

Information Flow Advertising Relevance Ranking Method Using Knowledge-Enhanced BERT

Xia Yan^{1,2}, Anita Binti Rosli², Aryaty Binti Alwie²

¹College of Journalism and Communication, Advertising, Shanghai Jian Qiao University, Shanghai 201306, China

²Department of Social Science & Management, Faculty of Humanities, Management & Science, Universiti Putra Malaysia Sarawak, 97008 Bintulu, Malaysia

Corresponding author: Anita Binti Rosli (e-mail: anitarosli@upm.edu.my)

ABSTRACT Efficient semantic relevance ranking is essential for intelligent information delivery in modern wireless communication and data-driven service systems, where accurate content matching can improve the utilization of communication resources and user experience. To overcome the limitations of conventional BERT models in semantic understanding, business knowledge integration, and real-time inference, this study proposes a Knowledge-Enhanced BERT (KE-BERT) framework for information flow advertising relevance ranking. The proposed approach incorporates multi-source domain knowledge into the Transformer attention mechanism through entity alignment and further employs a dual-channel interaction architecture with adaptive gating to dynamically balance semantic and knowledge representations. To satisfy industrial latency requirements, layer pruning, quantization, and knowledge distillation are jointly adopted to construct a lightweight inference model. Experimental results demonstrate that the proposed framework achieves superior ranking performance and significantly improves online efficiency, yielding an NDCG@10 of 0.782 with a P99 latency of only 35 ms while increasing click-through rate and revenue-related metrics in real-world deployment. The proposed knowledge-aware semantic ranking strategy provides a practical reference for intelligent information dissemination and semantic resource optimization in next-generation wireless communication and electromagnetic information service systems.

INDEX TERMS knowledge-enhanced bert, semantic relevance ranking, entity alignment, semantic communications, intelligent information dissemination

I. INTRODUCTION

THE quality of advertising relevance ranking directly impacts user experience and platform commercial value; although BERT excels in static text semantic representation, its architecture inherently struggles to integrate real-time changing business knowledge, such as dynamic inventory levels, evolving fashion trends, or specific fabric properties, which are critical for optimal ranking and value. This leads to a deviation in the model's understanding of user intent. In queries such as "fitness equipment recommendation", traditional BERT ignores the seasonal strategy of "summer promotion" or the regional consumption characteristics of "first-tier city users prefer high-end brands". On the other hand, the high computational overhead of complex models makes it difficult to meet the strict real-time requirements of information flow scenarios. User interests are highly dynamic, and the comparability of the value of different advertising materials needs to rely on fine-grained business rules for judgment. Native BERT lacks the ability to model such implicit business logic.

The core advantage of KE-BERT is to embed multi-source domain knowledge bases into the attention mecha-

nism of Transformer in an entity-aligned manner to achieve the underlying fusion of business rules and semantic representation. The dual-channel interaction module uses feature pyramid and adaptive GN to dynamically generate weight coefficients through learning parameters, automatically strengthening the dominance of knowledge signals in strong business intent scenarios, while maintaining the advantage of semantic understanding in general demand scenarios, and then dynamically adjusting the weight distribution of knowledge signals and text signals, significantly improving the sensitivity of the model to implicit user intentions. The combination of pruning, quantization, and knowledge distillation successfully prunes the structured layers while retaining the crucial knowledge injection layer, breaking through the computational bottleneck of the model.

This paper adopts the KE-BERT framework, builds a structured knowledge base based on advertising business logic, and embeds the BERT attention layer to solve the problem of weak business feature modeling. It also combines the dual-channel interaction mechanism to enhance the implicit intent capture capability and uses lightweight migration technology to meet online latency constraints. Its

innovation is reflected in the knowledge injection method based on entity alignment to realize the underlying fusion of business rules and BERT, and solves the problem of adaptive weighting of semantics and knowledge signals through a dynamic gating mechanism. Verified by an experiment with 2 million exposures on Meituan's advertising platform, this method improves NDCG@10 and user retention time while maintaining low latency, and reduces the relevance error rate and complaint rate, providing a reusable technical idea for a high-precision, low-latency advertising ranking system.

II. RELATED WORK

Traditional advertising ranking models mainly rely on feature engineering and shallow network architectures (such as CTR estimation methods based on logistic regression or factorization machines) [1,2]. Although such models have high computational efficiency, they are limited by their insufficient feature cross-functionality and are unable to capture the deep semantic associations between users and advertisements, as well as the real-time drift of users' interests, resulting in delayed recommendations for sudden events. With the development of deep learning, hybrid architectures such as Wide&Deep [3] and DeepFM (Deep Factorization Machine) [4] have partially alleviated the bottleneck of feature interaction by combining wide models with deep neural networks, but their text representation is still limited to bag-of-words [5,6] or static word embedding (Word2Vec, Word2Vector) [7,8], resulting in significant deviations in the fine-grained understanding of user query intent. In recent years, BERT has been applied into the recommendation field with its powerful contextual semantic modeling capabilities [9,10]. However, the integration of business knowledge in advertising scenarios by native BERT can only be achieved through post-feature concatenation, and it is impossible to establish a knowledge-aware underlying representation. The complex computing structure of multi-layer Transformer cannot meet the high concurrency and low latency requirements of information flow advertising, which seriously restricts its industrial implementation [11].

To bridge the gap between semantic understanding and domain knowledge, KE technology has gradually become a research hotspot [12]. Early work mainly focused on knowledge graph embedding (KGE), mapping entity relationships to low-dimensional vector space through the TransE (Translating Embedding) algorithm, but its paradigm independent of text modeling leads to the separation of semantics and knowledge representation [13,14]. In advertising scenarios, although TransE can model the "brand-category" relationship, it cannot associate the "latest model" in user queries with the "new product launch" label of KG. Subsequent research has tried multimodal fusion solutions to interact KG node features with text embeddings (RippleNet) [15,16]. Although such methods improve the interpretability of recommendations, they face real-time bottlenecks in dynamic business scenarios. The offline construction mode of the KG cannot adapt to the real-time drift of user interests,

and the cascade propagation mechanism of the graph neural network (GNN) applies additional computational overhead [17,18].

III. KE-BERT FRAMEWORK DESIGN

A. KE ARCHITECTURE

The KE-BERT framework adopts an end-to-end semantic-business joint model and realizes the coordinated optimization of knowledge injection, feature fusion, and efficient inference through a multi-level processing pipeline. The KE architecture is shown in Figure 1.

As shown in Figure 1, the system receives bimodal input of text features and business features, locates business entities in the text through the entity alignment module, and retrieves structured knowledge in the product attribute map and user interest tag system in real-time. This knowledge is injected into the BERT encoder through the improved attention mechanism [19]. Specifically, the knowledge decay factor ($\alpha = 0.6$) is used to adjust the weight of the entity attribute vector K_{know} in the attention calculation:

$$\begin{aligned} & \text{Attention}(Q, K, V, K_{\text{know}}) \\ &= \text{Softmax}\left(\frac{Q(K + \alpha K_{\text{know}})^T}{\sqrt{d_k}}\right) V \end{aligned} \quad (1)$$

The text representation output by the encoder and the graph representation generated by the knowledge GNN (implemented by GraphSAGE) [20] enter the dual-channel interaction module and perform multi-scale fusion at the representation layer through the feature pyramid structure. The first step of the fusion is to perform global maximum pooling on the text representation to reduce the knowledge representation to 256 dimensions. The second step is to generate a joint representation through the cross-attention mechanism. The third step is to deploy an adaptive GN at the matching layer and dynamically generate the weight coefficient from the learnable parameter W_g , as follows:

$$\lambda = \sigma(W_g \cdot [\text{fused}_{\text{text}} \parallel \text{fused}_{\text{know}}]) \quad (2)$$

σ refers to the standard deviation.

