborhood information but share the same rank-based interface. CPRR [Valem and Guimarães Pedronette 2016] accumulates pairwise affinity by computing the Cartesian product of each item's top-K reference set: for neighbors at ranks i and j, a similarity increment proportional to $(K-i) \times (K-j)$ is added symmetrically to an $N \times N$ affinity matrix, which is iteratively re-sorted over T iterations (we set $T=1$). BFS-Tree [Pedronette et al. 2021] first symmetrises the ranked lists through rank normalisation, then grows a breadth-first tree rooted at each item: neighbors are expanded when their list-level correlation—measured by Rank-Biased Overlap (RBO)—exceeds a threshold, and the accumulated affinities are propagated through matrix-based diffusion followed by probability normalisation. RDPAC [Guimarães Pedronette et al. 2021] formulates re-ranking as a lazy random walk on a K-nearest-neighbor graph with convergence guarantees: a transition matrix with exponential-decay weights ($p=0.6$, $p_l=0.99$) is built for increasing values of K, mixed with the identity at each step ($\alpha=0.95$), and amplified by a power-iteration step (P^3) that sharpens the separation between relevant and non-relevant neighbors. All three methods share two key properties relevant to this study: (i) they are fully unsupervised, requiring no relevance labels or training data; and

(ii) they operate on rank positions rather than feature vectors, which enables their direct application to text retrieval outputs without any modality-specific adaptation.

3.3. Implementation details

The entire pipeline is implemented in Python and orchestrated through Hydra configuration files, enabling reproducible composition of datasets, retrieval models, and re-ranking parameters via command-line overrides. The full implementation is publicly available.[1]

Contextual re-ranking is performed via the UDLF (Unsupervised Distance Learning Framework)[2] binary, interfaced through the pyUDLF Python wrapper (v0.8).[3] Each re-ranking method receives the initial ranked list in horizontal format and a list of all unique identifiers. We set a fixed window size of L=200 for all methods and vary the neighborhood parameter $K \in \{1, 3, 5, 10, 20, 30, 40, 50, 75\}$. Method-specific parameters follow default recommendations: CPRR uses t=1 iteration; BFS-TREE uses the Rank-Biased Overlap (RBO) correlation metric; RDPAC uses p=0.6, p_l=0.99, and l_mult=2.

4. Experimental Evaluation

This section presents a systematic evaluation of the proposed contextual manifold re-ranking strategies in text retrieval. We assess their impact on ranking effectiveness by comparing baseline retrieval results with their contextualized counterparts under both sparse and dense representations.

4.1. Datasets

The experiments were conducted on nine publicly available datasets covering diverse domains and retrieval settings. Five datasets are drawn from the BEIR benchmark—arguana, scifact, nfcorpus, scidocs, and fiqa—which provide standard query-document relevance judgments (qrels). To complement, we include four additional publicly available corpora—wos, bbc-news, huffpost-news, and mental-health—which are category-based datasets where document category attribute data are used to derive relevance signals.

To properly evaluate re-ranking effectiveness across diverse corpora, it is important to characterize how relevance is defined in each dataset, as the structure of the relevant sets directly impacts the computation of effectiveness metrics. Table 1 summarises the key properties of all datasets used in our experiments, while Table 2 details the distribution of relevant documents per query.

[1] https://github.com/unesp-text-retrieval/udlf-espresso
[2] https://github.com/UDLF/UDLF
[3] https://github.com/UDLF/pyUDLF

Table 1. Summary of datasets used in the experiments. BEIR datasets use standard relevance judgments (qrels); category-based datasets treat every document as a query and define relevance by shared category membership (excluding the query itself).

