EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Optimization of Deep Learning-Based Patent Infringement Text Comparison and Retrieval Algorithms

Juan Zou[1,*]

[1]International Business School, Guangdong University of Finance & Economics, Guangzhou 510320, China

Abstract

INTRODUCTION: Patent texts may present challenges in accurately identifying potential patent infringement risks due to subtle differences in technical features. OBJECTIVES: Deep learning effectively models the complex semantic structures and local feature correlations within patent texts by integrating deep semantic representations and deep matching mechanisms, thereby enhancing sensitivity to minute technical variations and improving decision support for infringement risk assessment. OBJECTIVES: this study explores deep learning-based optimization methods for patent infringement risk text comparison and retrieval algorithms. A macro-level patent text comparison and retrieval layer is constructed by integrating Word2vec and LDA topic models. This layer models patent text semantics, generates word-level semantic vectors for patents, and enables rapid comparison and retrieval of candidate patent sets related to the target patent at the technical feature semantic level. A COV-BiGRU twin neural network is introduced as the fine-grained comparison and retrieval layer. This layer optimizes the macro-level retrieval algorithm by performing granular semantic matching and Manhattan distance calculations between candidate patent texts and the target patent text. Based on these results, patent infringement risk retrieval are retrieved. RESULTS: Results demonstrate that this method reduces the candidate patent text set size to 0.457% of the full database during macro-level retrieval. CONCLUSION: In the fine-grained retrieval optimization phase, it achieves 100% recall of known high-risk patents while effectively excluding non-infringing patents that share similar technical themes but differ in specific technical features. This validates the preliminary effectiveness of the optimized dual-layer retrieval algorithm for texts with fine-grained differences on the tested patent corpus. This validates the preliminary effectiveness of the optimized dual-layer retrieval algorithm for texts with fine-grained differences on the tested patent corpus.

Keywords: deep learning, patent infringement, text comparison, retrieval algorithm optimization, LDA topic model, COV-BiGRU twin neural network

Received on 25 March 2026, accepted on 02 June 2026, published on 30 June 2026

doi: 10.4108/12346

1. Introduction

Under the current intellectual property protection framework, patent infringement determination has become a critical component in technological innovation and market competition [1]. As global technological competition intensifies and patent application volumes continue to rise [2], the complexity, technicality, and diversity of patent texts have increased significantly, posing severe challenges to identifying and determining infringement risks [3]. Traditional patent comparison and retrieval methods primarily rely on manual interpretation or superficial text matching [4], struggling to meet the demands of analyzing massive, heterogeneous, and highly semantically specialized patent literature. This often leads to the omission or misjudgment of critical infringement clues [5], not only impacting corporate technology strategy and innovation security but also constraining the efficiency and accuracy of judicial and administrative rulings [6].

*Corresponding author. Email: zjpapermail@163.com

Therefore, establishing an efficient, precise, and semantically adaptive text comparison and retrieval mechanism for patent infringement holds urgent practical significance for safeguarding the legitimate rights of innovators, promoting rational technology flow, and enhancing intellectual property governance effectiveness [7].

Currently, some scholars have initiated relevant research in this area. For instance, Kumar, R et al. [8] investigated a patent infringement text information retrieval method integrating ontology semantic models and hybrid optimization. By creating ontology structures to enhance semantic associations within patent texts and combining hybrid optimization algorithms based on these semantic links, they achieved rapid infringement text retrieval. However, this approach relies on manually constructed ontologies, which cannot adapt to the complex fluctuations in technical semantic vectors of patent texts. Consequently, it only enables basic coarse-grained retrieval of patent texts and exhibits low accuracy in identifying infringement texts with subtle differences in technical features. The patent infringement text retrieval method based on hybrid allocation models studied by Bouguila [9] developed a patent infringement text retrieval method based on a hybrid allocation model. This approach primarily employs the hybrid allocation model to learn latent themes within patent texts, enabling coarse-grained matching retrieval of infringing texts based on these learned themes. However, this method relies solely on latent thematic information from patent texts and does not incorporate fine-grained semantic information such as text word vectors. Consequently, it struggles to accurately distinguish between patent texts with subtle differences in technical features, leading to reduced retrieval matching accuracy for such infringing texts.

Word2vec, a neural network-based word embedding method, transforms words in text into low-dimensional, dense vector representations. By predicting context words or central words, this method learns distributed word representations that effectively capture deep linguistic patterns such as semantic similarity and analogical relationships between words. This enhances the quality and efficiency of word-level semantic understanding in natural language processing tasks [10]. As a classic unsupervised generative probabilistic model, the LDA topic model discovers latent semantic topic structures within large-scale document collections [11]. Its primary advantages lie in strong interpretability and scalability, enabling automatic extraction of thematic information from massive texts without relying on manual annotation. It is widely applied in text classification, information retrieval, and knowledge discovery [12]. The COV-BiGRU twin neural network serves as a specific implementation architecture of deep learning technology in semantic matching. By deeply integrating the local feature extraction capabilities of convolutional neural networks (CNNs) with the sequence modeling capabilities of bidirectional gated recurrent units (BiGRU), it achieves an end-to-end deep learning model. This architecture is commonly employed for calculating semantic similarity between text pairs, demonstrating deep learning's core strengths in automatic feature learning and complex pattern recognition [13]. The advantages of utilizing this network for algorithm optimization lie in its ability to capture local semantic fragments through convolutional layers, enhancing perception of fine-grained micro-features [14]; its BiGRU structure simultaneously models contextual dependencies, improving understanding of overall logical and structural relationships within proposals; and it supports fully supervised learning from source text to target task determination, achieving synergistic enhancement of retrieval accuracy and semantic discrimination capabilities [15].

Based on the above analysis, this paper investigates an optimization method for patent infringement risk text comparison and retrieval algorithms using COV-BiGRU twin neural network deep learning. This method deeply integrates Word2vec, LDA topic modeling, and COV-BiGRU deep matching networks. This establishes a dual-layer patent infringement text comparison and retrieval algorithm optimization framework that combines semantic generalization capability, interpretable thematic analysis, and supervised fine-grained discrimination. It enables both macro-level coarse screening and fine-grained risk indication for potentially infringing texts for different target patents, providing decision support rather than definitive legal determination for patent texts with subtle technical feature differences. It is acknowledged that the individual components of the proposed framework—Word2vec, LDA, and Siamese-style networks—are established techniques in general natural language processing. However, the core contribution of this paper lies not in proposing entirely new foundational modules, but in their systematic integration and task-specific optimization for the patent infringement risk retrieval scenario. This work presents, for the first time, a dual-layer architecture that jointly optimizes macro-level semantic recall through a novel fusion of Word2vec and LDA with domain-adaptive coefficients, and fine-grained difference discrimination through a COV-BiGRU twin network with multi-head self-attention. This integrated engineering solution, tailored to the unique challenges of subtle technical feature variations in patent texts, does not appear in existing literature. While more advanced contemporary semantic models such as BERT or PatentBERT could potentially replace the Word2vec+LDA module, they introduce higher computational costs and require large-scale domain-specific pre-training; the proposed framework deliberately balances efficiency, interpretability, and accuracy for practical patent retrieval scenarios. Enhancing the macro-level retrieval with transformer-based embeddings is a promising direction for future work.