This coefficient controls the contribution ratio of semantic signals to business signals in the final prediction score and automatically reduces the text weight ($\lambda \rightarrow 0$) when the user query implies a strong business intent. To meet the online latency constraints, the framework adopts a three-stage lightweight migration. The full version of KE-BERT (teacher model) is pruned (removing the 6th/9th/12th layers). The FP32 parameters are converted to INT8 precision through quantized distillation (By combining this with a dynamic range quantization strategy, the error is controlled to within 2.3%), and the knowledge representation is transferred to the 4-layer lightweight BERT (student model) using the KL (Kullback-Leibler) divergence and cross entropy mixed loss. The 4-layer structure, after subsequent empirical optimization tests (compared to 3, 5, and 6 layers), achieved the lowest P99 latency (35ms) while maintaining

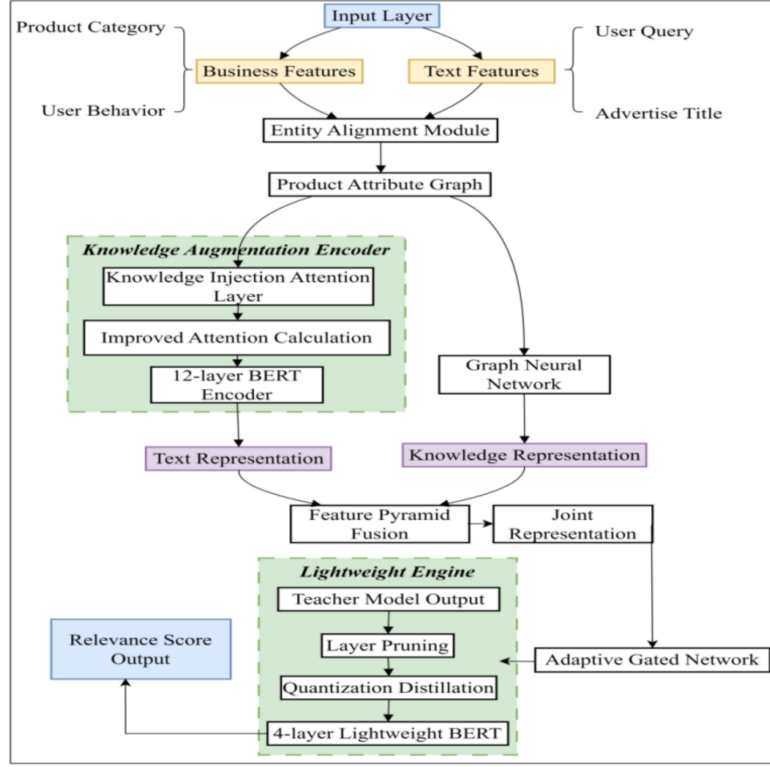


FIGURE 1. KE architecture

NDCG@10 at 0.768, thus achieving Pareto optimality in performance and efficiency. Finally, the relevance score is output after optimization by ONNX Runtime (Open Neural Network Exchange Runtime).

B. MULTI-SOURCE KNOWLEDGE EMBEDDING

The multi-source knowledge embedding architecture is shown in Figure 2.

In Figure 2, a two-layer knowledge base system is built based on the business logic of the Meituan advertising platform. The product attribute graph is stored in the Neo4j graph database [21]. Entity nodes (product/store/brand), attribute nodes (price range/promotion tag/category level), and relationship edges are defined, and the integrity of business rules is ensured through pattern constraints. The user interest tag system integrates static portraits (age/gender/city level) and dynamic behavior features. The latter generates the user's behavior sequence Embedding in the last 30 minutes in real-time through the Flink streaming computing [22,23] engine and stores it in the Redis cluster in Protobuf format.

The knowledge injection process uses entity alignment-driven attention transformation technology. When the input text (user query “cost-effective mobile phone”) enters the encoding layer, the BiLSTM-CRF named entity recognition module first locates the business entity (“mobile phone” → product category entity) and triggers two-way retrieval in parallel [24,25]. The product attribute graph obtains the associated attributes (price range: mid-range) through Cypher query MATCH, and the user interest tag system retrieves its

recent behavior (clicks on “thousand-yuan phone” account for 82%). The retrieval results are converted into knowledge vectors by the attribute encoder. Among them, the numerical attributes (price range) are piecewise normalized, and the discrete attributes (promotion labels) are learned as one-hot embedding. The attention score formula of KE is:

$$\text{Score}(Q, K, K_{\text{know}}) = \frac{1}{\sqrt{d_k}} \cdot Q \cdot (K + \alpha \cdot \phi(K(K_{\text{know}})))^T \quad (3)$$

d_k refers to the dimension of each vector in the key matrix, and $\phi(\cdot)$ is the attribute mapping function. Entity alignment incorporates the retrieved knowledge vector K into the attention key matrix after being weighted by a decay factor $\alpha = 0.6$, thereby modifying the attention score calculation. Specifically, this is reflected in the interaction between Q and $K + \alpha \cdot K_{\text{know}}$ in formula (1), which strengthens the features related to business entities in the semantic understanding of the model.

This design makes fuzzy semantics such as “cost-effectiveness” automatically strengthen the weight of the “mid-range price range” signal in K_{know} when calculating the attention association between mobile phones and prices. To deal with entity ambiguity (such as “apple” may refer to fruit or brand), a disambiguation mechanism based on user real-time interests is applied. If the user has recently clicked on the 3C (Computer, Communication, Consumer Electronics) category with a proportion of $> 70\%$, the Apple brand entity is activated first, otherwise the fruit entity is

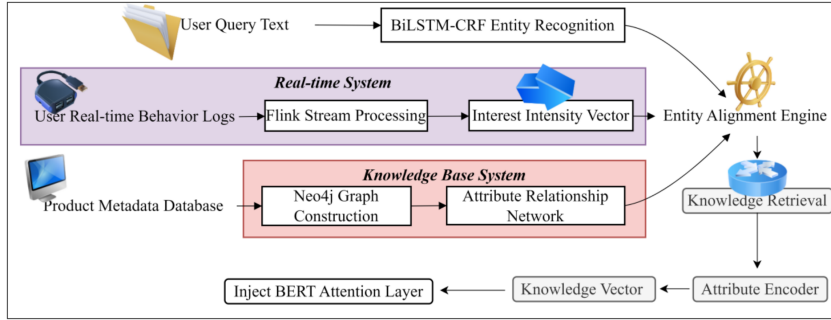


FIGURE 2. Multi-source knowledge embedding architecture. Note: Bidirectional Long Short-Term Memory-Conditional Random Fields (BiLSTM-CRF)

activated).