Dataset	#Docs	#Queries	Avg. Rel/Q	Eval. Type	Domain
ArguAna	8,674	1,406	1.0	Qrels	Argumentation
SciFact	5,183	300	1.1	Qrels	Scientific Claims
NFCorpus	3,633	323	38.2	Qrels	Biomedical
SciDocs	25,657	1,000	4.9	Qrels	Scientific Docs
FiQA	57,638	648	2.6	Qrels	Finance
BBC-News	2,225	2,225†	450.6	Category (5)	News
WOS	46,985	46,985†	389.0	Category (134)	Academic
HuffPost	32,730	32,730†	2,044.0	Category (42)	News
Mental-Health	20,353	20,353†	5,593.8	Category (4)	Health Forums

† Every document serves as both query and potential result. Avg. Rel/Q for category-based datasets counts same-category documents (excluding the query itself).

Table 2. Distribution of relevant documents per query (Rel/Q). For BEIR datasets, Rel/Q is the number of judged-relevant documents per query from official qrels. For category-based datasets, Rel/Q counts same-category documents excluding the query itself; percentiles are weighted by category size.

Dataset	Eval. Type	Mean	σ	Min	Median	P75	P95	Max
ArguAna	Qrels	1	0	1	1	1	1	1
SciFact	Qrels	1.1	0.5	1	1	1	2	5
NFCorpus	Qrels	38.2	71.7	1	16	39	158	475
SciDocs	Qrels	4.9	0.3	2	5	5	5	5
FiQA	Qrels	2.6	2.1	1	2	3	7	15
BBC-News	Category (5)	450.6	55	385	416	509	510	510
WOS	Category (134)	389	119.5	0	383	417	651	749
HuffPost	Category (42)	2,044	1,865.3	124	1,744	3,539	5,593	5,593
Mental-Health	Category (4)	5,593.8	1,386.8	2,581	5,451	7,053	7,053	7,053

4.2. Experimental Protocol and Parameters

Query construction and re-ranking input. For BEIR datasets, both the dedicated topic queries and every corpus document are issued as queries against the index; when an identifier appears simultaneously as a topic and as a document, the document text is preferred so that the resulting query set is identical across all retrieval models. For category-based datasets every document in the collection serves as a query. In both cases each query retrieves the top-L most similar documents (L=200). Because every document is also queried, the ranked-list file produced in Section 3.2 encodes the full retrieval neighborhood of every corpus element—a property essential for contextual re-ranking methods that exploit inter-document relationships across multiple ranked lists simultaneously.

Metrics and experiment grid. Effectiveness is measured with MAP, Precision, and Recall at cutoffs $k \in \{5, 10, 20, 50, 100, 200\}$. For BEIR datasets relevance is determined by the official qrels; for category-based datasets a document is relevant if and only if it shares the query's category label. Each re-ranking method exposes a neighborhood-size parameter K; we sweep $K \in \{1, 3, 5, 10, 20, 30, 40, 50, 75\}$ for every combination of dataset and retrieval model, yielding $9 \times 4 \times 3 \times 9 = 972$ re-ranking experiments. Baseline results (before re-ranking) are recorded under the same cutoffs for direct comparison.

4.3. Retrieval Results

Table 3 presents a compact view of the results distilled from the full suite of 972 re-ranking experiments. The table reports only the single best-performing method, with the neighborhood-size parameter K shown in parentheses chosen as the value that maximises relative P@20 improvement over the corresponding baseline. Baseline and re-ranked scores are listed side by side, with the relative change ($\Delta\%$) appended to each re-ranked cell so that improvements and regressions can be read at a glance.

Table 3. Re-ranking effectiveness across all datasets. For each retrieval model the single best UDLF method and K are selected by highest P@20 improvement.