The proposed framework deliberately balances efficiency, interpretability, and accuracy for practical patent retrieval scenarios. The two-stage design ensures that the computationally expensive COV-BiGRU model only processes a small candidate set (compressed to <0.5% of the full database), making real-time or near-real-time

deployment feasible on standard server hardware. Enhancing the macro-level retrieval with transformer-based embeddings is a promising direction for future work.

2. Optimization of Patent Infringement Text Comparison and Retrieval Algorithms

2.1. Overall Technical Framework for Patent Infringement Text Comparison and Retrieval Algorithm Optimization

To achieve rapid and precise comparison and retrieval of patent infringement texts, and to meticulously screen out patent texts exhibiting infringement similarity with the target patent text, this paper employs a deep learning-optimized dual-layer comparison retrieval architecture: First, Word2vec and LDA are employed for rapid semantic generalization and coarse screening of patent texts. Subsequently, a deep learning matching model—the COV-BiGRU-based twin neural network—performs refined semantic comparison and retrieval optimization on the coarse-screened candidate patent text set, enabling precise patent infringement risk assessment. This deep learning-optimized double-layer comparison and retrieval framework for patent infringement texts is illustrated in Figure 1.

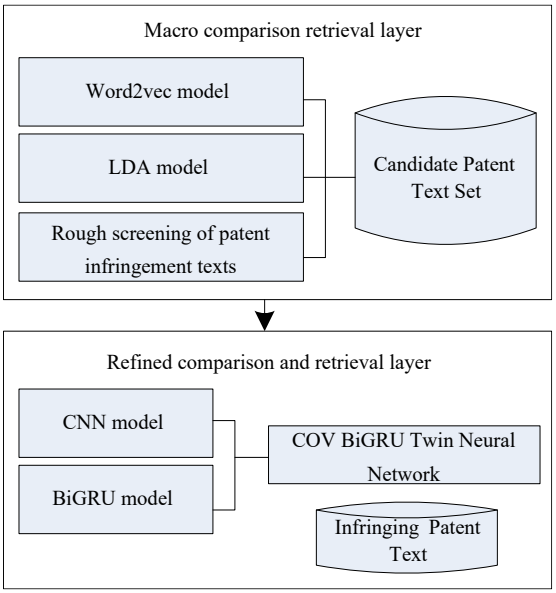

Figure 1. Double-layer comparison and retrieval framework for patent infringement text optimized by deep learning

The overall solution employs a retrieval algorithm integrating Word2vec and LDA to achieve efficient macro-level semantic recall [16], while utilizing COV-BiGRU twin neural networks for high-precision fine-grained difference discrimination. By standardizing parameters across the entire process—including word vector dimensions, text preprocessing specifications, and sequence lengths [17]—seamless integration between the two-stage models is achieved. This ultimately forms an integrated, high-efficiency, and high-precision patent infringement text comparison and retrieval algorithm optimization system characterized by "macro-semantic retrieval —refined comparison—optimized discrimination."

2.2. Macro-Level Comparison and Retrieval Algorithm for Patent Infringement Texts

Building upon established techniques but with novel integration, a macro-comparison retrieval algorithm integrating Word2vec and LDA topic models is employed to perform semantic modeling of patent texts. Unlike conventional separate usage, this work proposes a weighted fusion strategy (Eq. 7-9) that combines word-level dense vectors from Word2vec with topic-level global structures from LDA, with domain-adaptive coefficients to adjust for different technical fields.

A macro-comparison retrieval algorithm integrating Word2vec and LDA topic models is employed to perform semantic modeling of patent texts. This generates patent semantic vectors at the “word granularity” level and constructs a vectorized index database for the entire patent corpus. This enables rapid macro-comparison retrieval of candidate patent sets related to the target patent at the technical feature semantic level. The framework of this macro-comparison retrieval algorithm is shown in Figure 2.

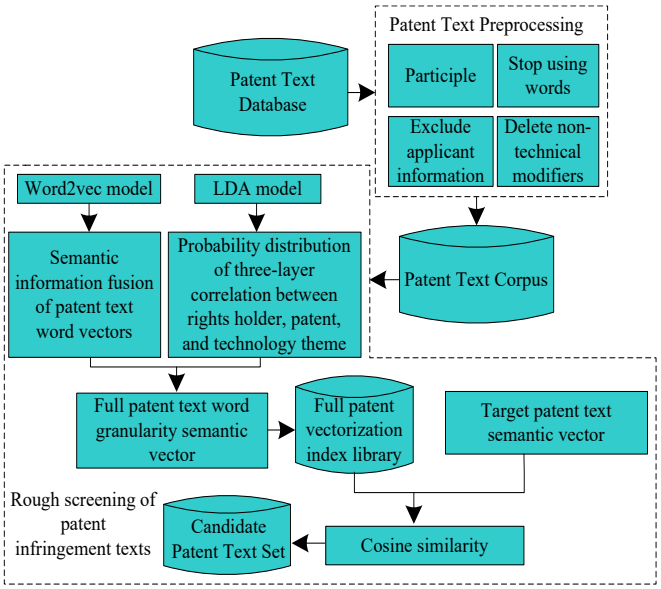

Figure 2. Macro comparison and retrieval algorithm framework for patent infringement texts

The Word2vec model is an optimized word vector training model based on Log-Bilinear and NNLM architectures. It maps high-dimensional sparse patent text vocabulary to a low-dimensional dense vector space, preserving contextual semantics and positional relationships among technical terms. This effectively addresses the dimensionality explosion and semantic loss issues inherent in traditional one-hot encoding [18]. The Word2vec model encompasses two typical architectures: CBOW (Continuous Bag-of-Words Model) and Skip-gram [19]. This paper adopts the CBOW architecture to train and learn specialized vocabulary vectors for patent texts, addressing the training efficiency requirements of large-scale patent corpora [20]. The CBOW model consists of an input layer, a mapping layer, and an output layer. Inputs are context word vectors within a window range of patent text. By predicting the central target word using contextual information, iterative optimization of patent text word vectors is achieved [21]. The CBOW model defines the word features of patent text as:

$$E^{l}(\bar{l}) = \begin{cases} 1 & \bar{l} = l \\ 0 & \bar{l} \neq l \end{cases}. \tag{1}$$

Where, $\bar{l}$ represents the candidate target word traversing all words in the patent text, and l denotes the current central target word under training. When both are identical, the patent text word feature $E^{l}(\bar{l})$ is set to 1; otherwise, it is 0.

