Automatic News Headline Generation and Topic Consistency Assessment Based on SimCSE and Knowledge Graph

Yun Liu[1]
[1]Department of Journalism, Sanda University, Shanghai 201209, China
Corresponding author: Yun Liu (e-mail: anlan1919@163.com)

ABSTRACT To address the problems of insufficient semantic alignment and topic drift in industry news and market report headline generation, this paper proposes a method that fuses SimCSE semantic constraints with knowledge graph entity associations, paying particular attention to key entities related to names of people, organizations, locations, and fields. Considering the growing demand for reliable information processing in intelligent communication systems and electromagnetic information transmission environments, the proposed framework provides a semantic enhancement mechanism that is beneficial for trustworthy content understanding in data-driven applications. Based on SimCSE, the news text is encoded at the sentence level, and contrastive learning is used to improve semantic representation consistency. Named entity recognition is performed and aligned with the Wikidata knowledge graph to construct a local knowledge subgraph containing entity relationships. A graph attention network is used to fuse SimCSE sentence embeddings with knowledge graph entity embeddings to generate knowledge-enhanced representations. The fused representation is applied during the Transformer decoding process to dynamically focus on key entities and core semantics, generating thematically consistent and factually accurate headlines. Experimental results show that the proposed method achieves an average BERTScore of at least 0.87 and an average BLEU score of at least 0.75 across different news domains. In terms of topic consistency, the Topic-F1 score remains at 0.835 ± 0.019 under the highcomplexity condition of four topics, and the topic deviation is as low as 0.165 ± 0.021. In scenarios involving the summarization of complex manufacturing processes or market trend analysis, this study effectively addresses the issues of insufficient semantic alignment and topic drift while providing a reference for semantic information processing in intelligent electromagnetic communication systems.

INDEX TERMS news headline generation, knowledge graph, semantic consistency, electromagnetic information processing, intelligent communication

I. INTRODUCTION

MARKET report headlines serve as semantic anchors for information dissemination, fulfilling the dual functions of condensing event cores and guiding cognitive pathways [1,2]. With the rise of automated content generation, building intelligent systems that accurately capture textual main ideas—whether on fast fashion trends or raw material price fluctuations—while preserving factual coherence has become a key NLP requirement [3,4]. High-quality headlines demand not only linguistic conciseness but also semantic fidelity and knowledge credibility to avoid loss of key elements during compression [5,6]. Integrating semantic representation with structured knowledge and injecting interpretable constraints into generative models is emerging as a vital direction for enhancing machine-generated text credibility [7,8].

This research proposes a semantically credible and factually controllable headline generation framework to address semantic drift and knowledge disconnection. Unlike sequence-only approaches, it introduces a dualtrack architecture: (1) SimCSE (Simple Contrastive Learning of Sentence Embeddings) encodes news texts into semantically compact representations, strengthening core proposition understanding; (2) named entity recognition aligns entities with Wikidata to build dynamic local subgraphs capturing explicit semantic relations. These tracks are fused in embedding space via a Graph Attention Network (GAT), producing a context vector enriched with semantic density and knowledge structure. This vector conditions the Transformer decoder, modulating attention weights to dynamically focus on key entities and their context. Beyond standard metrics, we design a topic fidelity scorer based on weighted semantic similarity and entity consistency. Contributions include: first, a SimCSE–knowledge graph

semantic enhancement mechanism improving internal consistency ; second, a GAT-based knowledge fusion module enabling fine-grained injection of structured knowledge (e.g ., linking fabric properties to manufacturing entities); third, a comprehensive topic fidelity evaluation scheme shifting quality assessment from surface matching to deep semantic alignment. This framework offers a scalable pathway for reliable automated summarization and news generation in the industry.

II. RELATED WORK

In recent years, researchers have sought to improve generation quality by integrating external knowledge and semantic modeling. Some approaches adopt pre-trained models like BART [9,10] and T5 [11] as backbones, achieving strong performance on large corpora but still struggling with generalization in domain-specific news. Mishra P proposed a hierarchical title generation method based on GPT-2 and T5 that produces semantically relevant, concise titles for learning paths and resources, showing high quality in both automatic metrics and human evaluation, thereby supporting topic preview and interest stimulation in online learning [12]. To enhance semantic consistency, sentence embedding models have been used for content matching supervision; for instance, Sentence-BERT [13,14] computes semantic similarity between generated titles and source summaries as a reinforcement learning reward. Wang R provided a systematic review of fine-grained entity type recognition, covering classification schemes, methods, resources, applications, corpus generation strategies, challenges from label noise, and future directions [15].

To jointly optimize semantic representation and knowledge fusion, recent work explores graph neural networks and contrastive learning. Contrastive learning has proven effective in reducing embedding distances between positive pairs while increasing separation from negatives, thus improving semantic discriminability [16,17]. He W introduced an unsupervised short text clustering model combining SimCSE with data augmentation and embedding optimization, alleviating feature sparsity and achieving strong results on multiple benchmarks [18]. In knowledge fusion, graph attention networks (GATs) [19,20] exhibit powerful relational reasoning by capturing high-order correlations from local knowledge subgraphs. Li S proposed an entity linking model integrating GNNs and cross-attention, which enhances entity embeddings with fine-grained semantics and re-ranks candidates, significantly boosting linking accuracy across domains [21]. While these methods confirm the viability of jointly modeling semantic enhancement and structured knowledge, two limitations persist: (1) knowledge subgraph construction depends heavily on accurate entity linking and is sensitive to recognition errors; (2) fusion between graph outputs and decoder states remains coarse-grained, lacking dynamic weighting based on semantic importance. These issues constrain model robustness and accuracy in complex news scenarios. This paper addresses them by deeply coupling SimCSE's high-fidelity sentence embeddings with a knowledge graph–driven entity relationship graph, using GAT for fine-grained multimodal interaction, and applying a joint attention mechanism during decoding to anchor generation to core semantics and key facts.

III. CONSTRUCTION OF TITLE GENERATION FRAMEWORK WITH SIMCSE ENHANCED SEMANTIC REPRESENTATION

Figure 1 shows the news headline generation architecture that integrates SimCSE with a knowledge graph.

The system first performs two-channel processing on the news text: the SimCSE encoder generates sentence embeddings through contrastive learning, while the Named Entity Recognition (NER) module extracts entities and aligns them with Wikidata to construct a knowledge subgraph. Subsequently, a graph attention network aggregates entity relationships to generate enhanced entity embeddings. In the fusion layer, semantic and entity attention interact through cross-modal interaction to generate context-aware representations. In the decoding phase, this representation serves as the initial state and key-value memory input to the Transformer decoder, where a dynamic two-channel attention mechanism enables the collaborative generation of semantics and knowledge. The final output is evaluated for both semantic consistency and entity fidelity to ensure the topical fidelity of the headline.