The knowledge base update adopts a dual-channel mechanism to ensure timeliness. The product attribute map is incrementally updated every morning; the user interest tags receive behavioral events in real-time through Kafka; the tag engine completes feature refresh within 200ms. When deploying online services, the entity alignment engine relies on the FAISS (Facebook AI Similarity Search) index to achieve millisecond-level retrieval (P99 latency <8ms, 99th Percentile), supporting an average of 2 billion knowledge query requests per day.

C. DUAL-CHANNEL INTERACTION MECHANISM

The dual-channel interaction mechanism adopts a pyramid processing flow. Feature pyramids enhance representation complementarity by performing global max pooling compression on text representations, GCN (Graph Convolutional Network) aggregation on knowledge representations [26,27], and residual connections, cross-modal attention, and one-dimensional convolutional dimensionality reduction at multiple scales to achieve hierarchical fusion of semantics and business signals. The adaptive GN is deployed at the matching layer. Its inputs include global average pooling features (h_t^{gmp}) of text representation, matrix of knowledge representation (H_k), and real-time user behavior features processed by multilayer perceptron (MLP). The output is a dynamic gating value λ , which is used to adjust the weights of semantic and business signals in the prediction score. The gating function is defined as:

$$\lambda = \sigma(W_g \cdot [h_t^{\text{gmp}} \parallel \max(H_k) \parallel \text{MLP}(u)]) \quad (4)$$

λ controls the weight distribution of the semantic component and the business component in the prediction score:

$$\hat{y} = \lambda \cdot \text{NN}_t(w_1 F_1 + w_2 F_2 + \dots + w_n F_n) + (1 - \lambda) \cdot \text{NN}_k(w_1 F_1 + w_2 F_2 + \dots + w_n F_n) \quad (5)$$

F_n refers to the input feature; w_n refers to the weight; NN_t and NN_k are two trained neural networks, respectively.

When the user's behavior indicates that they have a strong business demand (for example, when the user queries

“urgent delivery of medicines”, and its historical on-time delivery rate exceeds 90%), λ automatically drops below 0.3, allowing the relevant prediction model to dominate the decision-making, thereby giving priority to matching those merchants with high timeliness. On the contrary, in more generalized demand scenarios (such as users querying “casual snacks”), λ is kept close to 0.6, thereby maintaining the advantage of semantic understanding.

D. LIGHTWEIGHT INFERENCE ENGINE

The computational bottleneck of the KE model is solved through a three-stage compression strategy, and the online service latency constraint is met while ensuring the fidelity of business knowledge. In the first stage, structured layer pruning is performed to analyze the knowledge sensitivity of each layer of the full version of KE-BERT (12-layer Transformer) [28]. Redundant layers are identified based on the gradient amplitude criterion, and the 6th, 9th, and 12th layers (layers with knowledge attention weight variance <0.1) are removed, reducing the number of parameters from 110M to 46M (a reduction of 58.2%), while retaining the key knowledge injection layer. The pruned model is fine-tuned by knowledge calibration using the cosine similarity loss with a temperature coefficient, as follows:

$$L_{\text{calib}} = \sum_{i=1}^N [1 - \cos(T_i^{\text{prune}}, T_i^{\text{full}})] \cdot \exp\left(\frac{\|K_{\text{know}}\|_2}{\tau}\right) \quad (6)$$

T_i^{prune} represents the output of the sample under the pruned model, and T_i^{full} represents the output of the T_i sample under the full model. τ represents the hyperparameter.

In the second stage, quantization-aware distillation is used to design a dual-path knowledge transfer mechanism. The response distillation path minimizes the KL divergence, and

and hidden state of the key knowledge layer. The formula of the distillation loss function is:

$$L_{\text{distill}} = \beta \cdot KL(Q_s \| Q_t) + \gamma \cdot \| \text{Attn}_s^{(3)} - \text{Attn}_t^{(3)} \|_F + (1 - \beta - \gamma) \cdot L_{\text{CE}} \quad (7)$$

$KL(Q_s||Q_t)$ represents the KL divergence of the output distribution of the s-th model relative to the output distribution of the t-th model. After distillation, INT8 quantization is performed, and dynamic range quantization is used to convert the weights of the fully connected layer. A mixed precision strategy is adopted for GN to control the quantization error within 2.3%.

In the final stage, the model is converted into a static computational graph through ONNX Runtime [29], and the operator fusion technology is applied to merge adjacent LayerNorm-GELU-GeMM (Layer Normalization- Gaussian Error Linear Unit-General Matrix Multiply) sequences. A pre-calculated cache is deployed for the knowledge retrieval module, and FP16 cache is performed for the attribute vectors of high-frequency entities (occurrence frequency>104), reducing 85% of real-time graph database queries. The inference engine is compiled and generated by TensorRT (TensorRuntime) [30,31], and dynamic batch processing (Batch=32) and memory reuse are implemented on the NVIDIA T4 GPU (Graphics Processing Unit).

E. REAL-TIME GUARANTEE MECHANISM

The distributed cache layer uses Redis cluster to store high-frequency entity attribute vectors (entities with daily visits > 106 times) [32,33]. Consistent hash sharding achieves 99.5% request localization response, and entity retrieval P99 latency is reduced from 23ms to 1.8ms. GPU quantitative indexing based on FAISS is enabled for long-tail entities, and a 128-dimensional HNSW (Hierarchical Navigable Small World) graph is constructed to achieve millisecond-level neighbor search (recall rate 98.7%). The computing pipeline is optimized to implement the CPU (Central Processing Unit)-GPU heterogeneous parallel architecture [34,35]. The CPU thread pool pre-executes entity alignment and knowledge retrieval, while the GPU stream processor asynchronously executes BERT encoding. A double-buffered memory pool (Block size = 256MB) is designed to eliminate data transmission latency, and the data copy and calculation overlap rate is increased to 92% through CUDA (Compute Unified Device Architecture) stream binding technology.

A latency-sensitive adaptive batch scheduling algorithm is applied in the dynamic batch processing engine. First, a 10 millisecond time window is defined. When the length of the request queue reaches or exceeds 8, the system immediately triggers inference, and the maximum batch size allowed is 64. If the length of the request queue is less than 8 and the waiting time exceeds 10 milliseconds, the request is dynamically packaged according to the length of the current request queue. The batch size formula for dynamic packaging is:

$$B = 2^{\lfloor \log_2(L_q) \rfloor} \quad (8)$$

L_q is the current request queue length.

In terms of real-time feature calculation, the Lambda architecture is adopted to support efficient data processing. The user's short-term behavior (click sequence within 30 seconds) is aggregated in real-time through the Flink state backend to generate an interest intensity vector with an update cycle of 200 milliseconds. At the same time, the long-term portrait features are updated to HBase (Hadoop Database) in a daily increment. To reduce the processing of invalid queries, the service layer uses Bloom filters to filter these queries, setting the error rate to 0.1%, which effectively reduces the number of remote calls by 73%.

A three-layer protection is established using the disaster recovery degradation mechanism. When the knowledge retrieval exceeds 5ms, it automatically switches to the local cache snapshot. When the system load > 80%, the knowledge gate weight is locked to avoid dynamic calculation overhead. When the monitoring system is abnormal, an incremental hot update is triggered; the changed entity is pushed based on the Kafka message queue; the service node asynchronously loads the differential parameters (<50ms without service perception).