			Baseline			Re-ranked ($\Delta\%$)		
Dataset	Model	Method (K)	MAP	P@20	Recall	MAP	P@20	Recall
ArguAna	BM25†	RDPAC (30)	0.2191	0.0418	0.8357	**0.2857** (+30.4%)	**0.0464** (+11.1%)	**0.9282** (+11.1%)
	MiniLM	CPRR (1)	0.2988	0.0484	0.9680	**0.3086** (+3.3%)	**0.0486** (+0.4%)	**0.9723** (+0.4%)
	BGE	CPRR (1)	0.2988	0.0484	0.9680	**0.3086** (+3.3%)	**0.0486** (+0.4%)	**0.9723** (+0.4%)
SciFact	BM25†	BFS-TREE (20)	0.4812	0.0390	0.7083	**0.4853** (+0.8%)	**0.0415** (+6.4%)	**0.7411** (+4.6%)
	MiniLM	CPRR (5)	0.6005	0.0475	0.8373	0.5619 (-6.4%)	**0.0490** (+3.2%)	**0.8610** (+2.8%)
	BGE	BFS-TREE (30)	0.6791	0.0497	0.8753	**0.6851** (+0.9%)	**0.0508** (+2.3%)	**0.8933** (+2.1%)
	Contriever	BFS-TREE (20)	0.6128	0.0483	0.8602	**0.6151** (+0.4%)	**0.0493** (+2.1%)	**0.8650** (+0.6%)
NFCorpus	BM25	CPRR (30)	0.0817	0.1094	0.1147	0.0590 (-27.7%)	**0.1211** (+10.6%)	**0.1215** (+5.9%)
	MiniLM†	BFS-TREE (20)	0.1226	0.1782	0.1889	**0.1353** (+10.4%)	**0.2019** (+13.3%)	**0.2071** (+9.6%)
	BGE†	BFS-TREE (20)	0.1226	0.1782	0.1889	**0.1353** (+10.4%)	**0.2019** (+13.3%)	**0.2071** (+9.6%)
	Contriever	BFS-TREE (30)	0.1261	0.1701	0.1789	**0.1370** (+8.7%)	**0.1904** (+11.9%)	**0.2002** (+11.9%)
SciDocs	BM25†	CPRR (20)	0.0587	0.0326	0.1326	**0.0637** (+8.5%)	**0.0390** (+19.7%)	**0.1589** (+19.8%)
	MiniLM	BFS-TREE (20)	0.1432	0.0774	0.3137	**0.1487** (+3.8%)	**0.0796** (+2.8%)	**0.3228** (+2.9%)
	BGE	BFS-TREE (20)	0.1432	0.0774	0.3137	**0.1487** (+3.8%)	**0.0796** (+2.8%)	**0.3228** (+2.9%)
	Contriever	BFS-TREE (20)	0.1432	0.0774	0.3137	**0.1487** (+3.8%)	**0.0796** (+2.8%)	**0.3228** (+2.9%)
FiQA	BM25†	BFS-TREE (10)	0.0939	0.0212	0.1950	**0.0957** (+1.9%)	**0.0235** (+10.5%)	**0.2091** (+7.2%)
	MiniLM	BFS-TREE (40)	0.2744	0.0576	0.4777	**0.2806** (+2.2%)	**0.0590** (+2.4%)	**0.4821** (+0.9%)