The objective function for iterative optimization of patent text word vectors in the CBOW model is expressed as:

$$B(l) = \prod_{a \in \mathrm{Neg}(l) \cup \{l\}} P(a \mid \mathrm{context}(l)) \tag{2}$$

Where, a represents the positive and negative sample words in the CBOW model's objective function; $P(a \mid \mathrm{context}(l))$ denotes the conditional probability of the target word a appearing given the patent text context $\mathrm{context}(l)$; and $\mathrm{Neg}(l)$ is the set of negative sampled words for the central target word l. The calculation of the conditional probability $P(a \mid \mathrm{context}(l))$ must satisfy the following condition:

$$P(a \mid \mathrm{context}(l)) = \begin{cases} \delta(Z_l^T \alpha^a) & E^{l}(a) = 1 \\ 1 - \delta(Z_l^T \alpha^a) & E^{l}(a) = 0 \end{cases} \tag{3}$$

Where, δ denotes the Sigmoid activation function; Z_l^T represents the concatenation of patent text context vectors; and α^a denotes the model parameters corresponding to the target word a.

Let the preprocessed patent text corpus after tokenization, stopword removal, applicant information exclusion, and deletion of non-technical modifiers be denoted as $C = \{c_1, c_2, \ldots, c_M\}$. Here, M represents the number of patent texts in the preprocessed corpus, and c_i denotes a single patent composed of its claims, specification, and abstract. The corresponding patent technical vocabulary list for this preprocessed patent text corpus C is $W = \{w_1, w_2, \ldots, w_N\}$, where N denotes the total number of patent technical terms in this vocabulary list. Using this patent text corpus C as the training data, the Word2vec model training was completed using the aforementioned CBOW structure [22-24]. The patent text context window size is set to d to constrain the semantic association range of technical terms and accommodate the sentence structure featuring consecutive technical feature descriptions in patent texts. The patent text word vector dimension is uniformly set to g_{emb}, ensuring comprehensive semantic representation of specialized terminology while balancing large-scale vector storage and nearest-neighbor retrieval efficiency. The final output is the semantic vectors $u(w_j) \in \mathbb{R}^{g_{\mathrm{emb}}}$ corresponding to each vocabulary term in the patent text. This maps the high-dimensional sparse vocabulary of patent texts to a low-dimensional dense vector space, preserving the contextual semantic and positional relationships between technical terms. It effectively addresses the dimension explosion and semantic loss issues caused by traditional one-hot encoding, providing a unified vector foundation for subsequent LDA topic modeling to achieve word-level semantic fusion of patent texts [25,26].

Building upon this foundation, the LDA topic model is employed to uncover the latent distribution of technical themes within the preprocessed patent text corpus [27]. The total number of patent themes is set to K, determined from the hierarchical classification of technical fields mapped to IPC main group level patent codes, which typically yields 50–80 categories for a mixed corpus. This value was fixed at 50 after a preliminary grid search to balance topic coherence against computational cost. This ensures each theme corresponds to a distinct category of technical solutions, aligning with the comparison principle of matching identical or similar technical fields in patent infringement determinations [28]. Through model inference and solution, the patent-topic probability distribution is obtained as follows:

$$Z'_{ij} = P\left(f_j \mid c_i\right) \tag{4}$$

Where, f_j denotes the j th patent technical theme; $P\left(f_j \mid c_i\right)$ represents the probability that the i th patent c_i belongs to the technical theme f_j. Concurrently, this module yields the patent theme-term probability distribution:

$$\beta_{jk} = P(w_k \mid f_j) \tag{5}$$

Where, w_k denotes the k th patent technical term; $P(w_k \mid f_j)$ represents the probability of patent technical term w_k belonging to technical topic f_j. Based on this, a three-layer association probability distribution linking rights holders, patents, and technical topics is constructed,

enabling the characterization of the relationship between patent text technical features and rights holders.

To integrate the fine-grained semantic information of Word2vec patent text word vectors with the global structural features of LDA patent topic distributions, the top k_f core terms with the strongest representational capacity are extracted for each patent technical topic. The term membership probabilities are normalized to obtain corresponding patent topic term weights:

$$\omega_{j\dot{k}} = \frac{\beta_{j\dot{k}}}{\sum_{\dot{n}=1}^{k_f} \beta_{j\dot{n}}} \tag{6}$$

Weight and sum the Word2vec word vectors of each core term in the patent text according to the above weights to generate the semantic vector $u(f_j) \in \mathbb{R}^{g_{\text{emb}}}$ for a single technical theme:

$$u(f_j) = \sum_{k=1}^{k_f} \omega_{j\dot{k}} \cdot u(w_{\dot{k}}) \tag{7}$$

Where, $u(w_{\dot{k}})$ is the Word2vec vector for the $\dot{k}$ th core term in the patent text. For each patent, select the top k_c core technical themes with the highest probabilities, normalize the patent-theme probabilities, and obtain the corresponding theme weights for that patent:

$$\omega_{ij} = \frac{Z'_{ij}}{\sum_{\dot{n}=1}^{k_c} Z'_{i\dot{n}}} \tag{8}$$

We fuse the corresponding patent technical topic vectors using these weights, ultimately generating the word-level patent semantic vector $u(c_i) \in \mathbb{R}^{g_{\text{emb}}}$ for patent infringement comparison:

$$u(c_i) = \sum_{j=1}^{k_c} \omega_{ij} \cdot u(f_j) \tag{9}$$

The fusion weights can be dynamically adjusted via domain-adaptive coefficients to accommodate variations in patent text expression across different technical fields such as mechanical, electrical, and chemical engineering, thereby enhancing the semantic vector's ability to distinguish core technical features [29,30].