NEWS TEXT SENTENCE ENCODING BASED ON SIMCSE SENTENCE-LEVEL SEMANTIC ENCODING MECHANISM BASED ON CONTRASTIVE LEARNING

The original news text is automatically segmented into independent sentence units based on semantic boundaries, ensuring that each unit carries complete propositional information. These sentences are then sequentially fed into the SimCSE encoder, which is based on the BERT architecture and optimized using a contrastive learning strategy. During training, to mimic the semantic invariance of natural language, the system performs a double input process on the same sentence: after the first input generates a base sentence vector, a random dropout perturbation is applied to the second input, generating pairs of semantically equivalent positive examples with slightly different vector distributions. This mechanism eliminates the need for manual labeling and constructs high-quality semantic comparison groups, effectively enhancing the discriminative power of the embedding space.

The core optimization is achieved through a contrastive loss function:

$$L_{\text{cont}} = -\log \frac{\exp(\text{sim}(u, v)/\tau)}{\sum_{\text{negative samples}} \exp(\text{sim}(u, w)/\tau)} \quad (1)$$

In the formula, $\text{sim}(u, v)$ represents the cosine similarity between vectors u and v, τ is the temperature coefficient (set to 0.05), and the denominator summarizes the similarity

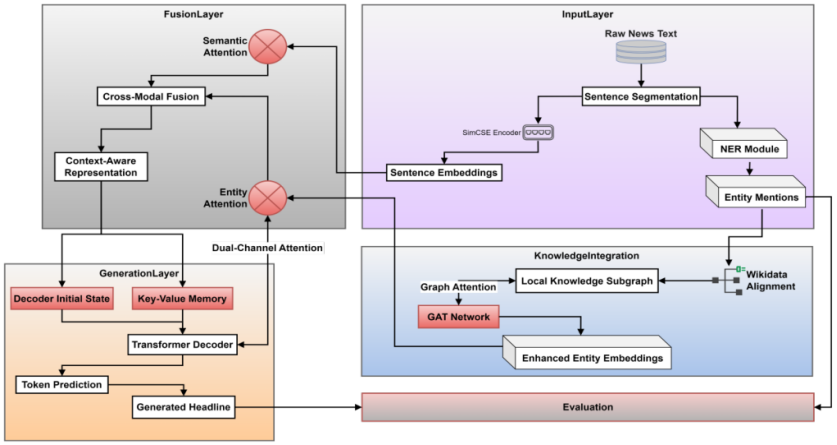

FIGURE 1. News Headline Generation Architecture Integrating SimCSE and Knowledge Graph

index of all negative sample pairs. Except for positive samples, all negative samples are constructed from other samples in the same batch. This function forces the model to cluster semantically consistent sentence vectors tightly together while pushing out unrelated sentence pairs, thereby constructing a semantic space with clear boundaries.

During the inference phase, the system only requires a single forward propagation to generate a dense vector representation for each sentence, avoiding repeated computational overhead.

SEMANTIC ANCHORING AND KEY INFORMATION ENHANCEMENT OF SENTENCE VECTOR SEQUENCES

Based on this, the system selects several sentences with the highest semantic centrality as topic anchor sentences. The number of these sentences is dynamically adjusted based on the length of the original text (typically retaining 3 to 5). The vector representations of these anchor sentences are individually labeled for key reference during the decoding phase. For the remaining non-core sentence vectors, a gated suppression mechanism is applied: a learnable semantic threshold is set. If the cosine value of a sentence vector with the highest similarity anchor sentence falls below this threshold, the sentence is considered to have low relevance, and its influence is significantly reduced during the subsequent fusion phase. Specifically, the inhibition weight is calculated as:

$$\alpha_t = \sigma\left(k \cdot \left(\cos(h_t, h_{\text{anchor}}) - \theta\right)\right) \tag{2}$$

The initial threshold θ controls the lower bound of semantic relevance retention, while the adaptive step size k adjusts the steepness of the inhibition function. The final weights are clipped to ensure the inhibition ratio does not exceed the maximum inhibition rate.

Table 1 shows the parameter configuration of the dynamic gated suppression mechanism, which controls the intensity of filtering low-relevance sentences during semantic encoding. The initial threshold (0.4–0.8) determines the minimum retention threshold for sentence vectors, the adaptive step size (0.001–0.03) adjusts the dynamic adjustment range of the threshold, the maximum suppression ratio (0.5–0.9) limits the upper limit of information loss, the smoothing factor (0.01–0.2) prevents sudden changes in weights, and the reset period (20–300 steps) periodically restores the initial state to prevent over-suppression.

KNOWLEDGE GRAPH ENTITY EXTRACTION AND ASSOCIATION CONSTRUCTION

NAMED ENTITY RECOGNITION AND STANDARDIZED MAPPING BASED ON MULTI-GRAINED MATCHING

A triple alignment strategy is designed that integrates string similarity, contextual semantic matching, and type constraints. First, each entity mention is used as a query item to retrieve a set of candidate entities with highly similar names or aliases in Wikidata; second, a lightweight semantic encoder is used to calculate the semantic similarity between the context window where the mention is located and the candidate entity description text, and the most semantically consistent matches are screened out; finally, a type consistency check is applied to exclude erroneous links that do not match the category, such as excluding "apple" from the "company" class entity if the context clearly points to the meaning of fruit. The process can be formalized as follows:

$$e^* = \arg\max_{e \in C(m)} \left[\alpha \cdot \text{strsim}(m, e) + \beta \cdot \text{semsim}(m, e) + \gamma \cdot I_{\text{type}}(m, e)\right] \tag{3}$$

In the formula, m represents the entity mention, $C(m)$ represents its candidate set, strsim represents the normalized edit distance, semsim represents the contextual encoding similarity, I_{type} represents the type matching indicator function, and α, β, and γ represent learnable weight coefficients. Through this joint decision-making mechanism, the system achieves standardized entity mapping with high recall and low misconnection, ensuring the accuracy of subsequent knowledge graph queries.

TABLE 1. Dynamic Gated Inhibition Parameters

Initial Threshold	Adaptive Step Size	Max Suppression Ratio	Smoothing Factor	Reset Cycle
0.6	0.01	0.7	0.1	100
0.5	0.02	0.6	0.15	200
0.7	0.005	0.8	0.05	50
0.4	0.03	0.5	0.2	300
0.8	0.001	0.9	0.01	20

To construct a semantic context window, this paper employs a text function classifier based on rule and keyword matching to automatically divide news texts into five functional paragraphs: introduction, event, location, participants, and result. This classifier is designed according to the inverted pyramid structure of news articles, combining location priors with domain keywords for sentence classification.