	BGE	BFS-TREE (30)	0.2915	0.0577	0.4695	**0.3008** (+3.2%)	**0.0597** (+3.5%)	**0.4851** (+3.3%)
	Contriever	BFS-TREE (30)	0.2421	0.0512	0.4350	**0.2470** (+2.0%)	**0.0530** (+3.5%)	**0.4377** (+0.6%)
WOS	BM25	BFS-TREE (10)	0.0171	0.3793	0.0220	0.0171 (-0.0%)	0.3792 (-0.0%)	0.0220 (-0.0%)
	MiniLM	CPRR (75)	0.0172	0.4164	0.0243	**0.0194** (+13.0%)	**0.4313** (+3.6%)	**0.0253** (+3.8%)
	BGE	CPRR (75)	0.0170	0.4153	0.0242	**0.0198** (+17.0%)	**0.4408** (+6.1%)	**0.0258** (+6.5%)
	Contriever†	CPRR (75)	0.0146	0.3649	0.0213	**0.0190** (+30.2%)	**0.4195** (+15.0%)	**0.0246** (+15.6%)
BBC-News	BM25†	CPRR (50)	0.0291	0.3990	0.0179	**0.0322** (+10.7%)	**0.4297** (+7.7%)	**0.0193** (+7.6%)
	MiniLM	CPRR (75)	0.0319	0.8574	0.0385	**0.0337** (+5.8%)	**0.8826** (+2.9%)	**0.0397** (+3.1%)
	BGE	CPRR (75)	0.0321	0.8603	0.0386	**0.0345** (+7.5%)	**0.8989** (+4.5%)	**0.0404** (+4.8%)
	Contriever	CPRR (75)	0.0325	0.8665	0.0390	**0.0351** (+7.9%)	**0.9076** (+4.7%)	**0.0409** (+4.9%)
HuffPost	BM25	CPRR (30)	0.0011	0.1746	0.0022	**0.0012** (+8.0%)	**0.1761** (+0.8%)	**0.0023** (+4.2%)
	MiniLM	CPRR (75)	0.0029	0.3237	0.0049	**0.0035** (+21.8%)	**0.3355** (+3.6%)	**0.0054** (+9.6%)
	BGE†	CPRR (75)	0.0029	0.3270	0.0050	**0.0038** (+32.9%)	**0.3579** (+9.5%)	**0.0059** (+17.8%)
	Contriever	CPRR (75)	0.0034	0.3578	0.0057	**0.0042** (+24.6%)	**0.3761** (+5.1%)	**0.0063** (+11.5%)
Mental-Health	BM25	CPRR (75)	0.0011	0.4553	0.0018	**0.0013** (+24.7%)	**0.5137** (+12.8%)	**0.0020** (+13.7%)
	MiniLM	CPRR (75)	0.0015	0.5619	0.0022	**0.0016** (+11.0%)	**0.5907** (+5.1%)	**0.0024** (+5.2%)
	BGE	CPRR (75)	0.0015	0.5544	0.0022	**0.0016** (+12.6%)	**0.5900** (+6.4%)	**0.0024** (+6.5%)
	Contriever†	CPRR (75)	0.0012	0.4772	0.0019	**0.0016** (+33.5%)	**0.5715** (+19.8%)	**0.0023** (+19.0%)

