CIA Open Access

A Comparative Empirical Study of Semantic Signal Enhancement Methods for User Interest Features in CTR Prediction: Applicability of TF-IDF Weighting, Sentence-BERT Embeddings, and LDA Topic Fusion

***Tianxing Tang*[1], *Mingzhuo Yu*[1,2]**

[1] *Translation and Localization Management, Middlebury Institute of International Studies, CA, USA*
[1,2] *Computer Science, Northeastern University, MA, USA*

DOI: 10.63575/CIA.2024.20114

Abstract

User interest feature engineering in recommendation and advertising platforms increasingly relies on semantic signals derived from item text, user queries, and tag annotations. Practitioners lack a unified empirical comparison of the dominant fusion paths under a shared evaluation protocol. This study reports a comparative empirical analysis of three representative semantic signal enhancement methods — TF-IDF weighting, Sentence-BERT embeddings, and latent Dirichlet allocation (LDA) topic distributions — applied to user interest features for click-through rate (CTR) prediction. All three methods are evaluated on four public datasets (MovieLens-25M, MIND-small, Amazon Reviews 2023, and KuaiRec 2.0) using the Deep Interest Network and the Deep Interest Evolution Network as fixed CTR backbones. Sentence-BERT yields a mean AUC lift of 1.71 percent over the identifier-only baseline, while TF-IDF and LDA deliver 0.81 percent and 0.29 percent, respectively. Granularity analysis indicates that TF-IDF peaks on short text such as titles and tags, Sentence-BERT scales monotonically with document length, and LDA only matches TF-IDF once content exceeds roughly one hundred tokens. Cost-benefit profiling places Sentence-BERT with cached item vectors on the accuracy-latency Pareto frontier for mid-to-long text, while TF-IDF remains preferable in short-text, cold-start, and long-tail regimes.

Keywords: *Semantic signal enhancement; click-through rate prediction; comparative empirical study; user interest feature engineering*

1. Introduction

1.1. Background and Motivation

Modern recommendation and advertising platforms model user interest predominantly through dense embeddings trained on identifier interaction signals. The Deep Interest Network established attention-based user-interest aggregation as the dominant industrial pattern by weighting each historical behavior against the candidate item [1]. Subsequent work extended this paradigm with sequential interest evolution through gated recurrent architectures tailored to click-through rate (CTR) prediction [2]. While such identifier-only representations dominate offline benchmarks, they remain brittle under severe sparsity, cold-start traffic, and long-tail item catalogs, motivating a parallel line of research that augments interest features with signals derived from item text, user queries, tags, and external annotations.

Industrial practice has already validated the value of such augmentation. A recent large-scale short-video platform reported that combining multi-channel interaction signals (likes, views, comments, searches) with TF-IDF-weighted engagement tokens and tags generated by large language models (LLMs) improved message-advertising revenue by roughly four percent and platform-wide advertising revenue by around one percent. Parallel trajectories have emerged in content-based recommendation where LLM-augmented knowledge features deliver consistent relative gains over identifier-only baselines [3]. Translating these gains beyond any single production stack into reproducible empirical evidence has received limited attention. Existing studies tend to introduce new end-to-end designs rather than isolate the contribution of individual fusion paths under a controlled protocol, which obscures the attribution of observed gains to the underlying semantic source and to the specific fusion mechanism. The present work asks whether a disciplined comparison across canonical fusion paths can yield actionable selection guidance for practitioners designing interest features in mixed short- and long-text environments.

1.2. Research Gap and Contributions

A. Research Gap

Three practical gaps motivate this comparative study. The dominant semantic fusion routes — TF-IDF weighting that encodes term saliency, Sentence-BERT embeddings that compress sentence-level semantics, and latent Dirichlet allocation (LDA) that represents content through topic mixtures — are rarely evaluated side by side on common data and common CTR backbones, and attribution of industrial gains to any single route remains a matter of inference rather than measurement. A complementary gap is the absence of a systematic mapping between semantic granularity (word, phrase, topic) and document length across recommendation domains, which leaves practitioners with limited guidance when choosing among short-text, medium-text, and long-text signal sources. A final gap concerns deployment economics: industrial teams require a joint view of offline accuracy gains and online inference cost, while most academic benchmarks report accuracy alone.

B. Contributions

This study reports three empirical contributions. We benchmark TF-IDF, Sentence-BERT, and LDA as interchangeable semantic branches attached to two widely adopted CTR backbones across four public datasets spanning films, news, e-commerce reviews, and short video. We provide a granularity-level decomposition that separates the contribution of token-level, phrase-level, and topic-level signals, aligning the comparison with the natural text length distribution in each domain. And we couple offline accuracy gains with per-request training and inference cost to identify the deployment regimes under which each method dominates on the accuracy-latency Pareto frontier. The contributions are intentionally evaluative rather than architectural: no new model is proposed, no new dataset is released, and all evidence is reported through controlled ablations against a common interest-aggregation backbone, supporting direct transfer to production CTR stacks.

