applied sciences

MDPI

Article

A Hybrid Deep Learning and Optimization Model for Enterprise Archive Semantic Retrieval

Xiaonan Shi [1,*], Junhe Chen [1], Yumo Wang [2] and Limei Fu [1]

[1] School of Artificial Intelligence and Computing, Xi'an University of Science and Technology, Xi'an 710054, China; changer_he@163.com (J.C.); fulimei@xust.edu.cn (L.F.)
[2] School of Electronics and Signals, Northwestern Polytechnical University, Xi'an 710072, China; 19407020420@stu.xust.edu.cn
* Correspondence: shi_xn@xust.edu.cn

Abstract

By searching for and summarizing the relevant information of the enterprise, we can build relevant knowledge maps, supplement and enrich the existing knowledge base, and support existing experiments and subsequent algorithm improvements. The extracted input text of enterprise archives is described via relation extraction and semantic analysis to improve the efficiency of archive retrieval and reduce the cost of communication. On the basis of the analysis of previous research, an enterprise archive semantic retrieval algorithm based on deep learning technology is constructed, that is, the BERT + BiGRU + CRF + HHO_improved model, to extract the relevant information of the enterprise. In the model, the Bidirectional Encoder Representations from Transformers (BERT) model is used to preprocess the Chinese word embedding, and the question-and-answer data are generated from the actual enterprise file database. Next, a Bidirectional Gated Recursive Unit (BiGRU) is used with the attention mechanism to capture the contextual features of the sequence. The Conditional Random Field (CRF) classifier is subsequently used to classify the text related to the enterprise archives, and the obtained data are labeled in sequence. Moreover, the swarm intelligence algorithm is introduced to dynamically optimize the model parameters and data processing strategies further to improve the generalization ability and adaptability of the model. The Harris Hawks Optimizer Improved (HHO_improved) algorithm is used to optimize the parameters of the CRF module to increase the performance and efficiency of named entity recognition. On the independently constructed dataset, the advantages of our algorithm are verified via comparative experiments with a variety of semantic matching algorithms and ablation experiments on the CRF and HHO_improved. The CRF and HHO_improved play essential roles in improving model performance. The obtained knowledge extraction results are expected to supplement and enhance the existing knowledge base, simplify the workflow, assist the enterprise's dynamic production task management, and improve the efficiency of enterprise operations. The proposed algorithm achieves an accuracy improvement of 36.33%, 43.88%, 15.24%, and 12.41% over the BERT, BiGRU, BERT + BiGRU, and BERT + BiGRU + CRF models, respectively.

Keywords: semantic retrieval; enterprise archives; deep learning; dynamic optimization; Harris Hawks Optimizer; Named Entity Recognition

Academic Editor: Douglas O'Shaughnessy

Received: 13 September 2025
Revised: 7 November 2025
Accepted: 14 November 2025
Published: 21 November 2025

Citation: Shi, X.; Chen, J.; Wang, Y.; Fu, L. A Hybrid Deep Learning and Optimization Model for Enterprise Archive Semantic Retrieval. *Appl. Sci.* **2025**, *15*, 12381. https://doi.org/10.3390/app152312381

1. Introduction

With the advancement of technology and the accumulation of archival information resources in the era of big data, the use of archival information resources to increase archival service capabilities and promote the intelligence of archival management has become an urgent task for data resource management [1]. Current computer retrieval systems use controlled languages, but compared with natural languages, controlled languages have the disadvantages of indexing difficulties, slow speed, delayed vocabulary updating, and high requirements for indexing and retrieval personnel [2]. Therefore, to change the traditional mode based on keywords and metadata detection in the past, these systems can retrieve unstructured text from enterprise document libraries; process the text through Natural Language Processing (NLP), deep learning, and other technologies; understand the meaning and strength of search objectives in multiple dimensions in document management systems; and determine the most suitable search results for users [3]

Deep learning is an emerging technology in the field of machine learning. In recent years, breakthroughs have been made in many application fields [4]. Unlike traditional machine learning models, deep learning can transfer learned features from similar tasks through multilayer feature extraction, thus showing unique advantages in complex document content analysis [5]. By collecting nonlinear deep network structures and distributed text data features, deep learning algorithms can accurately classify text data [6,7]. Convolutional neural networks (CNNs) achieved the earliest success in image classification. Subsequently, it was applied to NLP tasks to solve problems such as part-of-speech tagging, human–computer interaction question answering, text summarization, and named entity recognition [8].

Semantic retrieval is the development trend of information retrieval. As early as 1980, the concept of semantic retrieval emerged and has been studied in the field of information retrieval [9]. However, owing to the lagging development of multimedia utilization and retrieval technology for archival resources, the retrieval and utilization of archival multimedia resources are unsatisfactory [10]. To improve their effectiveness, scholars have proposed retrieval methods by considering different technical approaches. Zhou Jianfeng integrated the ontology concepts in semantic models into query extension technology and proposed a local document analysis query extension method based on ontology [11]. To enrich the index structure, Qi Baoyuan and Cao Cungen et al. proposed a semantic retrieval method for domain knowledge documents. By expanding the relationship between subject words, a secondary index structure was constructed from subject words to documents [12]. Then, Jin Biyi et al. proposed a semantic annotation strategy for ontology entities that maps entities in documents to instances in the ontology knowledge. Semantic queries were implemented by indexing user query conditions and instances [13]. In terms of digital archive resources, Lv Yuanzhi proposed cross-media aggregation from the perspective of semantic association. A specific semantic association aggregation implementation framework was constructed using the association data technology framework [14]. In terms of legal cases, Zhang Yunting et al. introduced case elements that can highlight legal semantics, and by modeling cases based on them, a semantic-based similar case retrieval algorithm was proposed [15].