Figure 2 systematically illustrates the semantic activation intensity of 10 key named entities within five semantic context windows. Semantic density is defined as the cosine similarity between the SimCSE-encoded entity mention vector and the corresponding context window sentence vector. The data is derived from the SimCSE model's context-aware encoding and entity matching of news articles. The horizontal axis represents the five functional contexts: introduction, event, location, participants, and outcome, while the vertical axis represents the identified entities, forming a 10×5 semantic density matrix. Each cell's color reflects its semantic relevance. The data distribution shows that "President" exhibits the strongest binding in the "Introduction" (0.92) and "Event" (0.88), indicating its role as the narrative thread running through the core of the report. "Prime Minister" reaches 0.93 in the "Participants" window, highlighting its key role in multiparty interaction. "Capital" peaks at 0.95 in the "Place" window, confirming the strong alignment between the geographic entity and the spatial description. Notably, "Election" reaches 0.89 in the "Result" window, but is below 0.7 in all other areas, reflecting its post-thematic nature. "Sanctions" is only 0.59 in the "Event" window, but rises to 0.78 in the "Result," reflecting its increasing influence as the event evolves. Overall, high values are concentrated in the first four columns, consistent with the "inverted pyramid" structure of news.

DYNAMIC CONSTRUCTION OF LOCAL KNOWLEDGE SUBGRAPHS AND RELATIONSHIP PATH EXTRACTION

After completing entity normalization, starting with each mapped entity, its directly associated adjacent triples from Wikidata are extracted to construct a local knowledge subgraph for the current news text. Specifically, the Wikidata SPARQL endpoint is called to query the outbound and inbound relationships of each entity node, its attribute values and relationships are obtained with other entities, and a first-order neighborhood structure centered on that entity is formed. To control the size of the subgraph and focus on topic-related knowledge, a semantic relevance threshold is set, retaining only relationship types with a high co-occurrence frequency with the core news event theme and removing meta-attributes that are too general. The knowledge base utilized in this study is a fully curated Wikidata open knowledge graph, neither pre-built nor trimmed. The local knowledge subgraph is dynamically generated based on entities identified in each input news article. Specif ically, it queries first-order neighbor triplets of each standardized entity in real-time and dynamically filters relationship edges according to contextual relevance. This ensures the subgraph maintains high alignment with the current news event while preventing cross-sample information leakage.

All extracted triples are weighted and aggregated according to the strength of entity co-occurrence to generate a connected knowledge fragment:

$$G_{\text{local}} = \bigcup_{e \in E} \{(e, r, e') \mid \text{sim}_{\text{ctx}}(e, e') > \delta\} \tag{4}$$

In formula, E is the set of identified entities, r represents the relationship type, sim_{ctx} is the semantic cooperation between the two entities within the context window in the original text, and δ is the dynamic threshold. This subgraph not only contains explicitly mentioned inter-entity relationships but also completes implicit associations through coreference resolution and contextual inference (Using a Chinese coreference resolution model based on SpanBERT).

In Figure 3, the node and edge density data are derived from real sampling sequences in the Wikidata knowledge graph construction log. The relation type distribution data comes from the actual co-occurrence relations of entity pairs in the Chinese news corpus constructed in this study, and is mapped to the Wikidata relation system after manual verification. Figure 3(a) shows the evolution of the number of entity nodes and edge density with the number of iterations during the knowledge subgraph expansion process. The horizontal axis represents the 1st to 5th iteration of the knowledge graph construction. The left vertical axis shows the increase in the number of nodes from the initial 8 to 32, showing a nearly linear growth trend, reflecting the cumulative effect of entity extraction. The right vertical axis shows the decrease in edge density from 0.25 to 0.10, which conforms to the characteristic decay curve of scale-free networks. Notably, an inflection point occurs during the third iteration, where a node count of 22 and an edge density of 0.15 achieve an optimal balance. This phenomenon indicates that the system has captured a significant number of key relationships within the text at this stage. The relationship type distribution in Figure 3(b) reveals the semantic structural characteristics of the local knowledge subgraph. The

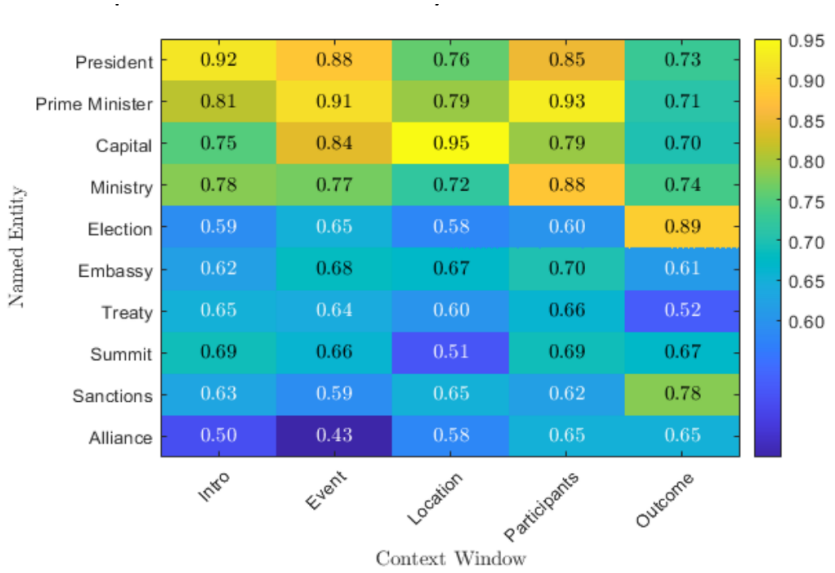

FIGURE 2. Contextual Semantic Density of Named Entities

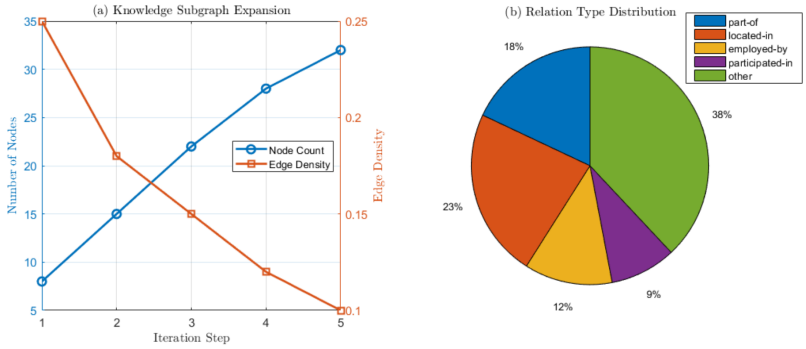

FIGURE 3. Correlation Analysis between Knowledge Subgraph Expansion Dynamics and Entity Relationship Distribution

data shows a high proportion (23%) of spatial relationships, confirming the spatial narrative preference of news texts; occupational relationships, such as "employed-by," account for 12%, reflecting a stable pattern of social relationship expression. Particularly noteworthy is the 38% proportion of the "other" category, a difference attributed to the dynamic thresholding strategy employed in this study, which preserves more marginal semantic relationships. Together, these data validate the trade-off between quality and scale in knowledge subgraph construction and provide an empirical basis for dynamic threshold selection.