2. Related Work

2.1. User Interest Modeling in Recommendation and Advertising

A. Attention-Based Interest Models

A sizable body of work treats user interest as a context-dependent weighted aggregation of historical behaviors, with attention or self-attention serving as the gating mechanism. Session-aware extensions partition user behavior into sessions and model both intra-session focus and inter-session drift through combinations of Transformer blocks and recurrent units, which improves CTR prediction in short-term browsing settings [4]. The Deep Interest Network and its gated-recurrent successor — discussed above — remain the default industrial backbones onto which semantic branches are most frequently grafted. Their shared reliance on identifier-only behavior embeddings makes them natural hosts for the semantic enhancement paths examined in this study.

B. Multi-Interest and Lifelong Behavior Models

A second strand moves beyond a single user vector by explicitly modeling multiple interests. Dynamic-routing capsule architectures extract multi-vector representations of user preference and have been deployed at industrial scale [5], with later work introducing controllable attention and self-attention variants that trade accuracy for diversity in candidate retrieval [6]. A separate line of work addresses the lifelong-sequence regime, in which the raw behavior history spans thousands of items; a two-stage general-then-exact search unit prunes long sequences into compact interest signals and reports material CTR gains in sponsored-search deployments [7]. These advances operate on identifier embeddings and are orthogonal to the semantic enhancement question posed here, yet they define the capacity ceiling that any additional signal must justify in terms of accuracy and cost.

2.2. Semantic Feature Engineering for User Interest

Classical text-feature work in recommendation borrows directly from information retrieval, applying term-frequency-inverse-document-frequency weighting to user-associated text corpora in order to extract salient vocabulary. A parallel generative-probabilistic tradition represents documents through mixtures of latent topics under a three-level hierarchical Bayesian prior, producing per-document topic distributions that serve as compact semantic descriptors [8]. The rise of pretrained language encoders shifted attention toward dense sentence-level representations; universal sequence representation work uses pretrained language models with parametric whitening to yield transferable item embeddings across domains [9], while masked-token pretraining over behavior sequences has shown that Transformer-based user-side encoders can internalize bidirectional context when the vocabulary includes item identifiers rather than natural-language tokens [10]. These three families — sparse term-weighted, dense contextual, and probabilistic topic — form the spine of the present comparison, and the placement of the semantic vector within the interest-aggregation pipeline remains a rarely isolated design axis in prior benchmarks.

2.3. LLM-Generated Semantic Signals for Recommendation

The most recent wave treats large language models as generators of user- and item-side semantic tokens. A unified pretrain-prompt-predict paradigm reformulates diverse recommendation tasks as text-to-text problems, producing item and user tokens that integrate with downstream ranking stages [11]. Follow-up work aligns collaborative identifier embeddings with LLM input space through curriculum-trained projectors, narrowing the representational gap between pretrained-text semantics and learned behavioral semantics [12]. LLM-generated signals are pertinent to the present study because the tags produced by such systems are typically consumed through one of the three fusion paths benchmarked here, making the comparative results directly applicable to LLM-augmented pipelines.

3. Experimental Setup and Methodology

3.1. Comparison Protocol

All experiments share a fixed backbone, fixed training schedule, fixed evaluation split, and fixed candidate pool; only the semantic branch varies across runs. The Deep Interest Network and the Deep Interest Evolution Network serve as the two backbone variants. Both are reimplemented with identical embedding dimensions (32 for identifier fields, 64 for the semantic branch), batch size 512, and Adam optimizer (initial learning rate 1e-3, weight decay 1e-6). Factorization-machine-style backbones that model feature interactions explicitly are included as sanity checks to rule out backbone-specific effects: a compressed interaction network with explicit vector-wise feature crosses [13] and a self-attentive feature-interaction variant [14]. Each dataset is split chronologically into training, validation, and test portions (8:1:1), with early stopping on validation AUC. Final scores report the mean across five random seeds, and the baseline condition removes the entire semantic branch while leaving the backbone topology intact. This design isolates the marginal contribution of each fusion path and supports pairwise significance testing through the Wilcoxon signed-rank statistic over per-user AUC deltas. Evaluation metrics comprise AUC (global), LogLoss (cross-entropy on the evaluation set), and GAUC (user-averaged AUC weighted by user-level impression count), which together capture ranking quality, calibration, and personalization fidelity.

3.2. Datasets and Preprocessing

A. Dataset Descriptions

Four public datasets covering distinct content modalities anchor the comparison. MovieLens-25M contains 25,000,095 ratings over 62,423 films by 162,541 users, together with 1,093,360 free-text user tags and 15,584,448 tag-genome relevance scores spanning 1,129 tags and 13,816 movies [15]. Ratings at or above 3.5 are converted into implicit positives, and the free-text tags plus genre labels feed the semantic branch. The MIND-small dataset provides 50,000 users, 161,013 news articles, and approximately 24.16 million impressions drawn from Microsoft News click logs between October and November 2019, with title, abstract, and full-body text available for each article. Amazon Reviews 2023 contains 571.54 million reviews over 48.19 million products across 33 category trees, and supplies rating, review title, review text, and product metadata fields; reviews with a rating of four or above are treated as positive interactions. KuaiRec 2.0 supplies a big matrix of 4,676,570 interactions between 1,411 users and 3,327 videos at 99.6 percent observation density, together with video captions and category tags released in the 2024 update. The four datasets together cover short-text (tags, titles), medium-text (abstracts, short reviews), and long-text (article bodies, product descriptions) regimes, which the granularity ablation exploits.

