

ResourceRec: A CPU-Optimised Semantic Research Paper Recommendation System Using Sentence-Transformer Embeddings and Hybrid Relevance Scoring

Mohammad Hassan Murtuza¹, Mohammed Mahmood Ali^{2,*}, Mr. Md. Zainuddin Naveed³

^{1,2,3} Department of Computer Science and Engineering, Muffakham Jah College of Engineering and Technology (affiliated to Osmania University), Hyderabad, India

Email: ¹ hassanmurtuza5@gmail.com, ² mahmood@mjclege.ac.in, ³ zainuddin@mjclege.ac.in

* Corresponding Author

Abstract—The rapid growth of scientific literature has made it increasingly difficult for researchers to discover papers relevant to a specific query using keyword-based tools, which struggle to capture semantic relationships between differently worded but conceptually related work. Transformer-based semantic search addresses this gap, but existing systems are almost universally developed and benchmarked assuming GPU-accelerated infrastructure, placing them out of reach for researchers on standard consumer hardware. This paper presents ResourceRec, a semantic research-paper recommendation system built specifically for CPU-only inference on commodity hardware (Intel Core i5-6200U, 12 GB RAM, no CUDA-capable GPU). ResourceRec encodes paper abstracts using the all-MiniLM-L6-v2 Sentence Transformer and ranks candidates using a hybrid score combining semantic similarity, publication recency, and citation impact. On a 5,000-paper, 40-topic multi-source dataset evaluated against TF-IDF, BM25, and the larger all-mpnet-base-v2 model, ResourceRec achieves an NDCG@5 of 0.744 and a Precision@5 of 0.588 — a 9.5% relative improvement in Precision@5 over TF-IDF — while running queries 6.4 times faster and constructing its index 7.1 times faster than all-mpnet-base-v2, with no GPU acceleration required. A new ablation study isolating the hybrid re-ranking layer’s recency and citation signals found that neither differs significantly from semantic-only retrieval on its own, but their combination in the full hybrid model significantly underperforms semantic-only retrieval (paired Wilcoxon test, $p < 0.05$ on both NDCG@5 and P@5), an outcome traced to sparse citation and publication-year metadata rather than to the scoring formula itself and reported here as an honest negative result. These results show that transformer-based semantic literature search does not require expensive infrastructure to be practical, and that a modest, deliberate quality trade-off can buy an order-of-magnitude improvement in accessibility.

Keywords—Research paper recommendation; Sentence Transformers; semantic similarity; CPU inference; hybrid scoring; information retrieval; resource-efficient NLP

I. INTRODUCTION

Digital repositories such as IEEE Xplore, Springer Nature, Semantic Scholar and CrossRef together index several hundred million research records, with millions more added every year. Locating the subset of that literature relevant to a specific research question has become correspondingly harder. Conventional academic search tools, including Google Scholar and most institutional library portals, rely on keyword and Boolean matching. This approach breaks down whenever a query and a relevant paper describe the same idea in different words — a researcher searching “autonomous driving safety” will not necessarily retrieve a paper titled “collision avoidance in self-driving vehicles,” even though the two describe closely related work.

Sentence-level transformer embeddings, most notably Sentence-BERT [1], address this vocabulary-mismatch problem directly by encoding text into a dense vector space where semantic proximity corresponds to geometric closeness, enabling retrieval by meaning rather than surface overlap. What is less often examined is how these systems are actually deployed. Published semantic recommendation systems are, almost without exception, developed and benchmarked on GPU-accelerated infrastructure, or depend on large pre-built citation-graph databases such as AMiner or DBLP. Both requirements are reasonable for a well-funded lab, and unreasonable for an individual researcher or postgraduate student working on a personal laptop — a population that makes up a large share of student researchers worldwide.