Bold = improvement over baseline. † = highest P@20 relative gain per dataset. $\Delta\%$ = relative change from baseline.

4.4. Qualitative Analysis

We qualitatively analyse the effect of re-ranking through two complementary lenses. A ranked-list illustration in Figure 2 shows a concrete example of how a single query's top-5 result list changes after UDLF re-ranking, mirroring what an end-user would experience. Each slot in the list displays a document identifier colour-coded by category: green for same-category (relevant) documents and red for cross-category (non-relevant) ones. In the baseline ranking relevant and non-relevant documents are interleaved, whereas after re-ranking the relevant documents are concentrated toward the top positions.

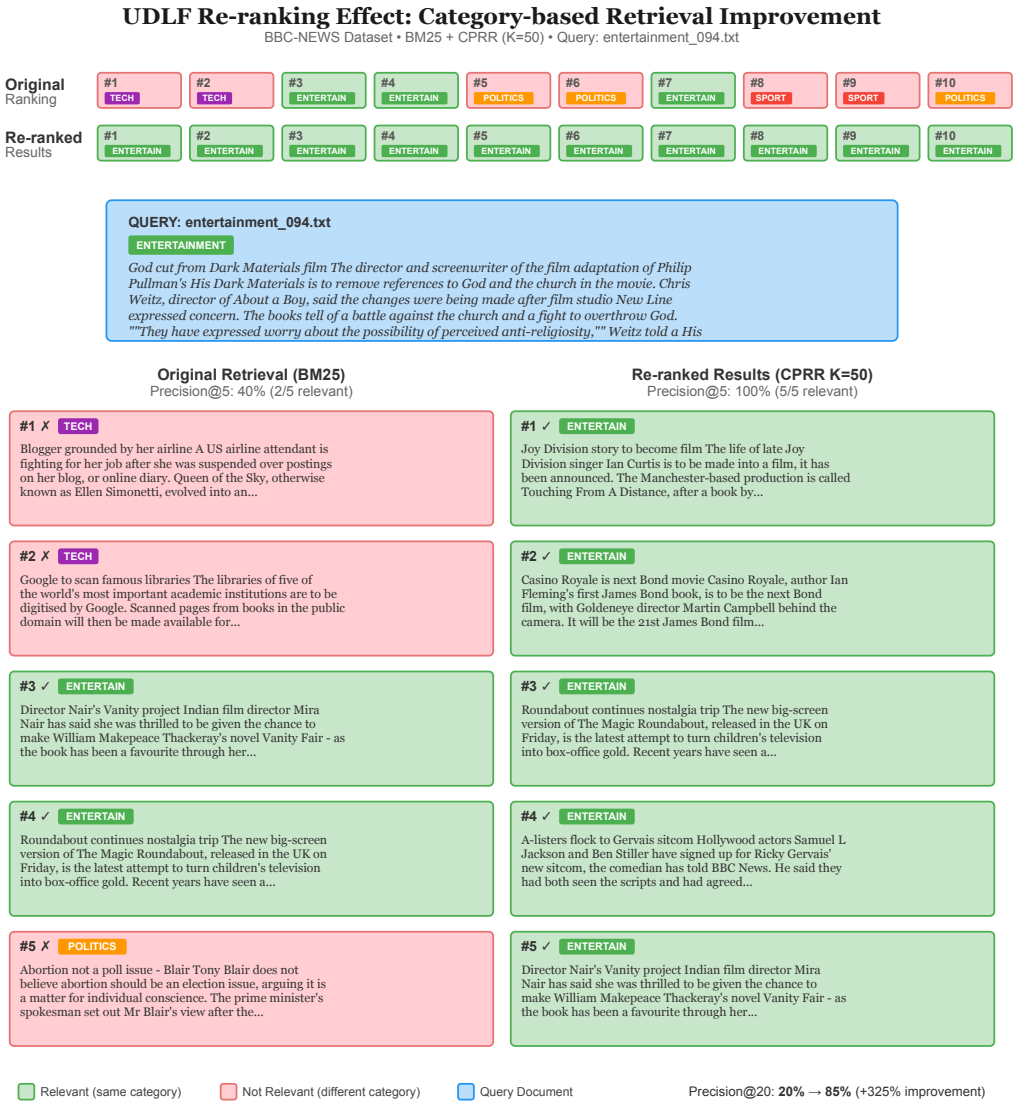

Figure 2. Retrieval and re-ranked results for a single query on BBC-News (BM25 + CPRR, K=50). Green slots denote same-category (relevant) documents; red slots denote cross-category results.

We present a manifold in Figure 3 that offers a geometric perspective on the same

phenomenon. All corpus documents are projected to two dimensions via UMAP and coloured by category; for a set of selected queries, the top-K neighborhood is highlighted before and after re-ranking on a shared viewport. In the *before* panel, the highlighted documents often span multiple category clusters, reflecting the noise in the initial retrieval. In the *after* panel, the highlighted neighborhood contracts toward the query's own category cluster, visually confirming that the re-ranker promotes intra-category coherence.