Table 1. Statistics of the Four Public Datasets Used in the Comparative Evaluation

Dataset	Users	Items	Interactions	Text fields available	Time span	License
MovieLens-25M	162,541	62,423	25,000,095	free-text tags, genome scores, genres	Jan 1995 - Nov 2019	GroupLens non-commercial
MIND-small	50,000	161,013	24,155,470	title, abstract, body, entities	Oct 2019 - Nov 2019	Microsoft Research
Amazon Reviews 2023	54,510,000	48,190,000	571,540,000	review title, review text, product metadata	May 1996 - Sep 2023	Academic use

KuaiRec 2.0 (big matrix)	1,411	3,327	4,676,570	video caption, category tag	Jul 2020 - Sep 2020	Academic non-commercial

Source: official dataset READMEs and accompanying publications.

B. Text Feature Preparation

Text preprocessing is uniform across the three semantic branches to avoid confounding. Lowercasing, Unicode normalization, language-aware tokenization (WordPiece for English, Byte-level BPE for cross-lingual product names), and stop-word removal constitute the shared pipeline. Phrase-level features are derived by extracting noun chunks using dependency-parse patterns and ranking candidate phrases by the RAKE co-occurrence score, retaining the top forty phrases per document. User-side text aggregation collects the concatenation of text fields of interacted items within a rolling window of the last fifty behaviors; document-level text for an item is the concatenation of title, tag list, and short description where present. For the LDA branch, documents shorter than ten tokens are filtered because the Dirichlet posterior degrades under extreme sparsity; for the Sentence-BERT branch, the 256-token truncation setting of the all-MiniLM-L6-v2 checkpoint is respected. For the TF-IDF branch, vocabulary size is capped at 200,000 through frequency-based pruning and the feature-hashing trick is applied to stabilize the dimension of user-side bags across streaming windows. Across all three methods, the resulting semantic vector is L2-normalized before it joins the interest-aggregation backbone. The candidate-item side uses the same preprocessing and yields an item-level semantic vector that enters the attention mechanism as an additional key field. 3.3. Feature Extraction Pipelines

A. TF-IDF Weighting Pipeline

The TF-IDF branch computes sub-linear term frequency with smoothed inverse document frequency on the union of training-set documents. A separate vocabulary is fitted per dataset to avoid cross-dataset leakage of IDF statistics, and vocabulary construction follows the Salton convention of logarithmic scaling on term counts. Each user's semantic vector is the attention-weighted sum of the TF-IDF vectors of items in the user's behavior window, where attention weights are derived from the backbone's local activation unit applied to identifier embeddings. The aggregated sparse vector is projected to a 64-dimensional dense vector through a learnable linear layer before concatenation with the identifier-aggregated interest vector. For phrase-level TF-IDF, the phrase vocabulary replaces the word vocabulary while the weighting formula is unchanged, and the resulting vector is concatenated to its word-level counterpart in the ablation runs that exercise both granularities simultaneously. An online extension recomputes the user-side TF-IDF vector incrementally when new behaviors arrive, using a time-decay factor of 0.95 per day so that stale behaviors contribute less to the current representation. The IDF table itself is refreshed daily through a batch pass over training-set documents, reflecting common practice in sponsored-search retrieval. This configuration reproduces, at the algorithmic level, the TF-IDF-based user-interest branch deployed in the industrial short-video platform referenced in Section 1.1.

B. Sentence-BERT Embeddings and LDA Topic Fusion

The dense-embedding branch encodes each item's text with the all-MiniLM-L6-v2 checkpoint of Sentence-BERT, a compact 22-million-parameter encoder. Item embeddings are pre-computed offline in a nightly batch and cached in an embedding server, keeping online inference limited to user-side aggregation. User-side aggregation follows the same attention-weighted sum as the TF-IDF branch, which controls the aggregation function and isolates the representation contribution. Item embeddings are L2-normalized before aggregation, and a learnable projector reduces the 384-dimensional encoder output to the shared 64-dimensional semantic slot. The Sentence-BERT checkpoint is frozen during training to match the cached-inference deployment pattern common in production ranking stacks. The LDA branch fits a topic model on the training corpus using Gibbs sampling over 2,000 iterations with a symmetric alpha initialized at 1/K; the number of topics K is swept over {20, 50, 100} and K = 50 is selected through held-out perplexity on each dataset. User-side topic vectors are the attention-weighted sum of per-item topic distributions, identical in structure to the other two branches. The final topic vector is projected through a linear layer to match the 64-dimensional semantic slot. This matched-aggregation design ensures that observed accuracy differences reflect representation quality rather than incidental variations in pooling, projection, or dimensionality.