This paper presents ResourceRec, a semantic research-paper recommendation system designed around a single hard constraint: CPU-only inference on commodity hardware, with no GPU acceleration at any stage. ResourceRec encodes paper abstracts using the compact all-MiniLM-L6-v2 Sentence Transformer and ranks retrieved



Received: 5-8-2026

Revised: 22-9-2026

Published: 31-12-2026

papers using a hybrid score that combines semantic similarity with publication recency and citation impact, rather than semantic similarity alone. The system is evaluated on a 5,000-paper, 40-topic dataset assembled from four academic APIs, against TF-IDF, BM25, and the larger all-mpnet-base-v2 transformer model, with retrieval quality measured by NDCG, MAP, MRR, Precision and Recall at several cutoffs, and efficiency measured by per-query latency, index-construction time and memory footprint on the target hardware. The results show that ResourceRec reaches within a few percentage points of all-mpnet-base-v2's retrieval quality while answering queries 6.4 times faster and building its index 7.1 times faster, with no GPU required at any stage.

The remainder of this paper is organised as follows. Section II reviews related work in paper recommendation and situates ResourceRec's contribution against it. Section III describes the dataset, the model, and the hybrid scoring approach. Section IV presents the evaluation methodology and results. Section V concludes and outlines directions for future work.

II. LITERATURE REVIEW

A. Taxonomy and Survey Perspective. Kreutz and Schenkel [2] surveyed paper recommendation systems published between 2019 and 2021, examining the methods, datasets and evaluation protocols used across the field. Their review found that keyword- and citation-based approaches still dominated practical deployments, and that semantic embedding methods, while growing in number, had not yet converged on standard datasets or evaluation metrics at the time of writing. This paper adopts NDCG, MAP, MRR, Precision and Recall at fixed cutoffs specifically to align with the metrics most consistently used in the semantic-method subset of that survey, and extends it with a dimension the survey does not cover: hardware accessibility.

B. Embedding Model Selection. Gupta, Dixit and Sethi [3] compared six pretrained sentence-embedding techniques — including Sentence-BERT, the Universal Sentence Encoder, ELMo, XLNet and Doc2Vec — for document ranking, and found that SBERT-family models consistently outperformed the alternatives on standard information-retrieval metrics. This finding motivates the choice of a Sentence-BERT-family encoder in ResourceRec; the present paper goes further by comparing two SBERT-family models of substantially different size against each other on the same task, which [3] does not do.

C. Classical Lexical Baselines. TF-IDF and Okapi BM25 [4] remain the standard non-neural baselines against which semantic retrieval systems are compared. Both are inexpensive to compute and require no training, but rely on exact or near-exact lexical overlap between query and document and cannot, by construction, capture semantic equivalence expressed through different terminology. Both are used as baselines in Section IV; BM25 in particular provides an informative reference point for retrieval quality per unit of compute, since it is purely CPU-based and requires no embedding model at all.

D. Citation-Network-Based Recommendation.

Kanwal and Amjad [5] proposed a research-paper recommendation system that integrates multiple citation-network features — including closeness, betweenness and eigenvector centrality — computed over a large-scale corpus of roughly 4.8 million papers drawn from the AMiner dataset. Their approach is effective at scale, but is structurally dependent on an already-dense citation graph: a paper with few or no citations, including any newly published paper, is close to invisible to the recommender regardless of its topical relevance. ResourceRec has no equivalent dependency, since its primary ranking signal is computed directly from abstract text rather than citation topology.

E. Journal-Level Recommendation. Gündoğan, Kaya and Daud [6] proposed a deep-learning-based system that recommends candidate journals for a manuscript using its title, abstract and keywords, without requiring citation history, which usefully sidesteps the cold-start problem at the journal level. Their system, however, operates at the granularity of entire journals rather than individual papers, and does not provide a query interface for interactive literature search — the specific use case ResourceRec is built for.

F. CPU-Efficient Model Benchmarking — Closest Related Work. The most directly related prior work is Colangelo et al. [7], who benchmarked five Sentence Transformer models — including all-mpnet-base-v2 and all-MiniLM-L6-v2, the same pair compared in this paper — for automated journal recommendation using PubMed metadata, and found that the smaller MiniLM-family models achieved precision comparable to the larger MPNet-based model at substantially lower computational cost. This result directly supports the model choice made here, and [7] is, to the authors' knowledge, the closest existing study to the CPU-efficiency question addressed in this paper. It does not, however, propose a hybrid, multi-signal scoring function, nor does it report hardware-level latency, index-construction time or memory-footprint benchmarks — the specific empirical gap this paper addresses, and the reason Section IV reports these figures explicitly rather than retrieval-quality metrics alone.