CONTEXTUAL REPRESENTATION GENERATION BY INTEGRATING SEMANTICS AND KNOWLEDGE

ENTITY SEMANTIC AGGREGATION MECHANISM BASED ON GRAPH ATTENTION NETWORK

To effectively utilize structured knowledge, this paper employs a graph attention network (GAT) to semantically aggregate entity nodes within a local knowledge subgraph. The input is the entity-relationship graph constructed in the previous stage, where each node corresponds to an aligned, standardized entity, and edges represent semantic relationships between them. To initialize the node representation, an initial vector for each node is extracted from the entity embedding space pre-trained on Wikidata. This embedding is obtained through self-supervised learning of large-scale knowledge graph triples and contains the entity's ontological attributes and general semantic roles. Based on this, GAT iteratively updates the node state through a multilayer attention mechanism to capture the implicit semantics in high-order relational paths. This study employs a two-layer GAT architecture, with each layer containing four attention heads and a hidden dimension of 768 aligned with SimCSE's sentence embedding dimension. The nonlinear activation function uses LeakyReLU (negative slope α=0.2), and residual connections are applied between layers to stabilize training. When the number of GAT layers increases from 1 to 2, the topic F1 score improves; however, performance tends to saturate or even slightly declines after three layers, indicating that two-hop relationship aggregation is sufficient to capture key semantic paths in local knowledge subgraphs, while deeper architectures may introduce noise.

In each layer of attention calculation, the target node updates its own representation by weighted aggregation of

information from its neighboring nodes. For node i, the attention weight between it and its neighbor node j is determined by the compatibility of their current representations:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^\top [Wh_i \| Wh_j]))}{\sum_{k \in N(i)} \exp(\text{LeakyReLU}(a^\top [Wh_i \| Wh_k]))} \quad (5)$$

In formula, h_i and h_j are the input vectors of the node, W is a learnable linear transformation matrix used to map the input to the attention space, $a^\top$ is the weight vector in the attention mechanism, $\|$ represents the vector concatenation operation, and $N(i)$ is the node's neighbor set.

CROSS-MODAL SEMANTIC ALIGNMENT AND FUSION CONTEXT GENERATION

After obtaining the enhanced entity representation, it needs to be deeply integrated with the text semantic space to construct a unified contextual representation. Because entity embeddings and sentence vectors originate from different modalities and encoding systems, direct concatenation can result in misaligned representation spaces. Therefore, a cross-attention alignment fusion mechanism is designed. Specifically, the entity vector set output by GAT is considered the "knowledge source," and the sentence vector sequence encoded by SimCSE is used as the "semantic sink." Semantic alignment is achieved through bidirectional attention interaction.

First, each sentence vector is used as a query vector, and a soft attention operation is performed on the entity set to obtain the key entity information of the sentence:

$$c_s = \sum_e \beta_{se} \cdot h_e \quad (6)$$

$$\beta_{se} = \frac{\exp(h_s^\top h_e)}{\sum_{e'} \exp(h_s^\top h_{e'})} \quad (7)$$

h_s is a sentence vector, h_e is the entity vector, β_{se} represents the sentence's attention intensity on entity e, and c_s is the generated entity context vector. This vector is then gated and fused with the original sentence vector to generate a hybrid representation that combines linguistic semantics and external knowledge. This operation on all sentences forms an enhanced context sequence. Finally, a global fused vector is generated by weighted pooling this sequence based on semantic centrality, which serves as the initial hidden state and additional conditional input for the subsequent decoder.

CAPTION DECODING AND GENERATION BASED ON FUSION REPRESENTATION

FUSION CONTEXT-GUIDED DECODING STATE INITIALIZATION AND CONDITIONAL INJECTION

To ensure that the generation process is controlled by the joint constraints of semantics and knowledge from the outset, this paper applies a fused context vector generated in the previous stage into the initial state construction of the Transformer decoder. This vector is mapped to the decoder's hidden space dimensions through a nonlinear transformation and injected as the hidden state bias term in the first time step of the first layer of the decoder, participating in the subsequent recursive calculation of autoregressive generation. In a specific implementation, a gated fusion mechanism is used to weightingly integrate the external context with the start mark ([BOS]) representation after standard position encoding to form an initial decoding state rich in topic priors.

KEY INFORMATION FOCUSING UNDER DYNAMIC DUAL-CHANNEL ATTENTION MECHANISM

To continuously track key semantic units during the generation process, this paper applies a dynamic dualchannel attention mechanism in the decoder's attention layer, focusing on core sentences and key entities, respectively. Specifically, at each decoding time step, the model performs two external attention calculations in parallel: first, soft attention to the enhanced sentence vector sequence based on the current decoding state to identify the most relevant semantic units at the moment; second, attention allocation to the knowledgeenhanced entity vector set to capture structured facts that should be emphasized. The two attention weights are independently calculated, normalized, and fused to form a comprehensive attention distribution:

$$\beta_t = \lambda \cdot \text{softmax}(q_t^\top K_s) + (1 - \lambda) \cdot \text{softmax}(q_t^\top K_e) \quad (8)$$

$q_t^\top$ is the query vector at the current time step, K_s and K_e are the key matrices consisting of sentence vectors and entity vectors, respectively, and λ is learnable balance coefficients that control the relative weights of semantic and knowledge attention. The fused attention result is used to weightedly aggregate the corresponding value vectors to generate a context-aware attention output, which is fed into the feedforward network layer.

IV. MULTI-DIMENSIONAL QUANTITATIVE EVALUATION OF GENERATED TITLES

EXPERIMENTAL DATA

This study constructs an experimental dataset based on a large-scale Chinese news corpus, covering six major areas: politics, economy, science and technology, sports, health, and education, with a total of 12,843 news samples. All texts are from public reports of authoritative media and have been processed through deduplication, cleaning, and format standardization to ensure the authenticity and timeliness of the content. The original text and manually written headline of each news article are retained as the generation target and evaluation benchmark. To ensure the quality of entity recognition and knowledge alignment, the SpaCy enhanced Chinese NER tool is used to extract key entities such as names, institutions, and locations, and stan dardized links are made through the Wikidata knowledge base to construct a high-precision entity mapping table. This dataset

combines domain breadth and semantic depth, providing a solid foundation for verifying the model's topic fidelity and factual fidelity under different text complexities. The model was trained for 3 epochs on a single NVIDIA A100 GPU using the Adam optimizer (learning rate 2e-5, batch size 32) and the early stopping policy was based on the validation set loss.