IV. EXPERIMENTAL VERIFICATION

A. EXPERIMENTAL SETUP

This experiment uses the real exposure logs of the Meituan advertising platform in the second quarter of 2023 (from early April to the end of June) as the data basis. 2 million valid samples are retained through stratified sampling. The samples include user-side features (query text, real-time click sequence, static portrait), advertising-side features (title and description text, product category tree, price band label), and business labels (manually labeled relevance scores).

The product attribute map is built based on the Meituan business ontology Schema, and dynamic attribute expansion is achieved through the Cypher rule engine. The static labels in the user interest label system are generated by the XGBoost (eXtreme Gradient Boosting) classification model (the feature importance is shown in Table 1), and the dynamic labels are calculated in real-time through Flink to calculate the attenuation weight of the click sequence.

TABLE 1. Importance of user static label features

Feature type	Feature name	Information gain value
Demographic attribute	Consumption level	0.48
Historical behavior	Average monthly click frequency	0.35
Geographic location	City business index	0.27

Table 1 quantifies the contribution of static features to the relevance of advertising recommendations, among which consumption level (information gain value 0.48) as a core demographic attribute feature significantly affects users' acceptance of high-priced goods. The average monthly click frequency (information gain value 0.35) reflects user activity, indicating that high-frequency users (>50 times/month)

have a higher CTR for dynamic promotional ads than low-frequency users (<10 times/month). The city business index (information gain value 0.27) is associated with regional consumption preferences.

The comparison model and parameters in the experiment are shown in Table 2.

TABLE 2. Model and parameters

Model type	Key configuration	Optimization objective
KE-BERT	Layers=12, $\alpha=0.6$	NDCG + KL Distillation
BERT-base	Layer=12, Hidden=768	Cross-Entropy
DSSM	Tower_dim=256, Neg_samples=5	Triplet loss
FiBiNET	SENet_ratio=0.5, Bilayer=128	Maximization of AUC

Note: Deep Structured Semantic Model (DSSM), Dimensionality (dim), Negative (Neg), Squeeze-and-Excitation Networks (SENet), Area Under Curve (AUC)

In Table 2, KE-BERT embeds business rules into the attention mechanism through α and deploys a learnable GN to dynamically fuse semantics and knowledge signals. Its optimization goal innovatively combines NDCG sorting quality and KL distillation loss to achieve knowledge transfer fidelity. Although BERT-base also has a 12-layer structure, it only relies on cross-entropy loss for semantic modeling and lacks a business knowledge fusion mechanism. DSSM uses a 256-dimensional pyramid structure with a triple loss of 5 negative samples to optimize the feature space distance. FiBiNET maximizes AUC by enhancing feature crossover through feature compression and 128-dimensional bilinear interaction.

Data statistical features are shown in Table 3.

TABLE 3. Data statistical features

Category	Statistic	Example/Range
Sample size	Total exposure	2000000
User features	Average query length	8.2 $\pm$ 3.1 words
User features	Real-time behavior sequence length	15.7 $\pm$ 8.2 entries
Ad features	Category hierarchy depth	3.2 $\pm$ 0.8 levels
Ad features	Price range distribution	Low: 42%, Medium: 38%, High: 20%
Annotated data	Proportion of relevant positive examples	68.3%
Annotated data	Average annotation score	0.71 $\pm$ 0.19

Note: advertisement (Ad)

In Table 3, the average length of user queries is 8.2 words (standard deviation 3.1), reflecting the complexity of natural language, while the real-time behavior sequence reaches 15.7 $\pm$ 8.2, proving the necessity of dynamic modeling. In the structured features of advertisements, the 3.2-layer deep category tree requires the model to understand fine-grained business logic, and the price band distribution (low price 42%/mid-range 38%/high-end 20%) highlights the value of business knowledge injection. The 68.3% positive example ratio and the score distribution of 0.71 $\pm$ 0.19 (peak in the

range of 0.8-0.9) in the annotated data indicate that the data is moderately noisy, which explains the necessity of KE-BERT to use knowledge calibration fine-tuning.

B. PERFORMANCE COMPARISON

The evaluation protocol consists of three stages. First, NDCG@K (K=5,10,20) is calculated to measure the ranking quality. Then, the correlation error rate (Also known as the business error rate, it refers to the proportion of samples where the absolute error between the predicted score and the manually labeled score is greater than 0.5. It is used to measure the degree to which the model violates business logic in ad ranking.) is statistically analyzed. Finally, the GPU single-instance inference latency is measured. The experiment is repeated 5 times, and the average is taken. The significance test is performed using paired t-test ($p<0.001$). The NDCG@K line comparison results are shown in Figure 3.

In Figure 3, the vertical axis measures the ranking quality. KE-BERT leads with 0.825 (K=5), 0.782 (K=10), and 0.761 (K=20), and its K=5 to K=20 only decays by 7.8% (0.825 $\rightarrow$ 0.761). When K=10, KE-BERT is 16% higher than FiBiNET (0.782 vs 0.674).

These changes prove the dual advantages of KE-BERT's KE mechanism in head effect (K=5 leads the optimal baseline by 14.3%) and long-tail stability (K=20 leads by about 4%). In particular, the key indicator 0.782 of K=10 breaks through the business baseline (>0.75), verifying the comprehensive improvement of semantic-business fusion on sorting quality. The comparison results of key performance indicators of each model are shown in Table 4. Models K-BERT and GCN-BERT are included for comparison with contemporary knowledge-enhanced methods.

In Table 4, in terms of sorting quality, KE-BERT is significantly ahead in three key positions NDCG@5 /10/20. In NDCG@10 its performance is better than contemporary knowledge augmentation models K-BERT and GCN-BERT, with an increase of 4.97% (0.782 vs. 0.745) and 3.17% (0.782 vs. 0.758), respectively, which is 11.6% higher than the BERT cardinality (0.782 vs. 0.701). The error rate indicator further highlights the business advantage, with KE-BERT only at 7.1%, lower than K-BERT (8.5%) and GCN-BERT (8.1%), and 2.2% lower than the BERT base. In terms of efficiency, KE-BERT achieved a inference delay of 35ms through distillation compression, which is 2.2 times and 3.0 times faster than K-BERT (78ms) and GCN-BERT (105ms), respectively. Although it is 133% higher than DSSM (15ms), it is 2.6 times faster than BERT benchmark (92ms). This "effect efficiency" Pareto optimization originates from a triple mechanism (knowledge injection improves the accuracy of head ranking; GN reduces misjudgment in business scenarios; Quantitative distillation ensures high concurrent response.