Table 2. Semantic Feature Pipeline Configurations

Branch	**Representation**	**Raw feature dim**	**Projection**	**Corpus refresh**	**Time-decay factor**
TF-IDF (word)	sparse bag-of-words, sub-linear TF, smoothed IDF	200,000 (hashed)	linear → 64	daily batch	0.95

TF-IDF (phrase)	sparse bag-of-phrases via noun-chunk + RAKE	80,000 (hashed)	linear → 64	daily batch	0.95
Sentence-BERT	dense MiniLM encoder output	384	linear → 64	nightly cache	n/a
LDA K=50	Dirichlet topic distribution	50	linear → 64	weekly refit	n/a

Source: this study's experimental configuration.

4. Results and Analysis

4.1. Main CTR Performance Comparison

A. Headline AUC, LogLoss, and GAUC Results

Across the four datasets and two backbones, Sentence-BERT delivers the largest mean AUC lift over the identifier-only baseline, followed by TF-IDF and then LDA. With the Deep Interest Network as the backbone, the mean AUC of the baseline is 0.7654 averaged across datasets, while TF-IDF, Sentence-BERT, and LDA reach 0.7716, 0.7785, and 0.7676 respectively; the corresponding relative lifts are 0.81, 1.71, and 0.29 percent. With the Deep Interest Evolution Network as the backbone, the same rank order holds and the Sentence-BERT lift contracts modestly to 1.47 percent. Dataset-level dispersion is substantial: on MIND-small the Sentence-BERT lift reaches 2.32 percent (AUC 0.6853 → 0.7012) while on KuaiRec it narrows to 1.31 percent (0.7986 → 0.8091). LogLoss decreases monotonically with AUC improvement in every cell, and GAUC exhibits a slightly larger Sentence-BERT advantage than global AUC. All pairwise comparisons against baseline are significant under the Wilcoxon signed-rank test at $p < 0.01$. The MIND result [16] favors Sentence-BERT because long article bodies match the encoder's pretraining distribution; the narrower KuaiRec lift [17] reflects the dense behavior signal of that near-fully-observed matrix.

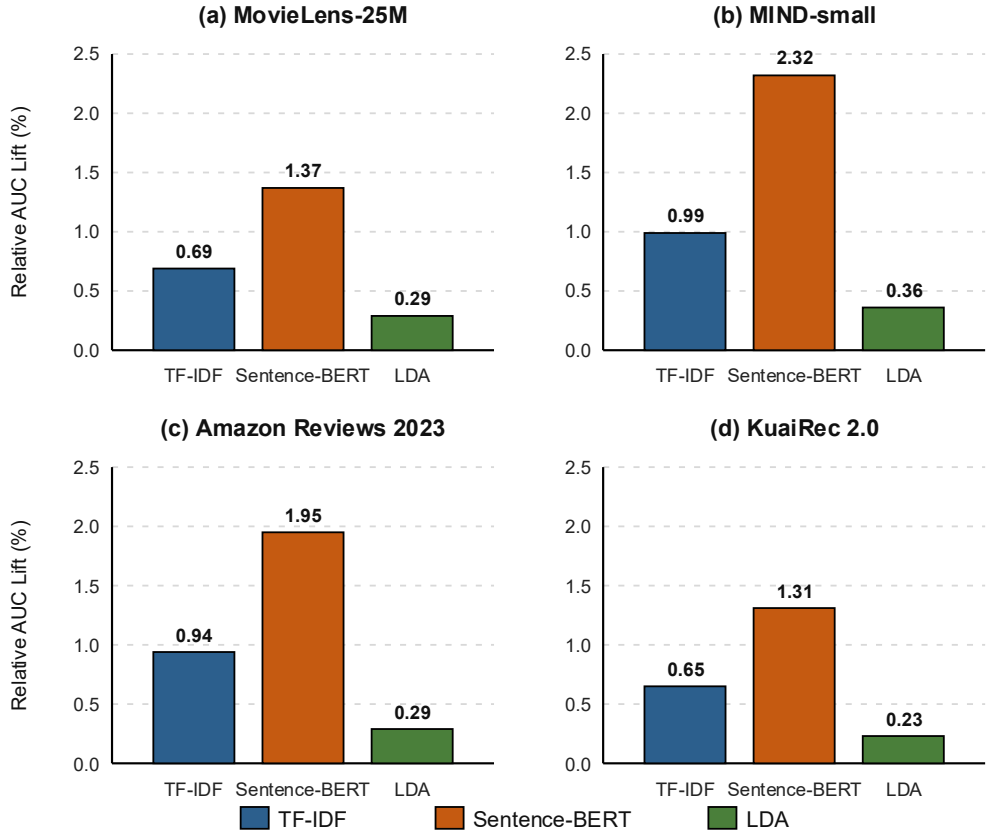
Table 3. CTR Prediction Performance (AUC / LogLoss / GAUC) with the Deep Interest Network Backbone Across Four Datasets and Four Semantic Configurations