To ensure rigorous evaluation and model generalization capabilities, this study employs a topic-entity joint deduplication strategy. First, stratified sampling is conducted by news topic categories to ensure proportional coverage across six major domains. Second, during partitioning, intersection detection is performed on standardized Wikidata entity sets to eliminate samples sharing core entities (e.g., enterprises, manufacturing processes, fabric types) between training, validation, and test sets. The final partition ratio of 7:1:2 has been validated to prevent theme or key entity overlaps across datasets, effectively avoiding performance overestimation caused by information leakage.

SEMANTIC RELEVANCE CHARACTERISTICS AND KEY SENTENCE SCREENING MECHANISM

Based on the SimCSE model, news texts are encoded at the sentence level to generate high-dimensional semantic vectors. A complete similarity matrix is constructed by calculating the cosine similarity between all sentence pairs. Core sentences are selected based on their average similarity exceeding a preset threshold.

A comprehensive scoring method incorporates three dimensions: semantic centrality (based on the global correlation strength of the similarity matrix), position weighting (to emphasize the importance of the first and last sentences), and a length normalization factor. The evaluation first verifies the effectiveness of SimCSE encoding by observing the cohesion of semantic clusters through heat maps. The score distribution in the scatter plot is then used to test the scoring function's ability to distinguish core sentences, ensuring that they are significantly higher than non-core sentences.

Figure 4 combines heatmaps and scatterplots to illustrate SimCSE's semantic relevance and key sentence selection. The symmetrical heatmap in (a) shows that the first 15 core sentences form a high-similarity cluster (red upper-left quadrant), reflecting SimCSE's contrastive learning, which pulls semantically equivalent sentences closer in embedding space. The scatterplot in (b) confirms the effectiveness of the scoring strategy combining semantic centrality and position weighting: key sentences (red) all score above 1—exceeding the mean (black dashed line). Together, these results show that SimCSE captures hierarchical text semantics, while the dynamic scoring accurately identifies core propositions, providing robust semantic anchors for headline generation and mitigating topic drift.

IMPACT OF THRESHOLD ON ENTITY LINKING AND KNOWLEDGE GRAPH STRUCTURE

Based on the entities identified in the news text, the entity matching threshold δ is gradually adjusted, and multiple rounds of matching experiments are conducted from lower to higher thresholds. The matching precision and recall rate are recorded at each threshold. At the same time, the matched entities and their relationships are mapped into the knowledge graph, and the corresponding local subgraph is constructed. The structural characteristic indicators of this subgraph, including the average node degree and graph density, are statistically analyzed. The calculation of precision and recall relies on a standard set of manually annotated entity pairs, while the structural indicators are directly derived from the graph adjacency matrix. Through this joint measurement, it is possible to quantify the combined impact of threshold changes on entity linking accuracy and knowledge graph structure.

Figure 5 shows the changing trends in entity recognition and mapping precision, recall, and local knowledge subgraph structural characteristics under different thresholds δ. Figure 5(a) illustrates the trade-off in entity linking performance with the threshold: as δ increases from 0.1 to 0.9, precision rises from 0.72 to 0.96, demonstrating that stricter matching criteria significantly reduce false positives. However, at the same time, recall drops from 0.95 to 0.70, indicating that some true entities are filtered out by overly strict criteria, resulting in reduced coverage. This illustrates the typical trade-off between precision and recall in entity matching. Figure 5(b) shows how the average degree and graph density of a local knowledge subgraph change with the threshold: As the threshold increases, the average degree decreases from 6.8 to 3.5, and the graph density decreases from 0.42 to 0.22, indicating a decrease in the number of edges and a gradual sparsification of the network. This sparsification helps eliminate noisy relationships and highlight key connections between core entities, but it can also weaken the subgraph's ability to capture complex event context. This result suggests that in knowledge graph construction, the threshold δ is a crucial parameter for balancing knowledge accuracy and structural richness, and its selection requires careful consideration and optimization based on specific task requirements.

SEMANTIC MATCHING COMPARISON

BERTScore and BLEU (Bilingual Evaluation Understudy) are used as metrics to quantify semantic alignment between the generated headline and the first paragraph of the original text. BERTScore quantifies semantic fidelity by measuring the cosine similarity of contextual embeddings, effectively circumventing the limitations of surface-level lexical matching. BLEU measures n-gram overlap, reflecting the standardization of language structure. The proposed SimCSE model is compared with currently popular standards SimCSE, TextRank, BART (fnlp/bart-base-chinese) and T5 (google/mt5-base) across six news categories: pol-

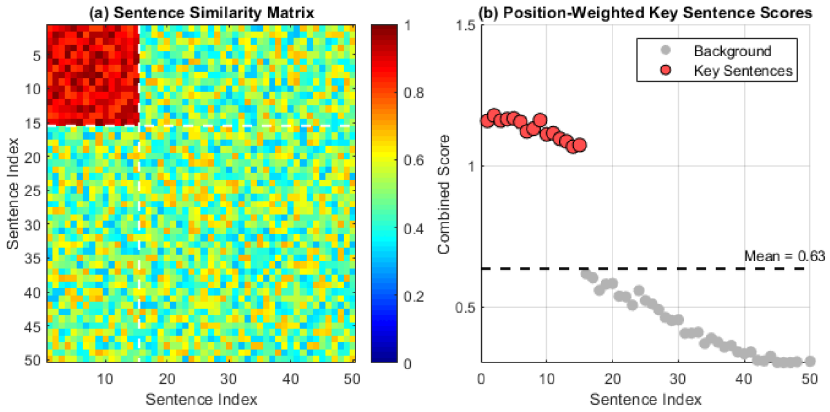

FIGURE 4. Semantic Association Characteristics and Key Sentence Screening Mechanism