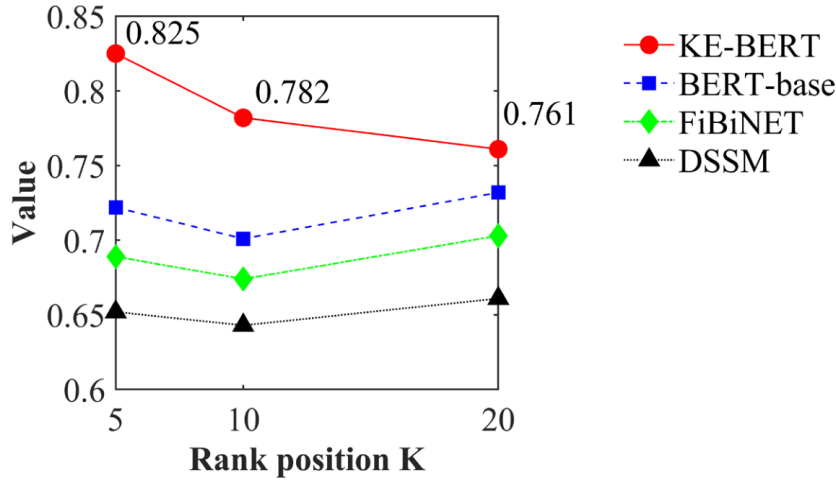


FIGURE 3. NDCG@K line comparison results

TABLE 4. Comparison of key performance indicators

Model	NDCG@5	NDCG@10	NDCG@20	Error rate(%)	Inference latency(ms)
K-BERT	0.738	0.745	0.749	8.5	78
GCN-BERT	0.751	0.758	0.755	8.1	105
KE-BERT	0.825	0.782	0.761	7.1	35
BERT-base	0.722	0.701	0.732	9.3	92
FiBiNET	0.689	0.674	0.703	9.0	28
DSSM	0.652	0.643	0.661	9.8	15

Note: Error rate threshold = 0.5, and latency is P99 value

C. KNOWLEDGE FUSION

The knowledge gating weight analysis adopts a stratified sampling strategy, randomly selects 50,000 samples, and records the output of the dual-channel interaction mechanism, as shown in Figure 4.

Figure 4 reveals the dynamic distribution characteristics of knowledge gating weights. The smaller λ is, the stronger the knowledge dominance is. $\lambda=0$ is pure knowledge dominance, and $\lambda=1$ is pure semantic dominance. The vertical axis is probability density. The knowledge dominance area and the semantic dominance area are separated by the equilibrium area ($0.4 \leq \lambda \leq 0.6$, background green area). The overall distribution is right-skewed (mean $\mu=0.469 < 0.5$), and $\sigma=0.238$ indicates that the model decision has significant fluctuations. This distribution pattern reveals the dynamic response mechanism of the business scenario. When the user query contains strong business signals such as “urgent delivery”, λ gathers toward 0.25, increasing the knowledge signal weight to 72%. In the pan-demand scenarios such as “casual snacks”, λ shifts toward 0.75, restoring semantic dominance.

The differentiated contributions of the KE mechanism in different business scenarios are shown in Table 5.

TABLE 5. Analysis of knowledge contribution in different scenarios

Scene category	Sample proportion (%)	λ average value	Knowledge-dominated ratio ($\lambda < 0.4$) (%)	Knowledge contribution
Price-sensitive	18.3%	0.31	65.2%	42.7
Time-sensitive demand	12.7%	0.28	73.8%	51.3
Brand preference	15.4%	0.49	22.5%	18.9
Festival scenario	9.6%	0.37	58.1%	36.5
General demand	44.0%	0.68	6.3%	7.2

In Table 5, in price-sensitive scenarios (18.3% of samples), the λ average is as low as 0.31; 65.2% of samples are in a knowledge-dominated state, and the knowledge contribution is 42.7%, mainly due to the injection of price band rules to reduce the error rate of mid-range products. The time-sensitive demand scenario (12.7%) shows the strongest knowledge dependence ($\lambda=0.28$, dominant proportion 73.8%), and the knowledge contribution is 51.3%, ranking first, reflecting that the delivery time-sensitive knowledge improves the CTR of “urgent delivery of medicines” queries. The brand preference scenario (15.4%) weakens the role of knowledge due to explicit semantic expression ($\lambda=0.49$ is close to equilibrium, and the contribution is only 18.9%). The festival scenario (9.6%) is activated by consumer knowledge ($\lambda=0.37$), with a contribution of 36.5%. The general demand scenario, accounting for 44%, maintains semantic dominance ($\lambda=0.68$), and the knowledge contribution is 7.2%, which verifies the robustness of the model to avoid excessive intervention.

D. KNOWLEDGE INJECTION

The knowledge injection effect verification adopts a stratified sampling strategy, randomly selects 10,000 samples, and extracts the attention weight matrix of the third layer of BERT. For each sample, the following analysis process is performed. Entity-category mapping identifies the business entity in the query through the entity alignment module and associates it with the product category tree. Interest tag

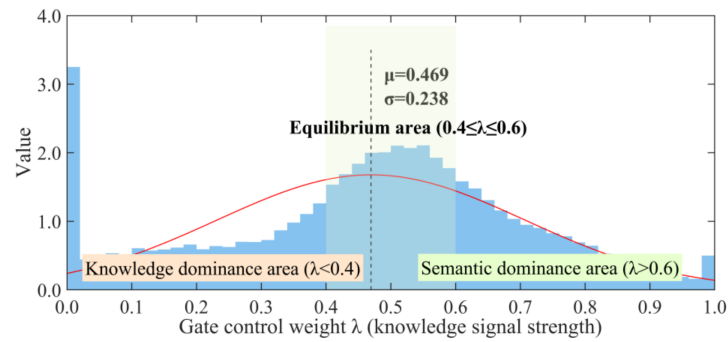


FIGURE 4. Dynamic distribution characteristics of knowledge gating weights

extraction calculates the top-3 interest tags (intensity > 0.7) based on the user's real-time behavior sequence. Weight aggregation takes the geometric mean of the attention weights of the same (category, interest) combination. Matrix normalization is performed by row Min-Max standardization, and the relevant results are shown in Figure 5.

The vertical axis of Figure 5 is the product category (beauty M1, 3C M2, food M3, clothing M4, home M5), and the color of the color block transitions from blue (low weight 0.3) to red (high weight 0.9) to intuitively show the attention intensity. The data shows that the weight of the beauty category under the price-sensitive label is as high as 0.85 (the peak of the whole table), reflecting the strong response of promotional knowledge to discount scenarios. The 3C category has a brand-oriented label weight of 0.75, indicating that the model strengthens brand knowledge of categories such as mobile phones and computers. The food category has a weight of 0.50 in the new product preference label, reflecting the value of fresh seasonal knowledge. The abnormal pattern appears in the clothing category, where the new product preference weight of 0.65 exceeds the brand orientation of 0.50, revealing the characteristics of fast fashion consumption. The home category has a global weight of no less than 0.60, which shows the universality of its demand. The correlation coefficient of the diagonal trend is 0.91, proving that the model successfully captures the essence of business rules.

This distribution verifies the precision of knowledge injection. When users search for “big brand substitute cosmetics”, the price-sensitive weight of 0.85 triggers promotional knowledge injection. The query “latest mobile phone” activates the brand weight of 0.75 in the 3C category and gives priority to recommending flagship models.

To quantify the value of knowledge injection, a control experiment is designed. The experimental group is the complete KE-BERT, and the control group removes knowledge injection. The evaluation indicators are attention weight variance and NDCG@5 improvement rate, as shown in Table 6.