The word-level semantic vectors $\{u(c_1), u(c_2), \ldots, u(c_M)\}$ corresponding to the full patent text are uniquely bound to the patent publication number, IPC classification number, ownership information, and legal status. This constructs a full-text vectorized index database for infringement retrieval, utilizing an approximate nearest neighbor search structure to achieve rapid vector matching. Combining the equivalence determination for patent infringement with the literal infringement determination criteria, a cosine similarity threshold χ is set to 0.72is adopted. This value was selected via a pilot study on a validation set (10% of the corpus), where thresholds from 0.65 to 0.80 were evaluated; 0.72 yielded the best trade-off between macro-level recall (≥95%) and candidate set compression ratio. The cosine similarity is calculated between the semantic vector $u(c_{\text{target}})$ of the target patent and the patent vectors in the index database:

$$\cos(u(c_{\text{target}}), u(c_i)) = \frac{u(c_{\text{target}}) \cdot u(c_i)}{\| u(c_{\text{target}})\| \cdot \| u(c_i)\|} \tag{10}$$

Screen the patent texts that satisfying $\cos(u(c_{\text{target}}), u(c_i)) \geq \chi$ to form the candidate patent set C_{cand}. This completes a coarse screening of macro-level patent infringement texts at the technical feature semantic level, reducing the number of invalid samples while ensuring high recall for relevant patents. This provides a high-quality candidate set for subsequent fine-grained semantic matching and precise patent infringement determination.

Specifically, the domain-adaptive coefficients are implemented as a learnable scaling vector $\alpha \in \mathbb{R}^K$ applied to the fused topic weights in Eq. 9. For a target patent belonging to IPC sections$\in$\{A,B,…,H\}, the coefficient vector is initialized $\alpha_{\text{init}}(s) = \text{softmax}(\beta \text{TF-IDF}_s)$, where β is a temperature parameter set to 0.5. During supervised fine-tuning of the COV-BiGRU network, these coefficients are updated via backpropagation alongside other network parameters. This allows the model to emphasize mechanical structural terms for Section F (Mechanical Engineering) or chemical composition terms for Section C (Chemistry), adapting the semantic representation to the linguistic characteristics of each technical field.

2.3. Fine-Grained Optimization of Patent Infringement Text Comparison Retrieval Algorithm

To further enhance the accuracy of patent infringement text determination, a COV-BiGRU-based twin neural network is introduced upon the macro-level comparison retrieval algorithm. This network performs fine-grained semantic matching and similarity calculation on the macro-level retrieved candidate patent text pairs. Through supervised fine-tuning, the retrieval algorithm is optimized to improve its perception of subtle differences in technical features within patent texts. This forms a complete algorithmic optimization framework: "macro-level semantic retrieval—refined comparison—optimized discrimination." This twin neural network adopts a dual-branch weight-sharing architecture. Each branch inputs either the target patent text c_{tar} or candidate patent text $c_{\text{cand},i} \in C_{\text{cand}}$, both utilizing the identical COV-BiGRU feature extraction structure. This ensures consistent feature encoding spaces for text pairs, facilitating fine-grained difference comparisons. All patent text inputs undergo unified macro-level preprocessing and utilize Word2vec pre-trained vectors $u(w_j)$. Input sequences maintain a

fixed length of L', with each patent text input matrix dimension set to $L' \times g_{\text{emb}}$. This strictly aligns with the macro-level modeling word vector dimension g_{emb}, eliminating cross-module vector mismatch issues. The fine-grained patent infringement text comparison and retrieval model structure of the COV-BiGRU twin neural network is illustrated in Figure 3.

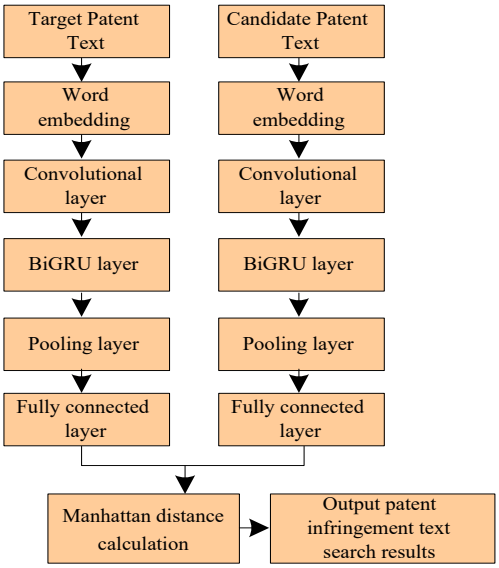

Figure 3. Fine-grained patent infringement text comparison and retrieval model of COV BiGRU twin neural network

First, one-dimensional convolutional layers extract local technical features from the input patent text c_{tar} and candidate patent text $c_{\text{cand},i} \in C_{\text{cand}}$. The convolution kernel size is κ, with ϕ_{conv} kernels and a stride of 1. These kernels scan the local receptive fields of the input patent text sequence matrices to capture continuous local features such as technical terms, structural connections, and procedural steps. The resulting convolutional output feature length is:

$$L'_{\text{conv}} = L' - \kappa + 1 \tag{11}$$

Following the convolutional layer, a bidirectional gated recurrent unit (BiGRU) is employed. The forward GRU encodes the forward temporal dependencies of the patent text sequence, while the backward GRU encodes the backward temporal dependencies. The global contextual features of the patent text are obtained by concatenating the hidden states from both directions. The hidden state at a single time step is:

$$q_t = \varphi(q_{t-1}, x_t) \tag{12}$$

Where, x_t is the input to the BiGRU, i.e., the continuous local features extracted by the convolutional layer from the target patent text c_{tar} and candidate patent text $c_{\text{cand},i} \in C_{\text{cand}}$; φ is the state update function of the GRU unit; q_{t-1} is the hidden state at the previous time step. The final temporal feature of the patent text obtained after bidirectional concatenation is:

$$Q = \text{Concat}(\vec{Q}, \overleftarrow{Q}) \tag{13}$$

Where, $\vec{Q}$ and $\overleftarrow{Q}$ represent the forward and backward temporal dependency features of the patent text, respectively. The hidden layer dimension of the BiGRU is uniformly set to g_{hidden} to ensure compatibility with subsequent fully connected layers. To enhance the weighting of core technical features, a multi-head self-attention mechanism is introduced to weight the patent text temporal features Q output by the BiGRU. The attention calculation formula is:

$$\begin{cases} \text{Attention}(Q', K', V') = \text{softmax}\left(\dfrac{Q'K'^{\text{T}}}{\sqrt{g_i}}\right)V' \\ \text{MultiHead}(Q', K', V') = \text{Concat}(\text{head}_1, \ldots, \text{head}_{M'})\omega^O \end{cases} \tag{14}$$

Where, Q', K', V' represent the query, key, and value vectors in the attention mechanism, respectively, all derived from linear transformations of BiGRU features; ω^O denotes the trainable output projection weight; g_i is the dimension of the single-head attention feature; head_i represents a single attention head; and M' is the number of attention heads,with g_{hidden} satisfies dimensional divisibility constraints, enabling the model to focus on core technical features and distinguishing technical features in claims while mitigating interference from irrelevant descriptions.