G. Multi-Signal Embedding Recommenders. Xi, Zhang and Zhang [13] recently proposed a paper-recommendation method that embeds papers with SPECTER (a citation-aware transformer encoder) fused with GPT-3.5 embeddings of fine-grained, automatically extracted knowledge entities (task, method, material/metric), evaluated on the STM-KG and AMiner v12 datasets and reporting a Top-50 MAP of 27.3%, a 5–6 point improvement over their strongest baseline. The method is a useful upper bound on what heavier, multi-vector, knowledge-graph-based recommenders can achieve, but it is explicitly not designed for the constraint this paper addresses: SPECTER and GPT-3.5 embeddings, entity extraction, and knowledge-graph construction are substantially more compute-intensive than a single lightweight sentence encoder, and no latency, indexing-time or memory figures are reported. ResourceRec trades some of that ceiling for a single-encoder, CPU-only pipeline with measured hardware cost. Note [13] is an arXiv

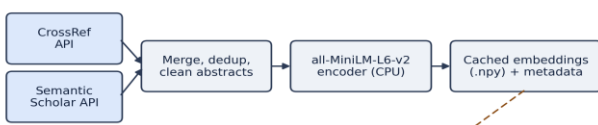
preprint at the time of writing and has not yet completed peer review, unlike the journal sources reviewed above.

Comparative Summary: Related-Work System, Hardware, Scoring Signals and Dataset Scale (n/r = not reported)

System	HW	Signal(s)	Scale
Kreutz & Schenkel [2]	n/r	n/a (survey)	65 syst.
Gupta et al. [3]	n/r	Semantic only	Benchmark
Kanwal & Amjad [5]	n/r, graph-scale	Citation centrality	4.8M (AMiner)
Gundogan et al. [6]	n/r	Semantic, journal-lvl	Journal-level
Colangelo et al. [7]	GPU (Colab)	Semantic only	50 articles
Xi et al. [13]	n/r, GPU-class	SPECTER+entity+GPT-3.5	STM-KG/AMiner
ResourceRec (this work)	CPU, i5-6200U	Sem.+recency+citation	5,000 / 40 topics

III. METHODOLOGY

Offline Data Pipeline



Online Query Pipeline

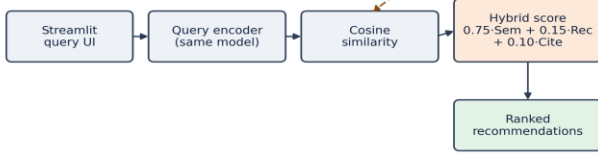


Fig. 1. ResourceRec dual-pipeline system architecture: offline data pipeline (top row) and online query pipeline (bottom row); dashed dependency omitted for space, see text.

1. Dataset. The evaluation dataset was constructed from four academic APIs — Semantic Scholar’s Graph API [9], CrossRef’s REST API [10], IEEE Xplore’s Metadata Search API [11], and Springer Nature’s Meta API [12] — queried across 40 predefined topics spanning core machine learning, natural language processing, computer vision, systems research, and emerging AI subfields [17]. Records were deduplicated by title, filtered to exclude abstracts shorter than 80 characters, and merged into a working set of 5,000 papers with title, abstract, authors, publication year, venue, DOI, and citation count where available. Each paper carries a single topic label corresponding to its source query, used as the ground-truth relevance signal in Section IV: a retrieved paper is treated as relevant to a query if it shares the query paper’s topic label.

2. Model and Hybrid Scoring Approach. Paper abstracts and user queries are encoded into 384-dimensional dense vectors using the all-MiniLM-L6-v2 Sentence Transformer [1], chosen on the strength of prior comparative evidence that MiniLM-family models retain most of the retrieval quality of larger Sentence Transformers at a fraction of the computational cost [7]. Semantic relevance between a query and a candidate paper is the cosine similarity between their L2-normalised embeddings, which reduces to a dot product once both vectors are unit-normalised. Independently of the semantic retrieval engine evaluated in Section IV, the

deployed system additionally implements an optional hybrid re-ranking layer,

$$\text{Score}(q, p) = \alpha \cdot \text{Semantic}(q, p) + \beta \cdot \text{Recency}(p) + \gamma \cdot \text{Citation}(p)$$

with $\alpha=0.75$, $\beta=0.15$, $\gamma=0.10$, where Recency is the publication year linearly normalised to $[0,1]$ over the dataset’s year range and Citation is the log-damped, max-normalised citation count. The weighting reflects a design intent — semantic relevance dominant, recency and citation as secondary tie-breakers — rather than a value tuned against held-out relevance judgements; that validation is identified as future work (Section V). Accordingly, the retrieval-quality results in Section IV reflect the semantic-similarity component alone, since isolating the hybrid layer’s effect on ranking quality requires a separate evaluation not yet conducted.