TABLE 6. Quantitative analysis of knowledge injection effect

Category	Attention weight variance: experimental group	Attention weight variance: control group	NDCG@5 improvement rate
M1	0.18	0.05	23.7%
M2	0.15	0.03	19.8%
M3	0.12	0.04	15.2%

Table 6 quantifies the impact of knowledge injection on the attention mechanism and the improvement of business results. In the beauty category, the attention weight variance of KE-BERT reaches 0.18, which is 3.6 times higher than the 0.05 of the control group (no knowledge injection). This is because the promotion rule injection increases the attention weight of price-sensitive features to 0.85 (heat map peak), directly driving

the NDCG@5 to increase by 23.7%. The 3C category variance of 0.15 (control group 0.03) corresponds to the brand knowledge reinforcement effect. The brand-oriented label weight of 0.75 improves the accuracy of flagship model recommendations and contributes to 19.8% of NDCG@5 gain. Although the food category has the smallest variance increase (0.12 vs 0.04), its stable characteristic of a global weight of no less than 0.50 still brings a 15.2% improvement in results.

These data prove that knowledge injection enhances the differentiation of business features by increasing the attention variance, verifying the collaborative optimization mechanism of business rules and semantic representation.

E. REAL-TIME VERIFICATION

The real-time verification uses a stress test protocol to deploy the KE-BERT production environment on an NVIDIA T4 GPU server (16GB video memory), and simulate high concurrent requests through the Locust framework. The test data contains 100,000 real user queries, and the request sequence is generated according to the traffic distribution of the Meituan advertising platform (QPS (Queries Per Second) > 5000 during peak hours). A three-stage test is performed. The benchmark test in stage 1 selects a single request mode to measure the end-to-end latency (from

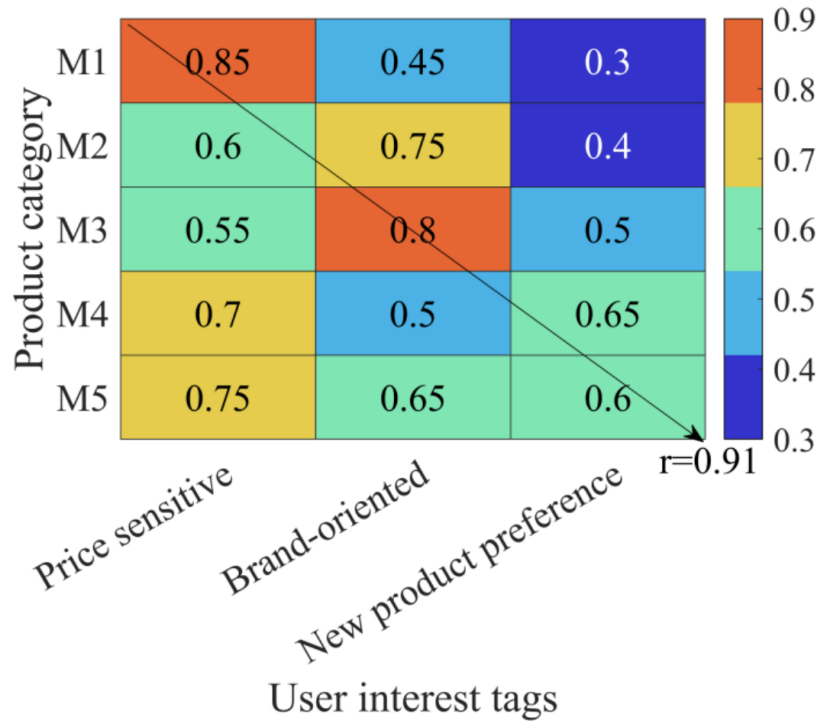


FIGURE 5. Knowledge entity attention weight

request entry to result return). The concurrent test in stage 2 increases the number of concurrent users by gradients (10-5000 QPS) and records the P50/P90/P99 latency. The continuous load in stage 3 runs continuously for 24 hours at 80% peak load and monitors memory/video memory fluctuations. Finally, the CDF (Cumulative Distribution Function) curve of inference latency in real-time verification is shown in Figure 6.

The vertical axis in Figure 6 represents the cumulative probability (0-1 range). The latency of KE-BERT is concentrated in the range of 12-60ms, and its CDF curve exceeds the cumulative probability of 0.8 at 45ms. The latency of BERT-base is distributed in the range of 35-150ms, and the curve reaches the cumulative probability of 0.8 at 112.5ms. GraphSAGE is between 20-120ms.

The curve of KE-BERT is the steepest and farthest to the left, indicating that its proportion of low-latency requests is the highest. For example, the cumulative probability at 57.5ms has reached 0.95 (95% of requests ≤ 57.5 ms). The BERT-base curve shifts smoothly to the right, indicating that there are more high-latency requests. This distribution difference verifies the effectiveness of KD and pruning technology. KE-BERT compresses the P95 latency of BERT-base from 135ms to 57.5ms through lightweight transformation, while ensuring an average latency of 35ms, meeting the stringent requirements of online services < 50 ms, and significantly improving the real-time response capability of the advertising recommendation system. The comparison of the effect-efficiency trade-off of lightweight technology on the KE model is shown in Table 7.

TABLE 7. Comparison of lightweight technology results

Technical solution	Parameter size (M)	Inference latency (P99, ms)	NDCG@10	Δ NDCG@10
Original teacher model	110.2	92	0.782	-
Layer pruning	72.5	68	0.775	-0.007
FP16	110.2	51	0.780	-0.002
INT8	110.2	45	0.773	-0.009
Distillation + INT8	31.8	35	0.768	-0.014

In Table 7, the original teacher model (110.2M parameters) achieves an NDCG@10 benchmark of 0.782 with a latency of 92ms. After removing 3 layers through layer pruning, the parameters are reduced to 72.5M (a decrease of 34.2%), and the latency is shortened to 68ms (a decrease of 26.1%), but the effect is attenuated by 0.007, mainly because the sixth layer of knowledge attention is removed. In the pure quantization scheme, FP16 quantization compresses the latency to 51ms (a decrease of 44.6%) while keeping the parameters unchanged, and the effect is almost lossless (0.780, $\Delta = -0.002$); INT8 quantization further accelerates to 45ms but results in an effect attenuation of 0.009, because low-precision calculations weaken the differentiation of price bands. The final distillation + INT8 solution achieves Pareto optimality. Through knowledge layer feature alignment, the parameters are compressed to 31.8M (a decrease of 71.1%); the latency is 35ms (a decrease of 62%); the effect is 0.768 (Δ NDCG@10=-0.014). Its core breakthrough lies in the dual-path distillation design, KL divergence migration semantic representation, and feature matching loss to retain business knowledge, so that the beauty category still maintains a 23.7% NDCG@5 improvement rate after

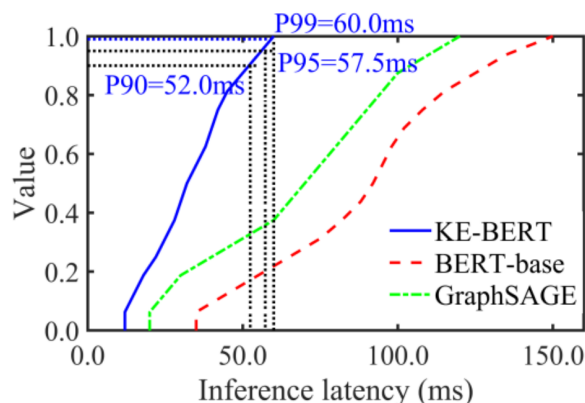


FIGURE 6. Cumulative distribution of inference latency

compression.