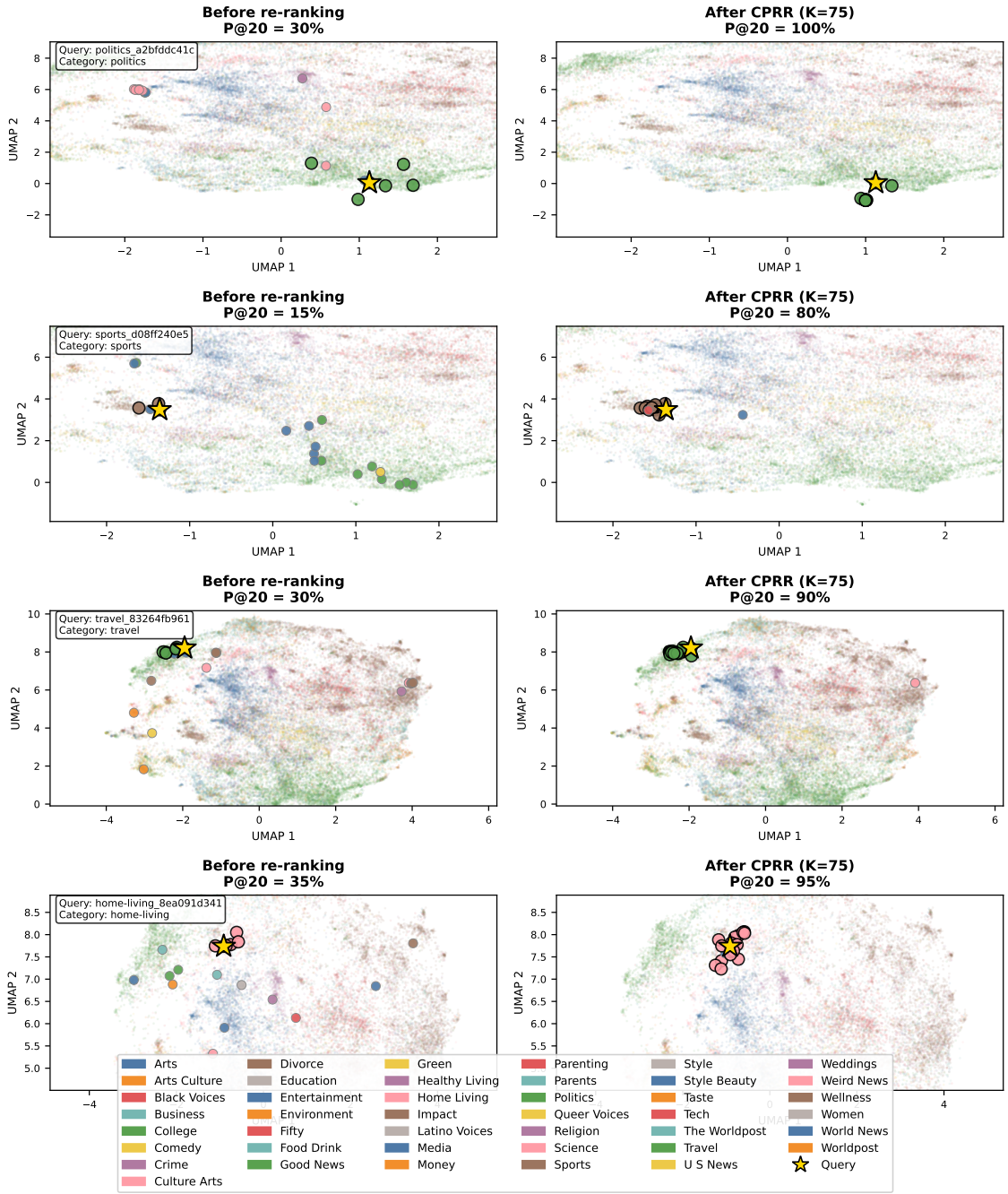

Figure 3. UMAP projection of HuffPost corpus with top-K neighborhood highlighted before and after CPRR re-ranking (K=75, Contriever).