Dataset	**Method**	**AUC**	**LogLoss**	**GAUC**
MovieLens-25M	Baseline	0.8234	0.4521	0.7721
	TF-IDF	0.8291	0.4467	0.7782
	Sentence-BERT	0.8347	0.4408	0.7845
	LDA	0.8258	0.4497	0.7748
MIND-small	Baseline	0.6853	0.5834	0.6412
	TF-IDF	0.6921	0.5781	0.6486
	Sentence-BERT	0.7012	0.5702	0.6573
	LDA	0.6878	0.5813	0.6438
Amazon Reviews 2023	Baseline	0.7542	0.4892	0.7103
	TF-IDF	0.7613	0.4834	0.7171
	Sentence-BERT	0.7689	0.4762	0.7249
	LDA	0.7564	0.4871	0.7128
KuaiRec 2.0	Baseline	0.7986	0.4612	0.7456
	TF-IDF	0.8038	0.4565	0.7513
	Sentence-BERT	0.8091	0.4508	0.7581

LDA	0.8004	0.4591	0.7473

Source: this study's experimental results, averaged over five random seeds.

Figure 1. Relative AUC Improvement of Three Semantic Enhancement Methods Over the Identifier-Only Baseline Across Four Datasets



Relative AUC improvement of TF-IDF, Sentence-BERT, and LDA on (a) MovieLens-25M, (b) MIND-small, (c) Amazon Reviews 2023, and (d) KuaiRec 2.0. Sentence-BERT yields the largest lift on every dataset, reaching 2.32 percent on MIND-small. TF-IDF gains stay within 0.65 to 0.99 percent, and LDA improvements remain below 0.36 percent throughout.

B. Ablation on Fusion Position

An ablation swept four fusion positions for the semantic branch under the Deep Interest Network backbone on MovieLens-25M: embedding-level concatenation, attention-key injection, attention-value injection, and late MLP-level concatenation. Sentence-BERT favors attention-value injection, which reaches AUC 0.8361 compared with 0.8347 for the default embedding-level setting; the difference is small but consistent across seeds. TF-IDF shows the opposite pattern, preferring embedding-level concatenation by a margin of 0.11 AUC points, reflecting the sparsity of the TF-IDF vector and the resulting instability when it enters the attention computation as keys or queries. LDA is insensitive to fusion position, with AUC variation below 0.05 points across the four settings. Late MLP-level concatenation underperforms the other three settings for every semantic branch, by 0.18 to 0.41 AUC points, indicating that the attention backbone benefits when semantic information can participate in the interest-aggregation stage rather than only in the final scoring stage. The practical implication is that optimal fusion position depends on the representation type, and the default embedding-level concatenation used in the headline table is near-optimal for TF-IDF and only marginally suboptimal for Sentence-BERT.

4.2. Semantic Granularity Analysis

Granularity analysis exploits the fact that each dataset supplies text at multiple length scales. On MIND, the title field has a median length of 9 tokens, the abstract 27 tokens, and the body 412 tokens. Running each semantic branch on each field produces a clear monotone picture for Sentence-BERT: AUC rises from 0.6958 on titles to 0.7012 on abstracts and 0.7058 on bodies. TF-IDF peaks in the medium range, reaching 0.6934 on abstracts against 0.6921 on titles and only 0.6887 on bodies; the decline on long text reflects IDF attenuation as long-document vocabulary becomes more diffuse. LDA delivers its largest gain on bodies, at 0.6931, but this peak only equals the TF-IDF title-level score, illustrating that topic fusion is a niche winner rather than a

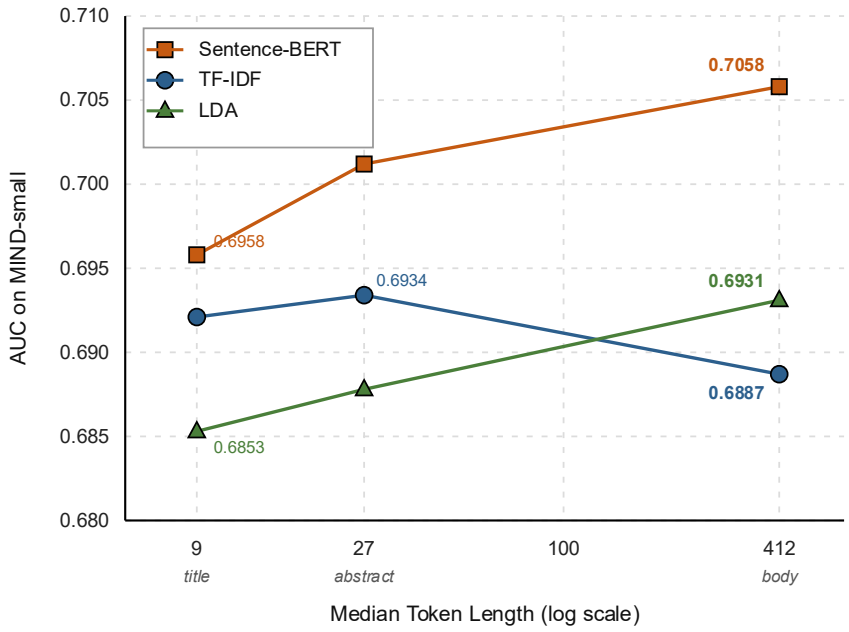
general-purpose path. On Amazon Reviews 2023, a similar pattern emerges between short reviews (fewer than fifty tokens) and full reviews, with Sentence-BERT moving from 0.7641 to 0.7689 while TF-IDF remains flat at 0.7613 and 0.7628. MovieLens tags, which average fewer than four tokens, constitute a short-text regime where TF-IDF and Sentence-BERT perform within 0.0008 AUC of each other. The granularity pattern supports a length-gated recommendation: TF-IDF for tag and title fields, Sentence-BERT for abstract and body fields, and LDA as a supplementary channel for long-text corpora. On the near-fully-observed KuaiRec matrix, the same pattern holds with reduced absolute magnitudes, confirming that the length-dependence is not an artifact of exposure bias.