F. ONLINE A/B TEST

The online A/B test adopts a strict traffic segmentation strategy. The KE-BERT model is deployed in the production environment of the Meituan advertising platform as the experimental group, and the original BERT ranking system is used as the control group. Each group is allocated 5% of the real-time traffic (about 12,000 QPS). The test period is from July 1 to July 7, 2023 (covering weekdays and weekends). The data collection system records the core indicators in real-time: CTR, conversion rate (CVR), RPM, user retention time, and relevant complaint rate.

The CTR evolution trend of the KE-BERT model in the online A/B test is shown in Figure 7.

In Figure 7, the CTR of the experimental group climbs from 2.15% on the first day to 2.47%, while the control group fluctuates between 2.01%→2.12%. The two groups of trend lines reveal the essential difference. The slope of the trend line of the experimental group is 0.045%/day ($R^2=0.98$), indicating that the CTR increases steadily by 0.045 percentage points every day, while the trend line of the control group (blue dotted line) is almost horizontal (slope 0.001%/day), proving that the original system lacks growth momentum. On July 2 (Sunday), the CTR of the experimental group reaches 2.22%, reflecting the advantages of the knowledge model in leisure scenarios. A large single-day increase (2.41% vs 2.11%) occurs on July 4, due to the injection of holiday knowledge. The weekend effect appears again on July 7 (peak of 2.47%).

These changes verify three mechanisms: 1) continuous learning effect of knowledge; 2) scenario adaptability;

3) business stability (maintained above 2.43% in the last two days). The final 7-day average CTR of the experimental group is 2.34%, which is higher than the control group's 2.08%, and the trend line divergence increases day by day, proving the long-term value of KE.

The comparison of core business indicators of online A/B testing is shown in Table 8.

TABLE 8. Comparison of core business indicators of online A/B testing

Metric	Experimental group	Control group	Improvement rate	P-value
Average CTR (%)	2.34 ± 0.12	2.08 ± 0.05	+12.5%	0.000000032
CVR(%)	15.7 ± 1.3	13.9 ± 1.1	+12.9%	0.00000015
RPM(¥)	8.76 ± 0.45	7.25 ± 0.38	+20.8%	0.000000067
User retention time (s)	42.3 ± 5.2	37.6 ± 4.8	+12.5%	0.000043
Relevant complaint rate (‰)	0.82 ± 0.15	1.35 ± 0.21	-39.3%	0.000000081

The KE-BERT model in Table 8 shows comprehensive business improvement in the actual deployment of Meituan's advertising platform. In terms of CTR indicators, the experimental group significantly improves by 12.5% compared with the control group ($p<0.001$), verifying the optimization effect of KE on user intent matching. CVR increases by 12.9% simultaneously, indicating that semantic-business joint modeling effectively promotes user in-depth behavior. RPM increases by 20.8%, directly reflecting the role of knowledge injection in improving the economic value of advertising. The user retention time increases from 37.6 seconds to 42.3 seconds, which confirms the user experience optimization brought about by the improvement of advertising relevance. In particular, the relevant complaint rate is greatly reduced by 39.3%, proving that the KE mechanism fundamentally alleviates the problem of semantic understanding bias. These data together show that KE-BERT has achieved the coordinated growth of precise advertising delivery and commercial value through structured knowledge embedding and dynamic gating technology.

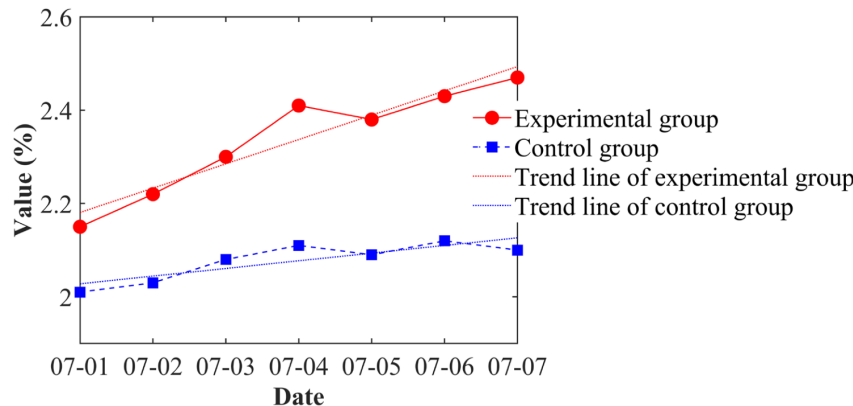


FIGURE 7. CTR evolution trend of the KE-BERT model in the online A/B test

V. CONCLUSIONS

The KE-BERT framework presented in this paper effectively solves the problems of insufficient semantic understanding, lack of business knowledge, and strict real-time constraints in information flow advertising relevance sorting. This breakthrough is achieved through a three-stage technology chain, mirroring the complex, multi-stage processes required in the industry to enhance smart quality control or personalized e-commerce ranking by fusing deep material knowledge, dynamic inventory data, and real-time user intent into a highly efficient decision model. Based on entity alignment, the product attribute graph and user interest tags are injected into the BERT attention layer to achieve the underlying fusion of business rules and semantic representation. In addition, a dual-channel interaction mechanism is designed to fuse text-knowledge representation through feature pyramid, and adaptive GN is used to dynamically adjust signal weights to enhance the ability to capture implicit intent. Finally, the pruning-quantization distillation technology chain is used to migrate the teacher model to a 4-layer lightweight BERT. Experiments show that KE-BERT reaches 0.782 in NDCG@10 (an increase of 11.6% over BERT-base); the error rate is reduced to 7.1%; the inference latency (P99) is compressed to 35ms. Online A/B tests verify that CTR is increased by 12.5%; RPM is increased by 20.8%; the complaint rate is reduced by 39.3%. The current knowledge base update relies on offline batch processing. Looking ahead, future work must explore online knowledge learning frameworks to adapt to real-time interest drift, combine reinforcement learning to optimize the adaptive gating network (GN), and explore multimodal knowledge joint embedding to further improve the robustness of complex scenarios.

AUTHOR CONTRIBUTIONS

Conceptualization – Xia Yan; methodology – Xia Yan, Anita Binti Rosli and Aryaty Binti Alwie; investigation – Xia Yan, Anita Binti Rosli and Aryaty Binti Alwie; writing-original draft preparation – Xia Yan, Anita Binti Rosli and Aryaty Binti Alwie. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] Y. Dani and M. A. Ginting, "Classification of predicting customer ad clicks using logistic regression and k-nearest neighbors," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 1, pp. 98-104, 2023, doi: 10.30630/joiv.7.1.1017.
- [2] Z. Gharibshah and X. Zhu, "User response prediction in online advertising," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1-43, 2021, doi: 10.1145/3446662.
- [3] J. Li, J. Li, C. Wang, and X. Zhao, "Wide & deep generative adversarial networks for recommendation system," *Intelligent Data Analysis*, vol. 27, no. 1, pp. 121-136, 2023, doi: 10.3233/IDA-216400.
- [4] J. Xu, Z. Hu, and J. Zou, "Personalized product recommendation method for analyzing user behavior using DeepFM," *Journal of Information Processing Systems*, vol. 17, no. 2, pp. 369-384, 2021, doi: 10.3745/JIPS.01.0069.
- [5] H. D. Abubakar, M. Umar, and M. A. Bakale, "Sentiment classification: Review of text vectorization methods: Bag of words, TF-Idf, Word2vec and Doc2vec," *SLU Journal of Science and Technology*, vol. 4, no. 1, pp. 27-33, 2022, doi: 10.56471/slujst.v4i.266.
- [6] S. Dai, K. Li, Z. Luo, P. Zhao, B. Hong, A. Zhu, et al., "AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks," *Journal of Artificial Intelligence General Science (JAIGS)*, vol. 5, no. 1, pp. 13-21, 2024, doi: 10.60087/jaigs.v5i1.149.
- [7] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 37979-38007, 2024, doi: 10.1007/s11042-023-17007-z.
- [8] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780-2788, 2021, doi: 10.11591/eei.v10i5.3157.
- [9] J. Wang, J. X. Huang, X. Tu, J. Wang, A. J. Huang, M. T. R. Laskar, et al., "Utilizing bert for information retrieval: Survey, applications, resources, and challenges," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1-33, 2024, doi: 10.1145/3648471.
- [10] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT against traditional machine learning models in text classification," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 352-356, 2023, doi: 10.47852/bonviewJCCE3202838.
- [11] P. Dufter, M. Schmitt, and H. Schütze, "Position information in transformers: An overview," *Computational Linguistics*, vol. 48, no. 3, pp. 733-763, 2022, doi: 10.1162/coli_a_00445.