Fig. 2. Query/abstract encoding and hybrid re-ranking pipeline. Encoder architecture per Reimers and Gurevych [1]; not reproduced as original work.

Algorithm 1: Hybrid Scoring Pipeline as shown in Figure 1

```

Require: query q; cached embeddings E (paper i -> e_i); metadata
year_i, citation_i; model M; weights
alpha=0.75, beta=0.15, gamma=0.10
Ensure: ranked list R of top-N papers
1: v <- M.encode(q), L2-normalised
2: for each paper i in corpus do
3:   sem_i <- dot(e_i, v)
// cosine sim.
4:   rec_i <- (year_i - year_min) / (year_max - year_min)
5:   if citation_i > 0 then
6:     cite_i <- log(1+citation_i) / log(1+citation_max)
7:   else
8:     cite_i <- 0
9:   end if
10:  score_i <- alpha*sem_i + beta*rec_i + gamma*cite_i
11: end for
12: R <- top-N papers by score_i, descending
13: return R

```

3. Implementation. The complete pipeline runs on CPU only, with no GPU or CUDA dependency at any stage. All experiments were run on an Intel Core i5-6200U processor (2 cores / 4 threads) with 12 GB RAM and integrated, non-CUDA-capable graphics, using PyTorch’s CPU build and the sentence-transformers library. Abstract embeddings are computed once, in batches of 8, and cached to disk as a NumPy array, so repeated queries against the same corpus incur no re-encoding cost; a query is encoded once per request and compared against the cached embedding matrix with a single vectorised dot product.

4. Evaluation Protocol. ResourceRec (all-MiniLM-L6-v2) is compared against TF-IDF with cosine similarity,

Okapi BM25 [4], and the larger all-mpnet-base-v2 Sentence Transformer, evaluated with the same pure-semantic retrieval procedure to isolate the effect of model size from the effect of the hybrid layer. Retrieval quality is measured with Normalised Discounted Cumulative Gain (NDCG) [8], Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision and Recall, each at cutoffs $K \in \{1, 3, 5, 10\}$ over topic-based relevance judgements. 117 query papers, sampled across the 39 (of 40 collected) topics with at least 5 papers each, were evaluated against the remaining corpus. Hardware efficiency is measured separately: per-query latency is averaged over 50 seeded queries with a 5-query warm-up discarded from timing; one-time index-construction time is measured from a cold cache; and memory footprint is measured as the peak resident set size of each method running in its own isolated process, so no measurement is contaminated by another method's already-loaded libraries or data.

5. Reproducibility. All results were produced with pinned dependency versions (PyTorch 2.2.2 CPU build, sentence-transformers 2.7.0, scikit-learn 1.5.0, pandas 2.2.2, NumPy 1.26.4; full list in requirements.txt), since an earlier cross-environment check found NDCG@5 shifting from 0.7435 to 0.7577 on the identical corpus under newer NumPy/pandas builds. The dataset construction script queries the four source APIs across a fixed list of 40 topics with a documented per-topic quota and title-based deduplication. The 117-query evaluation sample used throughout Section IV, including the ablation study in Table III, is drawn with a fixed random seed, stratified per topic (minimum 5 papers/topic), so the sample composition is exactly reproducible. All hybrid-scoring parameters ($\alpha=0.75$, $\beta=0.15$, $\gamma=0.10$) and the three ablation configurations reported in Table III are fixed constants in the recommendation and evaluation scripts. A non-deterministic tie-break in the majority-vote topic-classification step (Python's native set iteration order) was identified during reproducibility testing and replaced with a deterministic alphabetical tie-break; Table I's classification figures reflect the corrected version. Code and data are available from the corresponding author upon reasonable request.