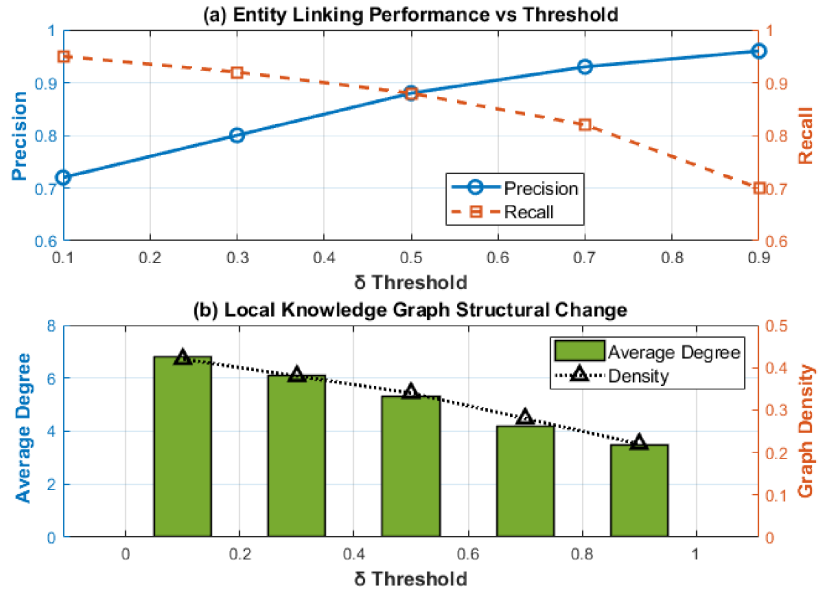

FIGURE 5. The Impact of Threshold on Entity Linking and Knowledge Graph Structure

itics, economy, science and technology, sports, health, and education.

Figure 6 shows the semantic alignment comparison results of the proposed SimCSE model in six news categories with four baseline methods. Figure 6(a) uses BERTScore to assess the deep semantic alignment between the generated headline and the first paragraph of the original article, while Figure 6(b) uses BLEU to measure the consistency of language structure at the n-gram level. The results show that our method achieves the best performance in all categories, with an average BERTScore of at least 0.87 and an average BLEU of at least 0.75. This is attributed to the model's fusion of sentence embeddings enhanced by SimCSE with knowledge graph-driven entity relationships, which effectively anchors the core semantics and key factual chains of events, avoiding the topic drift caused by semantic diffusion during information compression by traditional methods. In contrast, TextRank relies solely on surface word frequencies and co-occurrence statistics, making it difficult to capture deep semantic relationships. While BART and T5 possess strong language modeling capabilities, they lack explicit knowledge constraints, making them prone to generating fluent but factually biased content. While standard SimCSE improves semantic representation quality, it lacks structured knowledge guidance and remains inferior to our proposed method. The error bar distribution demonstrates that our proposed method maintains stable performance across different domains, validating its cross-domain adaptability and robustness. Although human-generated headlines serve as the final quality benchmark, the automated metrics (BLEU/BERTScore) use the first paragraph as a reference because it preserves complete event elements and avoids superficial mismatches caused by headline rhetoric, making it more conducive to objectively assessing semantic fidelity. It should be noted that the BLEU score in this paper uses the first paragraph of the original news article as the

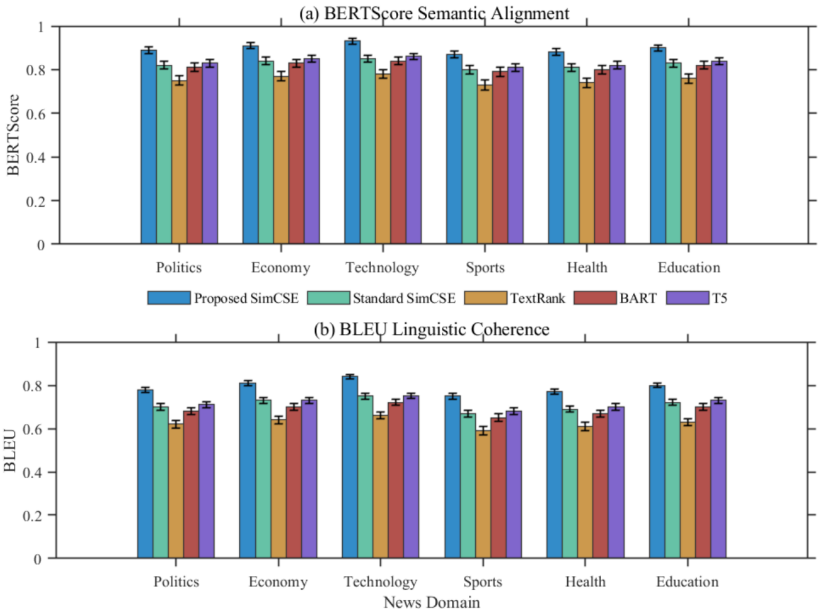

FIGURE 6. Semantic Fidelity and Language Coherence

reference text, rather than the full text summary. Since news headlines typically contain complete event elements and are highly condensed in language, they naturally overlap closely with human-generated titles, resulting in generally higher BLEU scores than general summary tasks. Nevertheless, the baseline model's BLEU score remains significantly lower than the proposed method, indicating that the performance advantage stems from the semantic-knowledge synergy mechanism's precise capture of core propositions, rather than loose evaluation criteria. Overall results confirm that the synergistic mechanism of semantic enhancement and knowledge infusion significantly improves the thematic consistency and factual credibility of generated titles.

KEY ENTITY RETENTION PERFORMANCE

To quantify the fidelity of generated headlines to the core facts of the original text, this paper defines two metrics: entity retention rate and entity accuracy, which respectively measure the completeness of key entity coverage and the correctness of reference. The system is evaluated on three types of news samples: short text ($\leq$100 words), medium text (200–500 words), and long text ($\geq$800 words). The proposed SimCSE model is compared with standard SimCSE, TextRank, BART, and T5.

Figure 7 presents the performance of various models in retaining key entities across three text lengths, measured by entity retention rate and entity accuracy. The proposed SimCSE model consistently outperforms baselines, especially in long texts, achieving a retention rate of 0.900 ± 0.016 and accuracy of 0.930 ± 0.014. This stems from its integration of semantic encoding with knowledge graphs, enabling precise identification and anchoring of recurring core entities during compression. In contrast, standard SimCSE lacks structured knowledge guidance and suffers from context dilution in long texts; TextRank relies on local salience and misses deeper semantic roles; BART and T5 often generalize or substitute entities, compromising factual fidelity. While other models degrade with increasing text length, our method maintains focus on key propositions by dynamically aggregating entity relationships via a graph attention network. Narrow error bars indicate high stability across samples, and the flat trend of the performance curve confirms robustness to textual complexity—highlighting the effectiveness of knowledge-augmented semantic modeling.