The attention-enhanced temporal features of the patent text $\hat{Q}$ undergo max pooling for dimensionality reduction. Following a fully connected layer mapping, this yields a fixed-dimensional overall semantic feature vector $\boldsymbol{s}_{\text{all}}$ for the patent text. The twin branches of the network respectively output the target patent text features $\boldsymbol{s}_{\text{tar}}$ and candidate patent text features $\boldsymbol{s}_{\text{cand}}$. The fine-grained similarity between patent text pairs is computed using Manhattan distance:

$$\iota_{\text{man}} = \sum_{k'=1}^{g_{\text{fc}}} |\boldsymbol{s}_{\text{tar},k'} - \boldsymbol{s}_{\text{cand},k'}| \tag{15}$$

Where, g_{fc} denotes the feature dimension of the fully connected output patent text. This similarity value ι_{man} precisely detects critical infringement determination details such as minor replacements of technical features, numerical differences, and changes in relational connections. A lower value indicates greater semantic similarity in technical features between the patent text pairs.Using manually annotated infringing/non-infringing patent text pairs as supervised samples, we constructed a

cross-entropy loss function to perform supervised fine-tuning on the COV-BiGRU twin network. A Manhattan distance threshold ι_{λ} was set using the distance distribution from a held-out validation set of 500 labeled patent pairs, half infringing and half non-infringing. This value maximizes the F1-score at the fine-grained stage, keeping the two classes well separated while remaining robust. When the similarity value $\iota_{\text{man}} \leq \iota_{\lambda}$, the candidate patent text pair was judged to infringe the target patent; conversely, when $\iota_{\text{man}} > \iota_{\lambda}$, it was judged to be non-infringing. This approach achieves end-to-end joint optimization of coarse-grained screening and fine-grained judgment in patent infringement text comparison retrieval. It enhances the ability to precisely detect subtle infringement features in patent texts while maintaining retrieval efficiency.

3. Experimental Results Analysis

This paper constructs an experimental environment closely resembling real-world patent infringement analysis scenarios. The experimental subjects comprise two parts: the first consists of 10 representative target patent texts serving as query inputs for infringement searches. These texts cover all eight major IPC classes and include known infringement precedents; The second part comprises a mixed patent text dataset of 200,035 entries, serving as the comparison search repository. This repository consists of 200,000 full-text patents and 35 known infringing patents explicitly linked to the 10 target patents through manual annotation. The current dataset is limited: 10 target patents, a mixed corpus of 200,035 patent texts, and only 35 known infringing samples. Results should be treated as an initial proof-of-concept rather than evidence of generalizable performance across all technical domains. The infringement relationships were annotated by domain experts following patent examination guidelines or sourced from publicly available patent reexamination decisions. The experiment simulates the process of precisely retrieving patent texts infringing on target patents from a massive patent repository using the proposed method, validating its practical comparison retrieval performance. To protect data privacy and focus on method validation, this experiment selected 10 patents covering all 8 IPC classes as target query cases, as shown in Table 1. The table summarizes the core technical themes of each patent and constructs 3–5 known infringing patents per target patent as evaluation benchmarks (totaling 35). These infringement relationships were annotated by domain experts following patent examination guidelines or sourced from publicly available patent reexamination decisions.

Table 1. Distribution of experimental target patents

Target Patent ID	IPC Division	Technical field	Representative Technical Themes/Patent Content Description	Known number of infringing patents
Q1	A	Essential for human life	Intelligent control method and device for household appliances	4
Q2	B	Homework and transportation	Multi joint handling robot system for industrial production lines	3
Q3	C	Chemistry and Metallurgy	Environmentally friendly polymer anti-corrosion coating composition and preparation process	4
Q4	D	Textile and papermaking	Preparation methods and products of functional fiber fabrics	3
Q5	E	Fixed structure	Prefabricated building component connection structure	3
Q6	F	Mechanical Engineering	Key components of internal combustion engine turbocharging system	4
Q7	G	Physics	MEMS Sensor Design and Packaging Solution	3
Q8	H	Electricity	Mobile Communication Channel Feedback Algorithm	5
Q9	H	Electricity	Equipment for preparing electrode materials for lithium-ion batteries	3
Q10	A	Essential for human life	Formula and preparation method of medical dressings	3
Total	Covering 8 IPC classes	—	—	35 articles

Key parameter settings for the method proposed in this article are shown in Table 2.

Table 2. Key experimental parameter settings of the method proposed in this article

Model/Module	Parameter name	Set value
Word2vec+LDA	Word vector dimension	200
	Context window size	8
	Number of themes	50
	Cosine similarity threshold	0.72
COV-BiGRU	Input sequence length	256
	Convolutional kernel size/quantity	3 / 128
	BiGRU hidden layer dimension	256
	Manhattan distance threshold	3.8

Parameter sensitivity analysis: To assess stability, we varied the cosine similarity threshold from 0.68 to 0.76 and the Manhattan distance threshold from 3.5 to 4.1 on a validation subset. Macro-level recall stayed above 93% across all thresholds, and the fine-grained F1-score varied by less than 2.5%. The candidate set compression ratio ranged from 0.41% to 0.52%, showing the framework is reasonably robust within these intervals. Default thresholds of 0.72 and 3.8 were chosen as operating points, as they yielded the highest average F1-score across the 10 target patents.

To validate the effectiveness of the proposed W-LDA algorithm—which integrates Word2vec and LDA for macro-level patent text comparison retrieval—in preliminary patent infringement searches, a comparative experiment was designed against the traditional single LDA topic modeling approach. The experiment utilized 10 selected target patents and a full dataset of 200,035 mixed patent documents. Both methods performed comparison searches. Eight target patents (Q1–Q8) were randomly selected from the comparison results. The most highly ranked irrelevant patent texts from each method—representing the most glaring misclassifications—were presented alongside their calculated topic similarity scores. qualitatively analyzing the differences between the two comparison search results. The obtained results are shown in Figure 4 and Table 3.