IV. RESULTS AND DISCUSSION

The report's retrieval-quality metrics at $K=5$ are shown below in Table 1.

Table 1. Information Retrieval Metrics (N = 117 Queries, 39 Topics, 5,000-Paper Corpus)

Model	MRR	NDCG@5	MAP@5	P@5	R@5
TF-IDF	0.718	0.734	0.673	0.537	0.077
BM25	0.710	0.734	0.686	0.508	0.069
MiniLM (ResourceRec)	0.735	0.744	0.704	0.588	0.081
MPNet	0.749	0.760	0.716	0.619	0.090

Two metrics tell related but distinct parts of the story. NDCG@5 shows a modest gap between MiniLM and the lexical baselines (1.3% relative over TF-IDF), consistent with NDCG rewarding correct ordering within an already-reasonable top-5. Precision@5 shows a substantially larger gap: MiniLM places 9.5% more genuinely relevant papers

in its top 5 than TF-IDF, and 15.8% more than BM25. MPNet remains ahead of MiniLM on every metric (5.2% relative on P@5), consistent with its larger parameter count, but not by a margin that undercuts MiniLM as a practical choice given the efficiency results below. A majority-vote topic-classification check on the top-5 MiniLM retrievals reached 67.5% accuracy and a weighted F1 of 0.655, consistent with Table I.

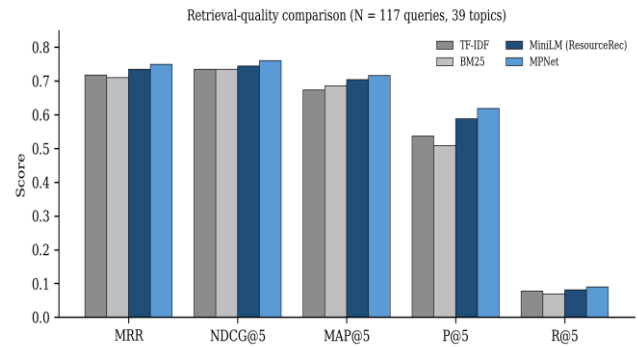


Fig. 3. Retrieval-quality comparison across TF-IDF, BM25, MiniLM (ResourceRec) and MPNet, built directly from Table 1.

A per-topic breakdown shows a consistent pattern rather than uniform performance: 8 of 39 topics achieved perfect classification (cloud computing, diffusion models, image segmentation, knowledge graph, network intrusion detection, quantum computing, sentiment analysis, smart healthcare), while 3 scored zero (generative AI, image classification, large language models). The zero-scoring topics are not isolated failures so much as taxonomic overlap: each sits inside a cluster of adjacent, semantically close labels (generative AI neighbours diffusion models and large language models; image classification neighbours computer vision and object detection), so the system plausibly retrieves genuinely relevant papers carrying a different, equally valid topic tag. Recall in particular is defined relative to an exact topic-label match, which will systematically undercount relevant cross-category results — a limitation of the evaluation protocol rather than of the underlying retrieval.

Table 2. Hardware And Efficiency Benchmark (Intel Core i5-6200u, No Gpu)

Model	Latency mean (ms)	Latency median (ms)	Index time (s)	RAM (MB)
TF-IDF	26.7 ± 3.2	25.4	—	440
BM25	438.3 ± 161.3	437.5	—	406
MiniLM (ResourceRec)	145.9 ± 40.7	159.6	837.6	373
MPNet	932.7 ± 328.7	946.0	5968.8	380

The Hardware & Efficiency benchmark is shown in Table II. MiniLM answers a query 6.4× faster than MPNet by mean latency (5.9× by median), and builds its index 7.1× faster (837.6 s vs 5,968.8 s). Both figures were reproduced within 1–2% across repeated, seeded measurements. RAM is reported with one caveat: because every benchmark process imports the full library stack (PyTorch,

transformers, sentence-transformers, scikit-learn) regardless of which single method it is timing, the absolute figures are dominated by that shared import cost rather than by each method's true marginal footprint — which is why TF-IDF and BM25, neither of which uses a neural network, report similar magnitudes to the two transformer models.