THEME CONSISTENCY ASSESSMENT

To assess the model's topic consistency within complex semantic structures, this paper applies two metrics: Topic-F1 and Topic Deviation. Topic-F1 measures the harmonized precision-recall ratio of shared topics be tween the generated headline and the original text. Topic Deviation is defined as the normalized KL (Kullback- Leibler) divergence; a lower value indicates closer distribution. The topic distribution of the original text and generated headlines is modeled using the Latent Dirichlet Allocation (LDA) topic model. This study employs a unified training LDA model (with $K \in \{2,3,4\}$ themes, consistent with experimental settings). After fitting the entire training news corpus, we perform topic inference on both the full text and generated headlines to obtain their respective topic distribution vectors. The topic F1 score is based on topics extracted by LDA, which is calculated by matching each topic in the title distribution with the most similar topic in the original text distribution one-to-one, and the harmonic mean of precision (number of matches/number of title topics) and recall (number of matches/number of original text topics) is calculated using

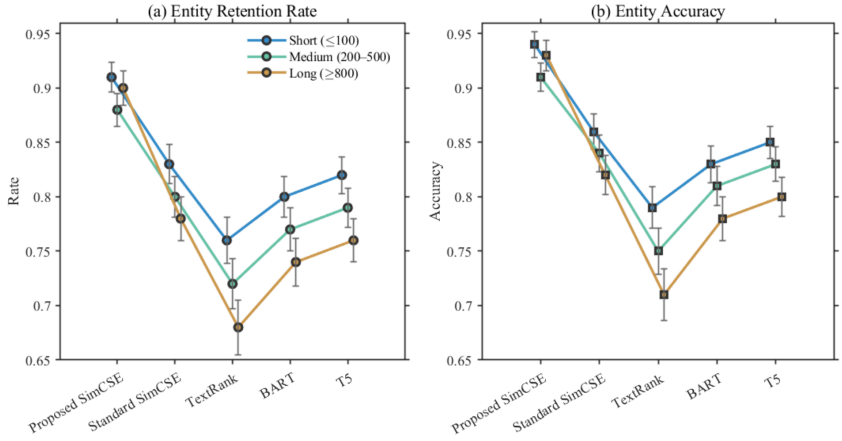

FIGURE 7. Key Entity Retention Performance

the number of matches. Topic deviation is defined as the KL divergence between the two distributions divided by log K, normalized to the [0,1] interval. Using complex news samples containing two, three, and four topics, it systematically analyzes each model's ability to capture and reconstruct multi-propositional structures. All data are presented as means ± standard deviations from five independent experiments.

Table 2 presents the quantitative performance of various models on multi-topic news samples, evaluated using Topic-F1 and Topic Deviation across texts containing two to four topics. The proposed SimCSE-based method consistently outperforms all baselines. Even under high complexity (four topics), it achieves a Topic-F1 of 0.835 ± 0.019 and a low Topic Deviation of 0.165 ± 0.021, with minimal variance—indicating strong stability. TextRank performs worst due to its reliance on surface-level word frequency and poor topic coverage. BART and T5, despite strong sequence modeling, often discard secondary topics during compression, increasing deviation. By fusing SimCSE semantic embeddings with knowledge graph structure, our approach enables fine-grained modeling of multi-propositional semantics, offering superior robustness and alignment accuracy in topic distribution—particularly well-suited for multi-event news summarization.

VERIFICATION OF FACTUAL ACCURACY

To systematically evaluate the factual credibility of generated headlines in multi-topic news, this paper employs a dual evaluation paradigm combining automatic verification and manual annotation. This paper adopts a fine-tuned Chinese version based on the FactCC architecture, trained on a self-built Chinese fact consistency dataset, and directly processes Chinese input. Based on the FactCC framework, the generated content is checked for contradictions, identifying assertions that are inconsistent with the semantics of the original text. The manually annotated Fact Error Rate (FER) is also applied to quantify the model's accuracy on key event elements. FER was completed independently by 3 annotators (each annotating 4,281 records), with an intergroup Kappa of 0.81; the Spearman correlation coefficient between FactCC and FER was 0.73. The proposed SimCSE model is compared with standard SimCSE, TextRank, BART, and T5 on a sample of complex news stories containing two, three, and four topics.

Table 3 presents factual accuracy results across models at varying topic complexity levels (2–4 topics), using both FactCC-detected contradiction rates and manually annotated factual error rates on composite news samples. The proposed SimCSE model consistently outperforms baselines, achieving the lowest error rates and maintaining stability as complexity increases. At four topics, it records a FactCC contradiction rate of 0.074 ± 0.013 and a factual error rate of 0.083 ± 0.015, demonstrating strong factual robustness. Its external knowledge graph validation effectively mitigates entity mismatches and relational misinferences compared to standard SimCSE, avoids semantic distortion inherent in TextRank's frequency-based approach, and reduces proposition distortion common in BART and T5 during compression. These results confirm that integrating semantic constraints with structured knowledge guidance significantly enhances factual consistency, offering a reliable pathway for high-fidelity headline generation.

V. CONCLUSION

To address semantic misalignment and topic drift in industry report headline generation, this paper proposes a framework integrating SimCSE semantic representations with Wikidata-based entity associations. It enhances semantic cohesion via contrastive learning and constructs a local entity subgraph to inject structured external knowledge. A graph attention network fuses semantic and entity embeddings, dynamically guiding the Transformer decoder toward key entities and core propositions. Experiments show consistent superiority over SimCSE, TextRank, BART, and T5 across news domains, achieving average BERTScore ≥ 0.87, BLEU ≥ 0.75, and under four-topic complexity, Topic-F1 = 0.835 ± 0.019 with topic deviation of 0.165 ± 0.021. The approach

TABLE 2. Thematic Consistency

Model	Topics	Topic-F1	Topic Deviation
Proposed SimCSE	2	0.892 ± 0.013	0.108 ± 0.015
	3	0.867 ± 0.016	0.133 ± 0.018
	4	0.835 ± 0.019	0.165 ± 0.021
Standard SimCSE	2	0.831 ± 0.017	0.169 ± 0.019
	3	0.784 ± 0.020	0.216 ± 0.023
	4	0.726 ± 0.024	0.274 ± 0.027
TextRank	2	0.763 ± 0.021	0.237 ± 0.025
	3	0.701 ± 0.025	0.299 ± 0.029
	4	0.642 ± 0.028	0.358 ± 0.032
BART	2	0.802 ± 0.018	0.198 ± 0.021
	3	0.743 ± 0.022	0.257 ± 0.025
	4	0.689 ± 0.026	0.311 ± 0.028
T5	2	0.815 ± 0.017	0.185 ± 0.020
	3	0.758 ± 0.021	0.242 ± 0.024
	4	0.703 ± 0.025	0.297 ± 0.027

TABLE 3. Factual Accuracy Assessment

Model	Topics	FactCC Contradiction Rate	Fact Error Rate
Proposed SimCSE	2	0.043 ± 0.008	0.051 ± 0.010
	3	0.058 ± 0.010	0.067 ± 0.012
	4	0.074 ± 0.013	0.083 ± 0.015
Standard SimCSE	2	0.089 ± 0.012	0.102 ± 0.014
	3	0.116 ± 0.015	0.131 ± 0.017
	4	0.148 ± 0.018	0.165 ± 0.020
TextRank	2	0.103 ± 0.014	0.124 ± 0.016
	3	0.137 ± 0.017	0.159 ± 0.019
	4	0.172 ± 0.020	0.194 ± 0.022
BART	2	0.095 ± 0.013	0.110 ± 0.015
	3	0.128 ± 0.016	0.143 ± 0.018
	4	0.161 ± 0.019	0.182 ± 0.021
T5	2	0.087 ± 0.011	0.105 ± 0.013
	3	0.121 ± 0.015	0.137 ± 0.017
	4	0.153 ± 0.018	0.176 ± 0.020

offers a credible, scalable solution for fact-sensitive, controllable headline generation.