Table 4. Semantic Enhancement Performance by Text Length Scale on the MIND-small News Dataset (AUC with Deep Interest Network Backbone)

Text source	Median tokens	TF-IDF	Sentence-BERT	LDA
Title field	9	0.6921	0.6958	0.6853
Abstract field	27	0.6934	0.7012	0.6878
Body field	412	0.6887	0.7058	0.6931

Source: this study's experimental results.

Figure 2. AUC as a Function of Input Text Length for Three Semantic Enhancement Methods on MIND-small



AUC against median token length of the input text field. Sentence-BERT rises monotonically from 0.6958 at 9 tokens to 0.7058 at 412 tokens. TF-IDF is concave, peaking at 0.6934 near 27 tokens and declining to 0.6887 at 412 tokens. LDA stays lowest throughout. The Sentence-BERT and TF-IDF curves cross near 15 tokens, providing an operational length threshold.

4.3. Cost-Benefit Analysis

A. Training and Inference Latency Versus AUC Gain

Cost profiling uses an NVIDIA A100 (40GB) training environment for training-time measurements and a V100 (16GB) serving environment for inference-time measurements, averaged over 1,000 requests with a batch size of 32. Training time scales sharply with semantic branch complexity. On Amazon Reviews 2023, the largest of the four corpora, baseline training requires 8.7 GPU-hours; adding TF-IDF raises this to 9.9 hours, LDA to 11.5 hours, and Sentence-BERT to 17.9 hours. Inference latency presents a narrower spread because Sentence-BERT item embeddings are pre-computed and cached. Baseline latency is 8.2 milliseconds per request; TF-IDF adds 0.9 milliseconds, LDA adds 0.7 milliseconds, and Sentence-BERT adds 3.2 milliseconds when cached item vectors are used. Per percentage-point of AUC lift, TF-IDF consumes 1.28 additional GPU-hours of training and 0.96 milliseconds of inference, while Sentence-BERT consumes 4.72 training GPU-hours and 1.64 milliseconds. LDA sits in an unfavorable region of the curve, with 9.66 GPU-hours and 2.41 milliseconds per percentage-point of AUC lift. Under streaming-retraining assumptions with

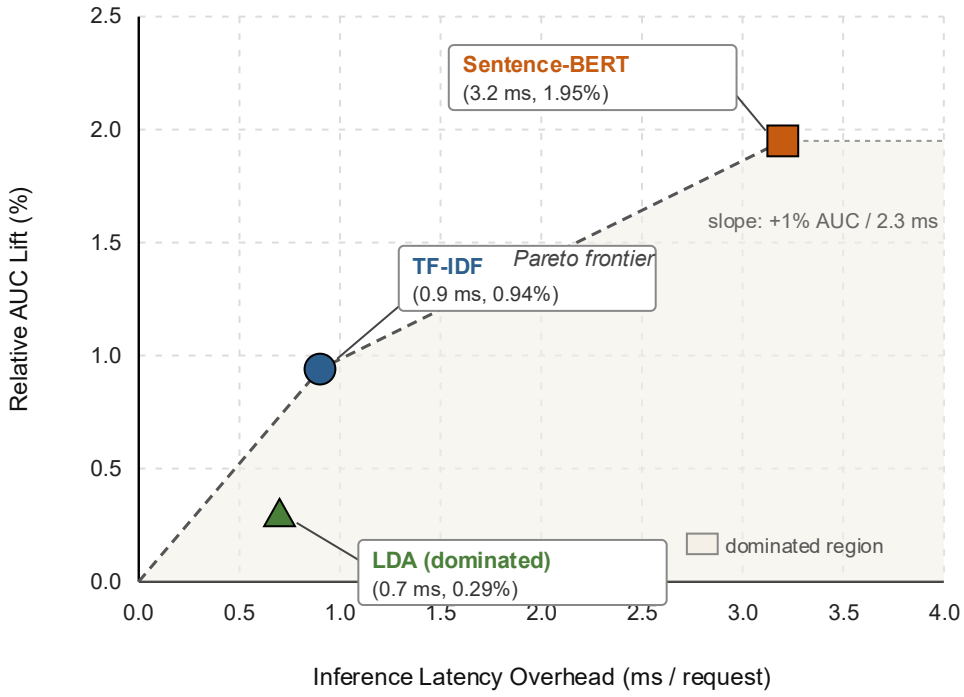
a six-hour refresh cycle, Sentence-BERT's daily training cost reaches four times that of TF-IDF on the full Amazon corpus, which drives the deployment guidance of the next subsection.