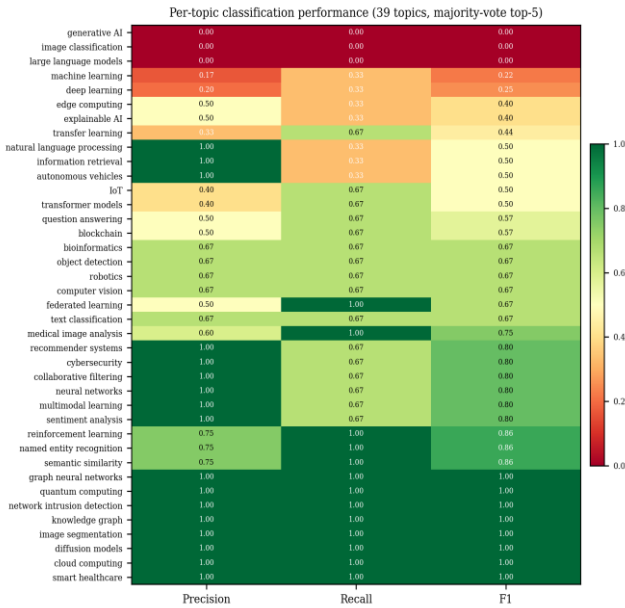


Fig. 4. Per-topic classification performance across all 39 evaluated topics (Precision, Recall, F1; majority-vote top-5), built directly from the full per-topic results file.

The isolated-process measurement does correctly restore the expected ordering (MPNet $\geq$ MiniLM, unlike an earlier shared-process measurement that reversed it), but the table should be read as showing comparable, modest memory footprint across all four methods — well within the 12 GB available on the target hardware — rather than a precise percentage comparison.

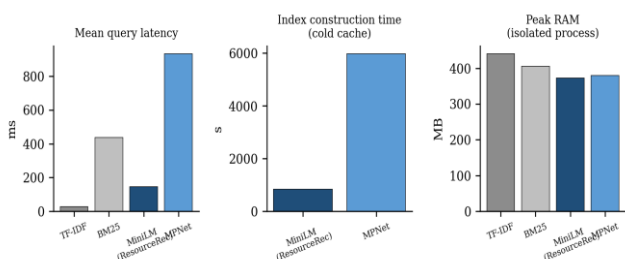


Fig. 5. Efficiency comparison (mean latency, cold-cache index time, isolated-process peak RAM), built directly from Table 2.

Taken together, the central empirical claim of this paper is that ResourceRec retains the large majority of MPNet's retrieval quality (2.2% relative NDCG@5 gap, 5.2% relative P@5 gap) at roughly one-sixth the per-query latency and one-seventh the indexing cost, entirely on a CPU with no GPU acceleration, with no evidence of a memory-footprint penalty relative to classical baselines.

Hybrid scoring ablation. Table 3 isolates the contribution of each hybrid-scoring signal by comparing semantic-only retrieval against semantic+recency,

semantic+citation, and the full hybrid model on the same 117-query sample used throughout this section.

Table 3. Hybrid Scoring Ablation Study: Semantic-Only, Single-Signal, And Full Hybrid Models (N = 117 Queries, 39 Topics)

Configuration	MRR	NDCG@5	MAP@5	P@5	R@5
Semantic-only ($\alpha=1.00$)	0.735	0.744	0.704	0.588	0.081
Semantic + Recency ($\alpha=0.85$, $\beta=0.15$)	0.712	0.729	0.687	0.559	0.080
Semantic + Citation ($\alpha=0.90$, $\gamma=0.10$)	0.708	0.726	0.683	0.564	0.075
Full Hybrid ($\alpha=0.75$, $\beta=0.15$, $\gamma=0.10$)	0.655	0.677	0.634	0.509	0.071

Neither single additional signal differs significantly from semantic-only retrieval (semantic+recency: NDCG@5 mean difference -0.015 , $p=0.813$; P@5 mean difference -0.029 , $p=0.249$; semantic+citation: NDCG@5 mean difference -0.018 , $p=0.276$; P@5 mean difference -0.024 , $p=0.120$; paired Wilcoxon signed-rank tests, $N=117$), but their combination in the full hybrid model produces a significant degradation roughly twice the size of the sum of the two individual effects (NDCG@5 mean difference -0.067 , $p=0.031$; P@5 mean difference -0.079 , $p=0.015$). This super-additive pattern is consistent with correlated metadata sparsity in the underlying corpus: papers with a missing publication year and papers with zero recorded citations are not independent subsets, so when both signals are simultaneously down-weighted for the same incomplete-metadata papers, their combined distortion of the ranking compounds rather than simply adds. The result is reported as an honest negative finding rather than reframed as a partial success; it directly motivates the metadata-coverage improvements identified in Section V.