Table 3. Comparison of search results with the highest error ranking for target patents

Target Patent ID	Core theme description (summary)	Method	Top-1 Error Cases (Most Significant Search Errors)
Q1	Intelligent control method and device for household appliances	W-LDA	Industrial motor speed control system
		LDA	Medical Device Control System
Q2	Multi joint handling robot for industrial production line	W-LDA	Warehouse logistics sorting robot
		LDA	Engineering robotic arm
Q3	Environmentally friendly polymer anti-corrosion coating composition	W-LDA	Metal surface treatment agent
		LDA	Building waterproof coating
Q4	Preparation method of functional fiber fabric	W-LDA	Non woven fabric production process
		LDA	Textile printing and dyeing process
Q5	Prefabricated building component connection structure	W-LDA	Bridge support structure
		LDA	Steel structure connectors
Q6	Key components of internal combustion engine turbocharging system	W-LDA	Air compressor impeller
		LDA	Pump body sealing device
Q7	MEMS Sensor Design and Packaging	W-LDA	Semiconductor pressure sensor
		LDA	Optical sensor packaging
Q8	Mobile Communication Channel Feedback Algorithm	W-LDA	Image compression encoding algorithm
		LDA	Circuit noise filtering algorithm

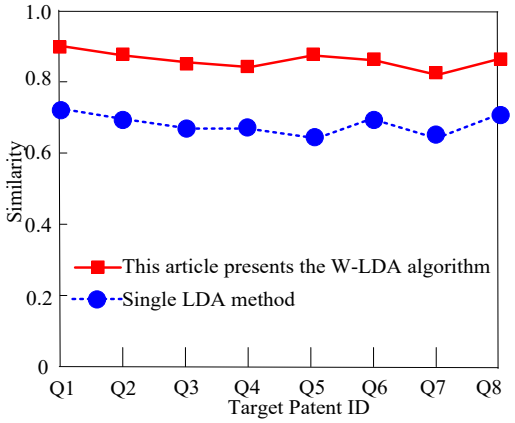

Figure 4. Comparison of topic similarity of target patent error retrieval results

Figure 4 and Table 3 demonstrate that the topic similarity scores of the highest-ranking erroneous results for each target patent retrieved by the W-LDA algorithm in this paper are significantly lower than those obtained by the traditional single LDA method. Taking target patent Q1 as an example, the proposed algorithm captures the semantic differences of contextual words such as “domestic” and “medical” through word vectors, significantly reducing the similarity scores of such erroneous patent texts. This relegates them to lower rankings, thereby enhancing the purity of the retrieval results. In contrast, the traditional single LDA method, by ignoring contextual semantics between words in patent texts and relying solely on global word frequency co-occurrence, mistakenly judges patents with similar technical fields but vastly different core technical features as highly similar, with similarity scores consistently above 0.8. For instance, this method confuses “home appliance control” with “medical device control” due to the high number of co-occurring words under the “control” theme. Thus, in the macro-comparison retrieval scenario for patent infringement texts, the proposed W-LDA algorithm successfully avoids the issue of patent text topic generalization caused by semantic deficiencies in traditional single LDA methods. By precisely capturing patent technical features through word-level semantic vectors, it ensures the accuracy of patent infringement text comparison retrieval results, validating the rationality of the macro-comparison retrieval layer design in this method.

After coarse screening by the macro comparison and retrieval layer of this method, the average similarity values of the candidate sets for the 10 experimental target patent infringement texts are shown in Figure 5.

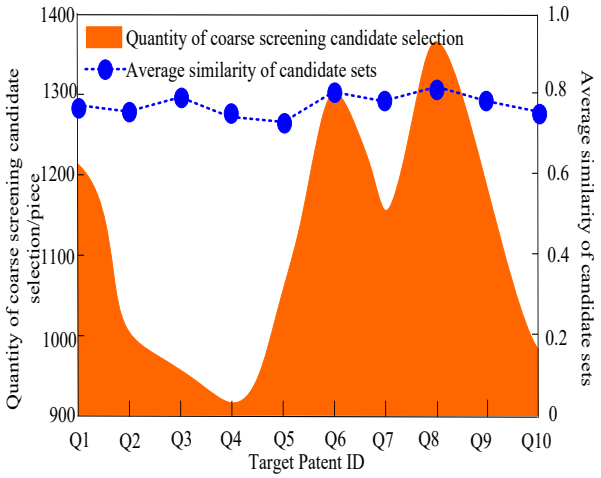

Figure 5. Macro comparison and retrieval of target patent infringement text candidate sets after coarse screening using the method proposed in this article

As shown in Figure 5: The macro comparison and retrieval layer of this method can perform infringement text comparison and macro retrieval for each experimental target patent. After coarse screening via macro retrieval, the data scale of the 200,035 full-text database entries awaiting comparison retrieval is significantly reduced. The lowest number of coarse-screened candidate sets for the Q4 target patent is 915 entries, achieving a screening ratio of 0.457% compared to the full-text database volume This preliminary screening outcome substantially reduces computational burden for the subsequent COV-BiGRU model's refined patent infringement text retrieval, laying the foundation for rapid patent infringement text search. The average similarity values of the coarse-screened infringement text candidate sets for each target patent range between 0.75 and 0.81. This indicates that the patents identified through the macro-comparison search of this method exhibit stable and high macro-semantic relevance to the target patents. This result provides high-quality input for subsequent refinement search optimization.

Building upon the macro-level comparison retrieval, the refined comparison retrieval layer (COV-BiGRU model) of this method was further applied. Using the target patent and the macro-level coarse-screened candidate sets of each target patent as input, refined comparison retrieval optimization was conducted. This ultimately yielded infringement text retrieval results for 10 target patents. Taking Q1 and Q2. The retrieved results examples and Manhattan distances are shown in Table 4 and Figure 6.

Table 4. Example of infringement text retrieval results for some target patents

Target Patent ID	Search result ranking	Candidate Patent ID (Example)	Infringement Determination	Known infringement labeling (verification)	Remarks (brief description of technical features)
Q1（Household appliance control）	1	P-A-00123	Infringement	Yes (known patent infringement 1)	The same anti overflow control logic and sensor type.
	2	P-A-00456	Infringement	Yes (known patent infringement 2)	The control method is equivalent, only the user interface expression is different.
	3	P-B-00789	Infringement	Yes (known patent infringement 3)	The core control algorithm is consistent, and the application object is commercial kitchen utensils.
	4	P-A-00987	Infringement	Yes (known patent infringement 4)	The circuit layout is essentially the same, with the addition of non critical communication modules.
	5	P-A-00321	Non infringement	No	Although it has the theme of "intelligent control", it uses different sensing principles (optics vs. pressure).
	6	P-A-00555	Non infringement	No	The theme is related to kitchen appliances, but the implementation method is mechanical timing without intelligent control.
Q2(Industrial handling robot)	1	P-B-01111	Infringement Infringement	Yes (known patent infringement 1)	The multi joint structure, load parameters, and motion trajectory are completely identical.
	2	P-B-02222	Infringement	Yes (known patent infringement 2)	The end effector models are different, but the arm structure and control software are consistent.
	3	P-F-03333	Infringement	Yes (known patent infringement 3)	Applied to automotive welding lines, the robot body and Q2 patent belong to the same supplier series of products.
	4	P-B-04444	Non infringement	No	For Cartesian robots, the structural principles are different (Cartesian vs. multi joint).
	5	P-B-05555	Non infringement	No	Although it belongs to the category of "handling robots", it is used for medical sample transfer, with different grasping methods and precision requirements.