4.5. Discussions and Limitations

The results in Table 3 provide consistent evidence that exploiting contextual neighborhood structure, rather than relying solely on independent query–document similarity, yields measurable ranking improvements across all nine datasets. Gains are observed for every retrieval paradigm: sparse BM25 rankings benefit from re-ranking just as dense embeddings do, often with complementary patterns, BM25 achieves the largest relative P@20 improvement on ArguAna (+11.1%) and Mental-Health (+12.8%), while dense models such as Contriever reach up to +19.8% on the same Mental-Health corpus and +15.0% on WOS. These findings suggest that the contextual signal captured by manifold-based diffusion is largely orthogonal to the information already encoded in the first-stage representation, whether lexical or neural.

A noteworthy aspect of this study is the breadth and diversity of the evaluation setting. The nine datasets span scientific, financial, biomedical, news, and health domains, with corpus sizes ranging from roughly 2 K to 58 K documents and relevance structures varying from single-relevant BEIR judgments (ArguAna) to large same-category pools exceeding 5 000 documents per query (Mental-Health). Despite this heterogeneity, all three re-ranking methods, originally developed and validated exclusively in the image retrieval literature, transfer to text without any modality-specific adaptation, since they consume only ordinal rank positions. This modality-agnostic property is, to our knowledge, demonstrated here for the first time at this scale for text retrieval, opening a path for the same techniques to be applied in further information-retrieval domains such as code search, legal retrieval, or multimodal settings.

Regarding the behavior of individual methods, no single technique uniformly dominates. CPRR tends to excel on category-based corpora with large relevant pools (BBC-News, HuffPost, Mental-Health, WOS), where dense neighborhood overlap among same-category documents amplifies the Cartesian-product affinity signal. BFS-Tree, conversely, proves the most effective on BEIR datasets with sparse and heterogeneous relevance judgments (SciFact, NFCorpus, FiQA, SciDocs), likely because its correlation-gated expansion is more conservative and avoids over-smoothing when few relevant documents exist. RDPAC shows its strength on ArguAna's BM25 ranking, where the convergence-guaranteed random walk recovers a substantial fraction of missed relevant documents (+30.4% MAP). We also note that occasional MAP regressions accompany P@20 improvements (e.g., NFCorpus BM25), indicating that contextual diffusion may redistribute relevance mass across cutoffs—a trade-off that warrants further investigation with per-query analysis.

The implementation of this study required substantial engineering beyond the re-ranking algorithms themselves, including format conversion between retrieval outputs and UDLF inputs, unified query construction for overlapping identifiers, and deterministic pipeline orchestration; these details are documented in the project repository[4] rather than in the pipeline figure (Figure 1), which focuses on the conceptual workflow.

From a practical standpoint, the re-ranking window was fixed at $L{=}200$ and the neighborhood parameter K was swept but not subjected to a validation-based selection protocol; a deeper exploration of these hyperparameters, as well as the inclusion of ad-

[4] `https://github.com/unesp-text-retrieval/udlf-espresso`

ditional re-ranking approaches, would further clarify the boundaries of applicability. Our primary effort was directed toward making large-scale experimentation convenient and reproducible, which motivated the publicly available pipeline and its configuration abstractions. Regarding computational cost, we did not conduct a dedicated efficiency analysis; however, prior work has investigated the runtime characteristics of the methods employed here [Valem et al. 2015, Pisani et al. 2017, Pisani et al. 2020], providing initial reference points for practitioners considering deployment.

The evaluation design also warrants discussion regarding potential bias. Category-based datasets, where relevance is defined by shared category membership, inherently produce well-defined neighborhood structures that amplify the contextual signal exploited by the proposed methods. Nevertheless, the five standard BEIR benchmarks—with sparse and heterogeneous relevance judgments—also exhibit positive results, indicating that the improvements are not confined to settings structurally favorable to neighborhood propagation and supporting applicability to more generic retrieval scenarios.