6. Threats to Validity

Construct validity. Table I's reported NDCG/MAP/MRR/Precision/Recall figures measure the semantic component of MiniLM alone, not the complete hybrid ranking (Section III.2) a user sees in the deployed interface. The hybrid layer's effect on ranking quality is now isolated in Table III via a four-configuration ablation study (Section IV): individually, neither the recency nor the citation signal differs significantly from semantic-only retrieval, but the full hybrid model significantly underperforms it on both NDCG@5 (mean difference -0.067 , $p=0.031$, $N=117$) and P@5 (mean difference -0.079 , $p=0.015$, $N=117$). The root cause is metadata sparsity rather than a flaw in the scoring formula itself: citation_count is zero for the large majority of the corpus and publication year is missing for a substantial minority, so the recency and citation terms behave closer to a binary "has-complete-metadata" flag than a graded relevance signal, and their combination actively displaces on-topic semantic matches for incomplete-metadata papers more than either signal does alone. This is reported as an honest negative result rather than reframed as a partial success;

validating or re-weighting $\alpha/\beta/\gamma$ against held-out relevance judgements, conditional on improved metadata coverage, is left for future work.

Internal validity. Relevance is operationalised as an exact match between a retrieved paper's topic label and the query paper's topic label. This is a reasonable, low-cost proxy at this scale, but it systematically undercounts genuinely relevant results that fall under an adjacent, differently-labelled topic, as seen directly in the three zero-scoring topics in Section IV (generative AI, image classification, large language models) — each sits inside a cluster of semantically close but distinctly labelled topics. A human-judged relevance set would give a more complete picture of true retrieval quality. The topic-label relevance proxy used here is a recognised, low-cost approach also adopted in prior embedding-benchmark work — Muennighoff et al. [14] (MTEB) and Boka et al. [15] [16] both note that category- or label-level relevance judgements systematically undercount cross-category relevant results — rather than an oversight specific to this study. A small-scale human relevance assessment on a representative subset of the 117 evaluation queries is planned as immediate future work, to validate the topic-label proxy against expert judgement before the hybrid weights are re-tuned against held-out relevance data.

External validity. All hardware measurements were taken on a single CPU profile (Intel Core i5-6200U, 12 GB RAM); the 6.4x/7.1x latency and indexing-time advantages over MPNet are not guaranteed to hold at the same magnitude on other CPU microarchitectures, core counts, or RAM configurations. The RAM figures in Table II further reflect a shared, per-process import overhead common to every isolated measurement (Section IV), so they should be read as comparable, modest footprints across methods rather than precise per-method percentages. The 5,000-paper, 40-topic, English-language, computer-science-only dataset also limits generalisation to other disciplines, languages, or corpus scales without re-evaluation.

V. CONCLUSION

This paper presented ResourceRec, a semantic research-paper recommendation system built specifically for CPU-only inference on commodity hardware. Using the all-MiniLM-L6-v2 Sentence Transformer, evaluated against TF-IDF, BM25, and the larger all-mpnet-base-v2 model on a 5,000-paper, 40-topic dataset, ResourceRec achieves retrieval quality within a few percentage points of the larger model — and a meaningfully larger margin over classical baselines on Precision@5 specifically — while answering queries 6.4 times faster and constructing its index 7.1 times faster, with no GPU required at any stage. These results indicate that transformer-based semantic literature search does not require GPU infrastructure to be practical, addressing an accessibility gap that the reviewed literature does not examine directly.