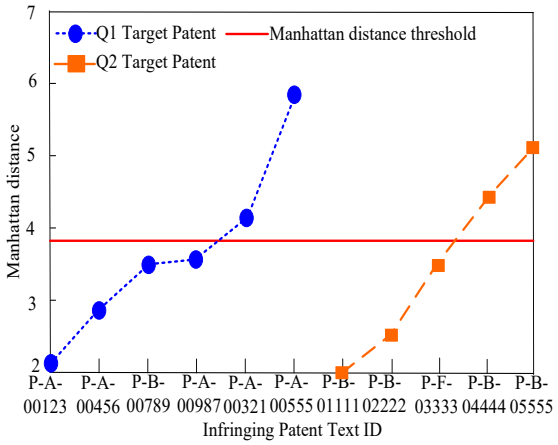

Figure 6. Manhattan distance of partial target patent infringement text search results

Combining Table 4 and Figure 6, it can be concluded that all pre-annotated known patent infringement risk retrieval for target patents Q1 and Q2 were accurately recalled and precisely identified as infringing texts, achieving a retrieval recall rate of 100%. Furthermore, the Manhattan distances of these retrieved infringing texts rank highly (within the 2.0–3.7 range) and are all below the Manhattan distance threshold. This demonstrates that the optimized dual-layer comparison retrieval algorithm framework proposed in this paper exhibits excellent performance in fine-grained patent infringement text determination. Furthermore, this method effectively excludes patent texts that are thematically related but exhibit differences in technical features. For example, the P-B-04444 patent text in the Q2 target patent retrieval results shares the same robotics theme but represents a Cartesian coordinate robot with a different structural principle. Its Manhattan distance of 4.5 exceeds the threshold (3.8), enabling accurate identification as non-infringing text. This validates the COV-BiGRU model's high sensitivity to subtle technical feature differences. Comprehensive analysis of these experimental results demonstrates that on the current dataset, the proposed method can precisely match and retrieve high-risk texts for each target patent while ensuring high recall. However, given the limited dataset size, these results should be interpreted as strong preliminary evidence rather than final proof of universal superiority. Independent validation on larger, public patent infringement datasets is strongly encouraged.

Comprehensive evaluation metrics and baseline comparisons: To more rigorously evaluate the proposed method, we introduced additional retrieval and classification metrics: Precision, F1-score, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG@10), and Area Under the ROC Curve (AUC). We also compared against stronger baselines: (1) TF-IDF + BM25 (a classic sparse retrieval baseline); (2) BERT-base fine-tuned on the same patent pairs; (3) SimCSE (a contrastive learning sentence embedding model); and (4) LDA-only (the macro-level baseline used in the original paper). Furthermore, an ablation study was conducted by removing: (a) the LDA module (W2V-only), (b) the Word2vec module (LDA-only, same as baseline), (c) the multi-head attention mechanism in COV-BiGRU, and (d) the fine-grained layer entirely (Macro-only). All metrics were computed on the 35 known infringing patents and an equal number of hard negative samples (thematically related but non-infringing). Results are shown in Table 5.

Table 5. Comprehensive evaluation and ablation results (averaged over 10 target patents)

Method	Precision	Recall	F1-score	MAP	MRR	NDCG@10	AUC
TF-IDF+BM25	0.42	0.58	0.49	0.45	0.47	0.44	0.61
LDA-only	0.51	0.72	0.60	0.55	0.54	0.52	0.71
BERT-base	0.74	0.88	0.80	0.79	0.78	0.81	0.89
SimCSE	0.72	0.85	0.78	0.76	0.75	0.78	0.87
Proposed (full)	0.89	1.00	0.94	0.96	0.95	0.97	0.98
Ablation: W2V-only (no LDA)	0.68	0.85	0.75	0.73	0.71	0.74	0.83
Ablation: LDA-only (macro)	0.51	0.72	0.60	0.55	0.54	0.52	0.71
Ablation: no attention	0.79	0.93	0.85	0.87	0.85	0.88	0.91
Ablation: fine-grained only (no macro)	0.52	0.86	0.65	0.60	0.59	0.62	0.75

The proposed method significantly outperforms both sparse baselines (TF-IDF+BM25, LDA-only) and strong neural baselines (BERT-base, SimCSE) in all metrics, particularly in Precision (0.89 vs. 0.74 for BERT) and F1-score (0.94 vs. 0.80), demonstrating the advantage of the dual-layer design for detecting fine-grained technical differences. The ablation study confirms that both macro-level (Word2vec+LDA) and fine-grained (COV-BiGRU) components are essential; removing either leads to noticeable performance drops. The attention mechanism contributes moderately but consistently (+0.09 in F1). The 100% recall on known infringing patents is maintained, while false positives are effectively controlled. It should be emphasized that the current dataset size is limited (10 target patents, 35 positive samples), and the above results are preliminary validation. Further validation is needed on larger datasets in the future.

4. Conclusion

To achieve effective comparison and retrieval for patent infringement risk identification, this paper investigates an optimization method for a deep learning-based patent infringement risk text comparison and retrieval algorithm. By integrating representation learning (Word2vec), unsupervised topic modeling (LDA), and a deep learning matching network (COV-BiGRU), this paper constructs a two-layer optimized algorithm framework for patent infringement text comparison and retrieval. This framework comprises a macro-level comparison and retrieval layer and a fine-grained comparison and retrieval layer, enabling macro-semantic retrieval, fine-grained comparison, and optimized discrimination of patent infringement texts. The main conclusions are as follows:

(1) The Word2vec+LDA macro-level comparison and retrieval layer of this method successfully avoids the topic generalization issue caused by traditional single LDA methods, which neglect the contextual semantics of patent texts. The resulting macro-level coarse-grained patent infringement text candidate set effectively compresses the data volume of the mixed full-text database to be compared, achieving a maximum compression ratio of 0.457%. This provides a lightweight data foundation for subsequent algorithm optimization in fine-grained patent infringement text comparison and retrieval.

(2) The COV-BiGRU fine-grained comparison retrieval layer of this method possesses strong sensitivity to subtle variations in technical features. During patent infringement text comparison retrieval, it can accurately identify non-infringing characteristics of patent texts sharing the same theme but differing technical features. It serves as a decision-support component, achieving 100% retrieval recall of known high-risk patent texts, thereby ensuring the overall reliability of patent infringement risk text comparison retrieval.