5. Conclusions

This study demonstrated that unsupervised contextual re-ranking methods can be effectively transferred to text retrieval with consistent effectiveness gains across nine diverse datasets and four retrieval models, totalling 972 experiments. Through a systematic sweep of the neighborhood parameter K, we identified configurations that improved Precision@20 in the vast majority of dataset–model combinations, with relative gains reaching up to +19.8% (Mental-Health, Contriever) and +30.4% in MAP (ArguAna, BM25), confirming that manifold-level signals complement both sparse and dense first-stage retrievers. These results pave the way for broader adoption of rank-based contextual re-ranking in text IR; however, important challenges remain. Scaling to million-document collections such as MS MARCO will require efficiency improvements in the diffusion and graph-construction steps, as well as approximate neighborhood strategies. Future work should also investigate which dataset properties—relevance density, category granularity, corpus homogeneity—best determine the relative advantage of each method, aiming toward principled method-selection guidelines rather than exhaustive grid search. A natural next step is to compare these unsupervised methods against supervised cross-encoder architectures, explicitly assessing the trade-off between effectiveness gains and the requirement of labelled training data. Furthermore, combining the outputs of multiple re-ranking methods through rank aggregation strategies may further improve robustness across diverse settings. Finally, extending the evaluation to additional retrieval domains, including cross-lingual and multimodal scenarios, will further test the generality of the modality-agnostic hypothesis supported by our findings.

6. Acknowledgments

The authors are grateful to the National Council for Scientific and Technological Development — CNPq (grant #313193/2023-1), the São Paulo Research Foundation — FAPESP (grant #2024/04890-5 and #2025/07171-2), and Petrobras (grants #2025/00498-6 and #2023/00095-3) for their financial support.

References

Ai, Q., Bi, K., Guo, J., and Croft, W. B. (2018). Learning a deep listwise context model for ranking refinement. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 135–144.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing Management*, 39(1):45–65.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Guimarães Pedronette, D. C., Pascotti Valem, L., and Latecki, L. J. (2021). Efficient rank-based diffusion process with assured convergence. *Journal of Imaging*, 7(3).

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

Lavrenko, V. and Croft, W. B. (2003). Relevance models in information retrieval. In *Language modeling for information retrieval*, pages 11–56. Springer.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Pedronette, D. C. G., Valem, L. P., and da S. Torres, R. (2021). A bfs-tree of ranking references for unsupervised manifold learning. *Pattern Recognition*, 111:107666.

Pisani, F., Pascotti Valem, L., Guimarães Pedronette, D. C., da S. Torres, R., Borin, E., and Breternitz Jr., M. (2020). A unified model for accelerating unsupervised iterative re-ranking algorithms. *Concurrency and Computation: Practice and Experience*, 32(14):e5702.

Pisani, F., Pedronette, D. C. G., Torres, R. d. S., and Borin, E. (2017). Contextual spaces re-ranking: accelerating the re-sort ranked lists step on heterogeneous systems. *Concurrency and Computation: Practice and Experience*, 29(22):e3962. e3962 cpe.3962.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994). Okapi at trec-3. pages 0–.

Rocchio Jr, J. J. (1971). Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4):288–297.

Valem, L. P. and Guimarães Pedronette, D. C. (2016). Unsupervised similarity learning through cartesian product of ranking references for image retrieval tasks. In *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 249–256.

Valem, L. P., Pedronette, D. C. G. a., Torres, R. d. S., Borin, E., and Almeida, J. (2015). Effective, efficient, and scalable unsupervised distance learning in image retrieval tasks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, page 51–58, New York, NY, USA. Association for Computing Machinery.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xu, Z., Mo, F., Huang, Z., Zhang, C., Yu, P., Wang, B., Lin, J., and Srikumar, V. (2025). A survey of model architectures in information retrieval. *arXiv preprint arXiv:2502.14822*.

Zerveas, G., Rekabsaz, N., Cohen, D., and Eickhoff, C. (2022). Coder: An efficient framework for improving retrieval through contextual document embedding reranking. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 10626–10644.

Zerveas, G., Rekabsaz, N., and Eickhoff, C. (2023). Enhancing the ranking context of dense retrieval through reciprocal nearest neighbors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10779–10803.

Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. (2024). Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4).