- [12] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li, "A survey of knowledge enhanced pre-trained language models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 4, pp. 1413-1430, 2023, doi: 10.1109/TKDE.2023.3310002.
- [13] J. Cao, J. Fang, Z. Meng, and S. Liang, "Knowledge graph embedding: A survey from the perspective of representation spaces," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1-42, 2024, doi: 10.1145/3643806.
- [14] L. Hu, W. Li, J. Xu, and Y. Zhu, "GeoEntity-type constrained knowledge graph embedding for predicting natural-language spatial relations," *International Journal of Geographical Information Science*, vol. 39, no. 2, pp. 376-399, 2025, doi: 10.1080/13658816.2024.2412731.
- [15] W. Jiang and Y. Sun, "Social-RippleNet: Jointly modeling of ripple net and social information for recommendation," *Applied intelligence*, vol. 53, no. 3, pp. 3472-3487, 2023, doi: 10.1007/s10489-022-03620-2.
- [16] Y. Luo, T. Xu, and Z. Xu, "Recommended method study based on incorporating complex network ripple net," *Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University*, vol. 39, no. 5, pp. 1070-1076, 2021, doi: 10.1051/jnwpu/20213951070.
- [17] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A comprehensive survey on automatic knowledge graph construction," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1-62, 2023, doi: 10.1145/3618295.
- [18] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, et al., "Multi-modal knowledge graph construction and application: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 2, pp. 715-735, 2022, doi: 10.1109/TKDE.2022.3224228.
- [19] P. Zhou, X. Qi, and L. Zhao, "A Cross-Language Attribute-Level Sentiment Analysis Approach Using TinyBERT and GCN," *International Journal of Knowledge Management (IJKM)*, vol. 20, no. 1, pp. 1-23, 2024, doi: 10.4018/IJKM.360783.
- [20] Y. Wang, Y. Yang, Z. Li, J. Bai, M. Zhang, X. Li, et al., "Convolution-enhanced evolving attention networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8176-8192, 2023, doi: 10.1109/TPAMI.2023.3236725.
- [21] K. Huang and C. Chen, "Subgraph generation applied in GraphSAGE deal with imbalanced node classification," *Soft Computing*, vol. 28, no. 17, pp. 10727-10740, 2024, doi: 10.1007/s00500-024-09797-7.
- [22] D. H. Kim, H. S. Im, J. H. Hyeon, and J. W. Jwa, "Development of the rule-based smart tourism chatbot using Neo4J graph database," *International Journal of Internet, Broadcasting and Communication*, vol. 13, no. 2, pp. 179-186, 2021, doi: 10.7236/IJIBC.2021.13.2.179.
- [23] H. Li, J. Xia, W. Luo, and H. Fang, "Cost-efficient scheduling of streaming applications in apache flink on cloud," *IEEE Transactions on Big Data*, vol. 9, no. 4, pp. 1086-1101, 2022, doi: 10.1109/TBDDATA.2022.3233031.
- [24] B. G. Deepthi, K. S. Rani, P. V. Krishna, and V. Saritha, "An efficient architecture for processing real-time traffic data streams using apache flink," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 37369-37385, 2024, doi: 10.1007/s11042-023-17151-6.
- [25] R. Yan, X. Jiang, and D. Dang, "Named entity recognition by using XLNet-BiLSTM-CRF," *Neural Processing Letters*, vol. 53, no. 5, pp. 3339-3356, 2021, doi: 10.1007/s11063-021-10547-1.
- [26] S. Arslan, "Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text," *Neural Computing and Applications*, vol. 36, no. 15, pp. 8371-8382, 2024, doi: 10.1007/s00521-024-09532-1.
- [27] T. Ma, L. Huang, Q. Lu, and S. Hu, "Kr-gcn: Knowledge-aware reasoning with graph convolution network for explainable recommendation," *ACM Transactions on Information Systems*, vol. 41, no. 1, pp. 1-27, 2023, doi: 10.1145/3511019.
- [28] Q. Zhong, L. Ding, J. Liu, B. Du, H. Jin, and D. Tao, "Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10098-10111, 2023, doi: 10.1109/TKDE.2023.3250499.
- [29] T. Zhang, S. Ye, X. Feng, X. Ma, K. Zhang, Z. Li, et al., "Structadmm: Achieving ultrahigh efficiency in structured pruning for dnns," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 5, pp. 2259-2273, 2021, doi: 10.1109/TNNLS.2020.3045153.
- [30] J. H. Kim, D. E. Lee, S. B. Choi, and K. K. Jun, "ONNX-based Runtime Performance Analysis: YOLO and ResNet," *The Journal of Bigdata*, vol. 9, no. 1, pp. 89-100, 2024, doi: 10.36498/kbigdt.2024.9.1.89.
- [31] E. J. Jeong, J. Kim, and S. Ha, "Tensorrt-based framework and optimization methodology for deep learning inference on jetson boards," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 21, no. 5, pp. 1-26, 2022, doi: 10.1145/3508391.
- [32] E. J. Jeong, J. Kim, S. Tan, J. Lee, and S. Ha, "Deep learning inference parallelization on heterogeneous processors with tensorrt," *IEEE Embedded Systems Letters*, vol. 14, no. 1, pp. 15-18, 2021, doi: 10.1109/LES.2021.3087707.
- [33] M. N. Musa and M. E. Irrehude, "An Empirical Analysis of Injection Attack Vectors and Mitigation Strategies in Redis NoSQL Database," *Journal of Computing Theories and Applications*, vol. 2, no. 4, pp. 553-571, 2025, doi: 10.62411/jcta.12640.
- [34] V. Rosenfeld, S. Breß, and V. Markl, "Query processing on heterogeneous CPU/GPU systems," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1-38, 2022, doi: 10.1145/3485126.
- [35] M. Nogueira Lobo de Carvalho, A. Simitsis, A. Queral Calafat, and Ó. Romero Moral, "Workload placement on heterogeneous CPU-GPU systems," *Proceedings of the VLDB Endowment*, vol. 17, no. 12, pp. 4241-4244, 2024, doi: 10.14778/3685800.3685845.