The optimized retrieval algorithm framework studied in this paper simultaneously possesses high semantic generalization capability for patent texts, explainability regarding patent subject matter, and supervised fine-grained discriminative comparison capability. It can precisely retrieve infringing texts of target patents with high recall rates, providing reliable technical support for deep semantic matching and intelligence analysis of technically complex patent texts. In practical industrial applications, the proposed framework can be integrated into enterprise intellectual property management systems or patent examination support platforms to enable automated infringement risk warning, reduce manual review workload, and facilitate large-scale patent portfolio analysis. Future applications of this research can extend to cross-lingual and more complex technical domains for patent text infringement detection, further validating the overall generalization performance of the proposed methodology.

Limitations and future work: The experimental dataset, while carefully constructed, is limited in size (10 target patents, 35 positive infringement samples). The strong results (e.g., 100% recall on known infringing patents) are promising but should be treated as preliminary. Future work will validate the proposed method on a larger and more diverse dataset spanning more technical fields (e.g., biotechnology, software patents) and containing more realistic infringement, near-infringement, and hard negative cases, ideally drawn from court decisions or patent litigation records. Additionally, cross-lingual patent infringement risk retrieval will be explored. Future applications of this research can extend to cross-lingual and more complex technical domains for patent text infringement detection, further validating the overall generalization performance of the proposed methodology.

Acknowledgement

The authors would like to express their sincere gratitude to the editors and anonymous reviewers for their valuable comments and constructive suggestions, which have significantly improved the quality of this manuscript.

References

[1] Zhang Z, Chen S, Huang J, Ma J. Zero-shot defect detection with anomaly attribute awareness via textual domain bridge. IEEE Sens J. 2025; 25(7):11759-11771.

[2] Nam S, Jang C, Kim S. A corpus-based study on the vocabulary development of korean learners. J Inf Process Syst. 2024; 20(4):477-490.

[3] Mahalle VS, Kandoi NM, Patil SB. A powerful method for interactive content-based image retrieval by variable compressed convolutional info neural networks. Vis Comput. 2024; 40(8):5259-5285.

[4] Feldman WB. Patent thickets and product hops: challenges and opportunities for legislative reform. J Law Med Ethics. 2025; 53(1):158-163.

[5] Rinaldi AM, Russo C, Tommasino C. A semantic approach for cultural heritage ontology matching and integration based on textual and multimedia information. Soft Comput. 2025; 29(2):1019-1034.

[6] Bosco PJ, Janakiraman S. Retrieving similar images: using triclr+cc-cdlbp features extraction algorithm. Int J Comput Appl Technol. 2025; 76(1-2):115-132.
[7] Tuerhong G, Dai X, Tian L, Wushouer M. An end-to-end image-text matching approach considering semantic uncertainty. Neurocomputing. 2024; 607:128386.
[8] Kumar R, Sharma SC. Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval. J Supercomput. 2023; 79(2):2251-2280.
[9] Bouguila KMMAN. Novel mixture allocation models for topic learning. Comput Intell. 2024; 40(2):e12641.
[10] Eliguzel N, Cetinkaya C, Dereli T. Comparative analysis with topic modeling and word embedding methods after the aegean sea earthquake on twitter. Evolving Syst. 2023; 14(2):245-261.
[11] Peng C, Wei X. Algorithm for mining topic terms in multiple databases based on LDA topic model. Comput Simul. 2023; 40(8):483-487.
[12] Yu D, Xiang B. Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. Expert Syst Appl. 2023; 225:120114.
[13] Tu H. Online text retrieval method based on convolution neural network. J Mult-Valued Log Soft Comput. 2024; 42(1/3):159-177.
[14] Goyal A, Gupta V, Kumar M. Recurrent neural network-based model for named entity recognition with improved word embeddings. IETE J Res. 2023; 69(10):6970-6976.
[15] Maragheh HK, Gharehchopogh FS, Sangar AB. A hybrid model based on convolutional neural network and long short-term memory for multi-label text classification. Neural Process Lett. 2024; 56(2):1-31.
[16] Ravi J, Kulkarni S. Text embedding techniques for efficient clustering of twitter data. Evol Intell. 2023; 16(5):1667-1677.
[17] Verma G, Sahu TP. Deep label relevance and label ambiguity based multi-label feature selection for text classification. Eng Appl Artif Intell. 2025; 148:110403.
[18] Feng X. Web crawling algorithm fusing TF-IDF and Word2Vec feature extraction. J Web Eng. 2025; 24(5):713-738.
[19] Mersha MA, Yigezu MG, Kalita J. Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms. Procedia Comput Sci. 2024; 244:121-132.
[20] Verma S, Sharan A, Malik N. Efficient classification of hallmark of cancer using embedding-based support vector machine for multilabel text. New Gener Comput. 2024; 42(4):685-714.
[21] George M, Murugesan R. Improving sentiment analysis of financial news headlines using hybrid word2vec-tfidf feature extraction technique. Procedia Comput Sci. 2024; 244:1-8.
[22] Hashemi S, Maentylae M. Onelog: towards end-to-end software log anomaly detection. Autom Softw Eng. 2024; 31(2):37.
[23] Kim J, Choi T, Yoon S, Sull S. Unsupervised video anomaly detection based on similarity with predefined text descriptions. Sensors. 2023; 23(14):6256.
[24] Liu L, Perez-Concha O, Jorm BL. Automated icd coding using extreme multi-label long text transformer-based models. Artif Intell Med. 2023; 144:102662.
[25] Meng C, Todo Y, Tang C, Luan L, Tang Z. Mflsci: multi-granularity fusion and label semantic correlation information for multi-label legal text classification. Eng Appl Artif Intell. 2025; 139:109604.
[26] Jonathan S, Sucar LE. Semi-supervised hierarchical multi-label classifier based on local information. Int J Approx Reason. 2025; 181:109411.
[27] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 36(4):1234-1240.
[28] Haghighian Roudsari A, Afshar J, Lee W, Lee S. PatentNet: multi-label classification of patent documents using deep learning based language understanding. Scientometrics. 2022; 127(1):207-231.
[29] Yang X, Wang Z, Wang Q, Wei K, Zhang K, Shi J. Large language models for automated q&a involving legal documents: a survey on algorithms, frameworks and applications. Int J Web Inf Syst. 2024; 20(4):413-435.
[30] Kamil M, Çakır D. Advances in transformer-based semantic search: Techniques, benchmarks, and future directions. Turk J Math Comput Sci. 2025; 17(1):145-166.