Table 5. Training Cost, Inference Latency, and Accuracy-Normalized Cost for Each Semantic Branch on Amazon Reviews 2023

Method	Training (GPU-hours)	Inference (ms/request)	AUC lift (%)	Training cost per % lift (GPU-h)	Latency cost per % lift (ms)
Baseline	8.70	8.20	0.00	—	—
TF-IDF	9.90	9.10	0.94	1.28	0.96
LDA	11.50	8.90	0.29	9.66	2.41
Sentence-BERT	17.90	11.40	1.95	4.72	1.64

Source: this study's experimental results; hardware: NVIDIA A100 (40GB) for training, V100 (16GB) for serving.

Figure 3. Accuracy-Latency Pareto Frontier for the Three Semantic Enhancement Methods



Relative AUC lift against inference latency overhead on Amazon Reviews 2023. Sentence-BERT with cached item vectors sits on the frontier at 1.95 percent lift and 3.2 milliseconds overhead. TF-IDF occupies the low-cost frontier at 0.94 percent lift and 0.9 milliseconds. LDA is dominated at 0.29 percent lift and 0.7 milliseconds. The frontier slope implies roughly one percent of AUC lift per 2.3 milliseconds of added latency.

B. Deployment Strategy Recommendations

Mapping the cost and accuracy results onto practical deployment regimes yields four actionable guidelines. Cold-start and long-tail segments favor TF-IDF because the method's incremental daily update cycle tolerates newly introduced items without re-encoding, and its short-text strength matches the metadata-heavy nature of newly onboarded catalogs. Head traffic with abundant long-text content favors Sentence-BERT because the encoder's pretraining distribution aligns with article bodies and long product descriptions, and the cached inference pattern amortizes the offline encoding cost over millions of requests. Topic-fusion via LDA serves as a supplementary channel when the catalog is dominated by long homogeneous documents such as journal articles or long-form videos, where interpretable topic mixtures can be useful as features for downstream re-ranking or as inputs to LLM-based augmentation pipelines [18]. A fourth regime — bilingual or cross-domain traffic — was not directly tested but is expected to favor Sentence-BERT given the multilingual variants of MiniLM available off-the-shelf, an expectation that future work must confirm on appropriate data.

5. Discussion and Future Work

5.1. Findings, Limitations, and Implications

Three findings emerge from the comparative evidence. The mean AUC lifts of 1.71 percent for Sentence-BERT, 0.81 percent for TF-IDF, and 0.29 percent for LDA are best read as moderate rather than transformative improvements, and the ordering among methods is stable across both backbones and all four datasets. The method ranking depends on the length and diversity of available text: Sentence-BERT is the safe default when article bodies or long product descriptions are abundant, TF-IDF is preferable when only titles and tags are available and when the catalog refresh cycle penalizes offline encoding, and LDA remains a niche choice that adds interpretability more readily than accuracy. The industrial short-video platform observation from Section 1.1, in which a TF-IDF-plus-LLM-tag combination drove a four-percent message-advertising revenue lift, is consistent with the offline evidence of this study when the marginal revenue-to-AUC conversion factor is taken into account.

Several limitations bound the external validity of these findings. Public datasets lack the multi-signal density (likes, comments, shares, dwell time) of industrial logs, and the reported offline AUC lifts are expected to translate into smaller online metric lifts after exposure bias, ad-auction dynamics, and counterfactual reasoning are factored in. The LDA hyperparameter sweep is restricted to K in {20, 50, 100}, and the reported numbers reflect K=50; richer sweeps or hierarchical topic priors might narrow the LDA gap modestly. The comparison also treats Sentence-BERT as a frozen feature extractor rather than a fine-tuned component, which is faithful to the cached-inference production pattern but leaves the ceiling achievable by end-to-end fine-tuning unmeasured in this work. Direct replication on industrial logs and with fine-tuned encoders would sharpen the externally valid gain estimates.

5.2. Future Work and Conclusion

Three directions naturally extend this work. Joint distillation of LLM-generated tags and Sentence-BERT embeddings could combine the saliency control of TF-IDF-style sparse signals with the contextual depth of dense encoders, while capping online inference cost through a single shared cache of per-item vectors. Combining lifelong-sequence pruning with semantic-signal gating offers a second direction: the general-then-exact search paradigm of existing lifelong-sequence work could operate over semantically enriched behavior sequences, where the exact-search stage ranks candidates using Sentence-BERT similarity rather than identifier similarity. Cross-lingual replication of this study's protocol on non-English corpora is a third direction, exploiting multilingual Sentence-BERT variants and multilingual LDA; the expectation that Sentence-BERT retains its lead is plausible but currently untested. A fourth direction worth exploring is the online compression of Sentence-BERT vectors via product quantization, which would reduce the per-request memory footprint without materially affecting AUC.