AUTHOR CONTRIBUTIONS

Yun Liu designed, collected and analyzed the data, and drafted the manuscript. Yun Liu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Yun Liu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding .

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] K. Singh R, S. Khetarpaul, R. Gorantla, et al., “SHEG: summarization and headline generation of news articles using deep learning,” Neural Computing and Applications, vol. 33, no. 8, pp. 3251-3265, 2021, doi: 10.1007/s00521-020-05188-9.

[2] O. Turner S, L. Buttle, D. Linden S L V, and et al."More" Is More Interesting When Framing Headlines: The Influence of Comparative Language Framing and Consistency With Prior Beliefs on Engagement With Health-Related News, “Applied Cognitive Psychology, 2025, 39(5):e70118,”, doi: 10.1002/acp.70118.

[3] R. Das and D. Singh T, “Assamese news image caption generation using attention mechanism,” Multimedia Tools and Applications, vol. 81, no. 7, pp. 10051-10069, 2022, doi: 10.1007/s11042-022-12042-8.

[4] A. Table, “Generating headlines for Turkish news texts with transformer architecture based deep learning method,” Journal of the Faculty of Engineering and Architecture of Gazi University, vol. 39, no. 1, pp. 485-495, 2024.

[5] N. Hagar, N. Diakopoulos, and B. DeWilde, “Anticipating attention: On the predictability of news headline tests,” Digital Journalism, vol. 10, no. 4, pp. 647-668, 2022, doi: 10.1080/21670811.2021.1984266.

[6] D. Ismayanti, R. Said Y, N. Usman, et al., “The Students Ability in Translating Newspaper Headlines into English A Case Study,” IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, vol. 12, no. 1, pp. 108-131, 2024, doi: 10.24256/ideas.v12i1.4767.

[7] H. Gonçalo Oliveira, “Automatic generation of creative text in Portuguese: an overview,” Language Resources and Evaluation, vol. 58, no. 1, pp. 7-41, 2024, doi: 10.1007/s10579-023-09646-3.

[8] T. Shohan F, T. Nayeem M, S. Islam, et al., “XL-HeadTags: Leveraging Multimodal Retrieval Augmentation for the Multilingual Generation of News Headlines and Tags,” Findings of the Association for Computational Linguistics ACL 2024, pp. 12991-13024, 2024, doi: 10.18653/v1/2024.findings-acl.771.

[9] D. Vincentio A and S. Hansun, “A Fine-Tuned BART Pre-trained Language Model for the Indonesian Question-Answering Task,” Engineering, Technology & Applied Science Research, vol. 15, no. 2, pp. 21398-21403, 2025, doi: 10.48084/etasr.9828.

[10] R. Goyal, P. Kumar, and P. Singh V, "Automated question and answer generation from texts using text-to-text transformers," Arabian Journal for Science and Engineering, vol. 49, no. 3, pp. 3027-3041, 2024, doi: 10.1007/ s13369-023-07840-7.
[11] N. Mulla and P. Gharpure, "Leveraging well-formedness and cognitive level classifiers for automatic question generation on Java technical passages using T5 transformer," International Journal of Information Technology, vol. 15, no. 4, pp. 1961-1973, 2023, doi: 10.1007/s41870-023-01262-2.
[12] P. Mishra, C. Diwan, S. Srinivasa, et al., "Automatic title generation for learning resources and pathways with pretrained transformer models," International Journal of Semantic Computing, vol. 15, no. 04, pp. 487-510, 2021, doi: 10.1142/ S1793351X21400134.
[13] B. Juarto and S. Girsang A, "Neural collaborative with sentence BERT for news recommender system," JOIV: International Journal on Informatics Visualization, vol. 5, no. 4, pp. 448-455, 2021, doi: 10.30630/joiv.5.4.678.
[14] J. Pijeira-Díaz H, S. Subramanya, J. van de Pol, et al., "Evaluating Sentence-BERT-powered learning analytics for auto-mated assessment of students' causal diagrams," Journal of Computer Assisted Learning, vol. 40, no. 6, pp. 2667-2680, 2024, doi: 10.1111/jcal.12992.
[15] R. Wang, F. Hou, F. Cahan S, et al., "Fine-grained entity typing with a type taxonomy: a systematic review," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 5, pp. 4794-4812, 2022, doi: 10.1109/TKDE.2022.3148980.
[16] S. Tang, Q. Yang, L. Fan, et al., "Contrastive learning-based semantic communications," IEEE Transactions on Communications, vol. 72, no. 10, pp. 6328-6343, 2024, doi: 10.1109/TCOMM.2024.3400912.
[17] T. Xiao, S. Liu, S. De Mello, et al., "Learning contrastive representation for semantic correspondence," International Journal of Computer Vision, vol. 130, no. 5, pp. 1293-1309, 2022, doi: 10.1007/s11263-022-01602-y.
[18] W. He, C. Wu, and S. Zhou, "Study on short text clustering with unsupervised SimCSE," Computer Science, vol. 50, no. 11, pp. 71-76, 2023.
[19] Y. Ye and S. Ji, "Sparse graph attention networks," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 905-916, 2021, doi: 10.1109/TKDE.2021.3072345.
[20] Y. Liu, S. Yang, Y. Xu, et al., "Contextualized graph attention network for recommendation with item knowledge graph," IEEE Transactions on knowledge and data engineering, vol. 35, no. 1, pp. 181-195, 2021, doi: 10.1109/ TKDE.2021.3082948.
[21] S. Li and Y. Zhang, "Improving entity linking by combining semantic entity embeddings and cross-attention encoder," Journal of Intelligent & Fuzzy Systems, vol. 46, no. 1, pp. 2899-2910, 2024, doi: 10.3233/JIFS-233124.