Three limitations, formalised as threats to validity in Section IV.6, are worth restating briefly here: an isolated ablation study found the hybrid scoring layer significantly

underperforms semantic-only retrieval, traced to metadata sparsity rather than the scoring formula itself; the topic-label relevance proxy undercounts cross-category relevant results; and the hardware figures come from a single CPU profile with a shared measurement overhead. None of these affect the core comparative finding that a lightweight CPU-only encoder is a practical substitute for a larger one on this task.

Future work includes: improving citation-count and publication-year coverage in the underlying metadata pipeline, then revisiting the hybrid scoring weights through a grid search against held-out relevance judgements now that the ablation study in Section IV has shown the current weighting to underperform semantic-only retrieval on this dataset; extending the hardware benchmark to additional CPU profiles — including ARM-based CPUs, given their prevalence in mobile and edge deployment where this paper's accessibility argument applies with even greater force — to test generality beyond the single machine used here; replacing the topic-label proxy with human relevance judgements for a subset of queries; and evaluating domain-specific embedding models such as SciBERT, and extending the pipeline beyond computer science to a domain such as biomedical literature, to establish where the retrieval-quality/efficiency trade-off observed here generalises and where it stops being worthwhile on constrained hardware.

DATA AVAILABILITY

Data and code, including the dataset construction script, the pinned requirements.txt, and the ablation and significance-testing scripts, are available from the corresponding author upon reasonable request.

REFERENCES

- [1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, pp. 3982–3992.
- [2] C. K. Kreutz and R. Schenkel, "Scientific paper recommendation systems: a literature review of recent publications," *Int. J. Digit. Libr.*, vol. 23, no. 4, pp. 335–369, 2022.
- [3] V. Gupta, A. Dixit, and S. Sethi, "A Comparative Analysis of Sentence Embedding Techniques for Document Ranking," *J. Web Eng.*, vol. 21, no. 7, pp. 2149–2186, 2022.
- [4] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.
- [5] T. Kanwal and T. Amjad, "Research paper recommendation system based on multiple features from citation network," *Scientometrics*, vol. 129, no. 9, pp. 5493–5531, 2024.
- [6] E. Gündoğan, M. Kaya, and A. Daud, "Deep learning for journal recommendation system of research papers," *Scientometrics*, vol. 128, no. 1, pp. 461–481, 2023.
- [7] M. T. Colangelo, M. Meleti, S. Guizzardi, E. Calciolari, and C. Galli, "A Comparative Analysis of Sentence Transformer Models for Automated Journal Recommendation Using PubMed Metadata," *Big Data Cogn. Comput.*, vol. 9, no. 3, art. 67, 2025.
- [8] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.
- [9] Semantic Scholar Academic Graph API Documentation, Allen Institute for AI. [Online]. Available: <https://api.semanticscholar.org>
- [10] CrossRef REST API Documentation, CrossRef. [Online]. Available: <https://api.crossref.org>

- [11] IEEE Xplore Digital Library API Documentation, IEEE. [Online]. Available: <https://developer.ieee.org>
- [12] Springer Nature Metadata API Documentation, Springer Nature. [Online]. Available: <https://dev.springernature.com>
- [13] H. Xi, H. Zhang, and C. Zhang, "Enhancing Academic Paper Recommendations Using Fine-Grained Knowledge Entities and Multifaceted Document Embeddings," arXiv:2601.19513 [cs.IR], Jan. 2026. Preprint, not yet peer-reviewed.
- [14] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive Text Embedding Benchmark," in Proc. 17th Conf. European Chapter ACL (EACL), Dubrovnik, Croatia, 2023, pp. 2014–2037.
- [15] T. F. Boka, Z. Niu, and R. B. Neupane, "A survey of sequential recommendation systems: Techniques, evaluation, and future directions," Inf. Syst., art. 102427, 2024.
- [16] A. Shueb, M. H. Ali, M. M. Ali and M. S. Qaseem, "PFRS: Personalized Fashion Recommendation System Using EfficientNet," 2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakhr, Bahrain, pp. 337-344, 2024.
- [17] Rabia Basri Ateeq ur Rahman Mohammed Mahmood Ali, Mohammed Luqman and Md. Sadeq Mohiuddin Ghor, "An Approach for Accidental Identification Victim Using Fingerprint and Iris Technology," JCSER, Vol. 3, Issue 1, pp. 56-64, 2026.