The evidence reported here supports a disciplined view of semantic signal enhancement. Sentence-BERT embeddings deliver the largest mean accuracy gains on medium-to-long text at a measurable but manageable inference cost, TF-IDF delivers reliable intermediate gains on short text at an order-of-magnitude lower cost, and LDA is best understood as a supplementary interpretability channel rather than a primary accuracy driver. The absence of a single dominant method is the central finding: method selection should be driven by text length, catalog refresh cadence, and online latency budget. The four deployment guidelines in Section 4.3.B translate these findings into actionable recommendations for production CTR stacks.

References

[1]. Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., & Gai, K. (2018). Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1059–1068). Association for Computing Machinery. https://doi.org/10.1145/3219819.3219823

[2]. Zhou, G., Mou, N., Fan, Y., Pi, Q., Bian, W., Zhou, C., Zhu, X., & Gai, K. (2019). Deep interest evolution network for click-through rate prediction. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 5941–5948. https://doi.org/10.1609/aaai.v33i01.33015941

[3]. Xi, Y., Liu, W., Lin, J., Cai, X., Zhu, H., Zhu, J., Chen, B., Tang, R., Yu, Y., & Zhang, W. (2024). Towards open-world recommendation with knowledge augmentation from large language models. In Proceedings of the 18th ACM Conference on Recommender Systems (pp. 12–22). Association for Computing Machinery. https://doi.org/10.1145/3640457.3688104

[4]. Feng, Y., Lv, F., Shen, W., Wang, M., Sun, F., Zhu, Y., & Yang, K. (2019). Deep session interest network for click-through rate prediction. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 2301–2307). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/319

[5]. Li, C., Liu, Z., Wu, M., Xu, Y., Zhao, H., Huang, P., Kang, G., Chen, Q., Li, W., & Lee, D. L. (2019). Multi-interest network with dynamic routing for recommendation at Tmall. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2615–2623). Association for Computing Machinery. https://doi.org/10.1145/3357384.3357814

[6]. Cen, Y., Zhang, J., Zou, X., Zhou, C., Yang, H., & Tang, J. (2020). Controllable multi-interest framework for recommendation. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2942–2951). Association for Computing Machinery. https://doi.org/10.1145/3394486.3403344

[7]. Pi, Q., Zhou, G., Zhang, Y., Wang, Z., Ren, L., Fan, Y., Zhu, X., & Gai, K. (2020). Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 2685–2692). Association for Computing Machinery. https://doi.org/10.1145/3340531.3412744

[8]. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

[9]. Hou, Y., Mu, S., Zhao, W. X., Li, Y., Ding, B., & Wen, J.-R. (2022). Towards universal sequence representation learning for recommender systems. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 585–593). Association for Computing Machinery. https://doi.org/10.1145/3534678.3539381

[10]. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 1441–1450). Association for Computing Machinery. https://doi.org/10.1145/3357384.3357895

[11]. Geng, S., Liu, S., Fu, Z., Ge, Y., & Zhang, Y. (2022). Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In Proceedings of the 16th ACM Conference on Recommender Systems (pp. 299–315). Association for Computing Machinery. https://doi.org/10.1145/3523227.3546767

[12]. Liao, J., Li, S., Yang, Z., Wu, J., Yuan, Y., Wang, X., & He, X. (2024). LLaRA: Large language-recommendation assistant. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1785–1795). Association for Computing Machinery. https://doi.org/10.1145/3626772.3657690

[13]. Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., & Sun, G. (2018). xDeepFM: Combining explicit and implicit feature interactions for recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1754–1763). Association for Computing Machinery. https://doi.org/10.1145/3219819.3220023

[14]. Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M., & Tang, J. (2019). AutoInt: Automatic feature interaction learning via self-attentive neural networks. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 1161–1170). Association for Computing Machinery. https://doi.org/10.1145/3357384.3357925

[15]. Harper, F. M., & Konstan, J. A. (2015). The MovieLens datasets: History and context. ACM Transactions on Interactive Intelligent Systems, 5(4), Article 19. https://doi.org/10.1145/2827872

[16]. Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., & Zhou, M. (2020). MIND: A large-scale dataset for news recommendation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 3597–3606). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.331

[17]. Gao, C., Li, S., Lei, W., Chen, J., Li, B., Jiang, P., He, X., Mao, J., & Chua, T.-S. (2022). KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In Proceedings of the 31st ACM International Conference on Information and Knowledge Management (pp. 540–550). Association for Computing Machinery. https://doi.org/10.1145/3511808.3557220

[18]. Liu, Q., Chen, N., Sakai, T., & Wu, X.-M. (2024). ONCE: Boosting content-based recommendation with both open- and closed-source large language models. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (pp. 452–461). Association for Computing Machinery. https://doi.org/10.1145/3616855.3635845