Recent research on text embedding has demonstrated remarkable progress. For instance, Li Zehan constructs a General Text Embedding (GTE) model based on multi-stage contrastive learning, which integrates multi-source heterogeneous data and an improved contrastive objective to achieve leading performance across both text and code tasks [16]. Similarly, Chankyu Lee significantly enhances the performance of Large Language Model (LLM)-based embedding models by introducing a latent attention layer, removing causal masks, designing a two-stage training strategy, and optimizing data construction, achieving

state-of-the-art results on the Massive Text Embedding Benchmark (MTEB) and Audio Instruction Benchmark (AIR-bench) [17]. In addition, Wang Liang proposes a new paradigm for text embedding training based on synthetic data, adopting a simplified contrastive learning framework that eliminates manual data annotation while improving multi-task semantic representation [18]. This rapid evolution is further evidenced by studies combining synthetic data with LLMs [19], adapting large models for dense retrieval [20], and leveraging entailment signals for fine-tuning [21]. These advancements provide robust, contemporary baselines that are highly relevant to real-world enterprise needs.

However, research in the domain of semantic retrieval for enterprise archives remains limited. Due to factors such as enterprise implementation gradients, most traditional enterprises continue to rely on conventional semantic retrieval methods, including vector space models [22], query expansion and knowledge bases [23], and latent semantic analysis [24]. These traditional approaches rely solely on surface-level lexical features, lack deep semantic understanding, cannot model dynamic contextual information, and incur high maintenance costs for rules and knowledge bases, all of which constitute barriers to enterprise development [25]. Some large-scale internet enterprises have begun experimenting with large language model-based semantic retrieval, such as Amazon's utilization of BERT for understanding ambiguous queries [26]. However, large-scale deployment in traditional enterprises still faces significant obstacles. Therefore, the construction of novel semantic retrieval systems for enterprise archives has become an urgent imperative [27].

Given the shortcomings of traditional semantic retrieval approaches, it is essential to adopt models capable of understanding complex semantic relations and adapting to diverse enterprise text structures. Therefore, this study integrates BERT, BiGRU, CRF, and the HHO_improved algorithm. BERT provides deep contextual representations suitable for enterprise-specific terminology, BiGRU captures sequential dependencies in textual records, CRF maintains label consistency in structured information extraction, and the HHO_improved algorithm adaptively optimizes model parameters for enhanced performance. The selection of these models is thus driven by their compatibility with the semantic, sequential, and dynamic characteristics of enterprise archival data.

To address the limitations in enterprise archive semantic retrieval, this paper makes the following key contributions:

1. We propose a hybrid semantic retrieval model that integrates BERT, BiGRU, and CRF for enterprise archive retrieval. The framework leverages BERT for deep semantic representation, employs BiGRU to capture contextual dependencies in document sequences, and utilizes CRF for structured entity labeling. An HHO_improved algorithm is further incorporated to dynamically tune model parameters, enhancing both retrieval accuracy and cross-scenario robustness.
2. We incorporate a knowledge graph into the semantic retrieval pipeline to enrich entity-relation understanding. This integration helps bridge the gap between lexical-level matching and true semantic comprehension, enabling more accurate reasoning under complex or ambiguous enterprise queries.
3. We introduce an enhanced HHO strategy for adaptive optimization of the CRF model parameters. By simulating collective hunting behaviors with improved convergence properties, the algorithm effectively avoids local optima and strengthens model stability when applied to dynamic enterprise corpora.
4. We evaluate the proposed approach on a large-scale enterprise log dataset, demonstrating consistent improvements in retrieval accuracy and stability over strong baseline methods.

This paper is organized as follows: beginning with an exposition of the proposed model in Section 2, proceeding to the experiments and results in Section 3, and concluding with a discussion of the findings in Section 4.

2. Research on Enterprise Archives Semantic Retrieval Algorithm

2.1. Bert + BiGRU + CRF + HHO_Improved Architecture

This study develops an independently constructed enterprise profile retrieval dataset and a novel Bert + BiGRU + CRF + HHO_improved architecture.

The BERT + BiGRU + CRF + HHO_improved model is a deep learning model that combines BERT, BiGRU, CRF, and the Harris Hawk Optimization Improved algorithm for Named Entity Recognition (NER) tasks. Specifically, in the BERT + BiGRU + CRF + HHO_i mproved model, the input text is first encoded via BERT to obtain a representation of the text. Next, the representation of the BERT output is input into a bidirectional gated cyclic unit, and the BiGRU model learns the forward and backward information of the textual representation and concatenates the results of the bidirectional representation. Finally, the cascade representation is input into a conditional random field with a recursive structure. The Harris Hawk Optimization Improved algorithm is subsequently used to optimize the parameters of the CRF model to improve the performance of NER. The BERT + BiGRU + CRF + HHO_improved model is more effective in the NER task because it can accurately identify entities. By integrating a variety of technologies, the model fully excavates the semantic and contextual information of the text and comprehensively considers the relationships between the labels, which improves the accuracy and efficiency of the NER task. The model diagram is shown in Figure 1.

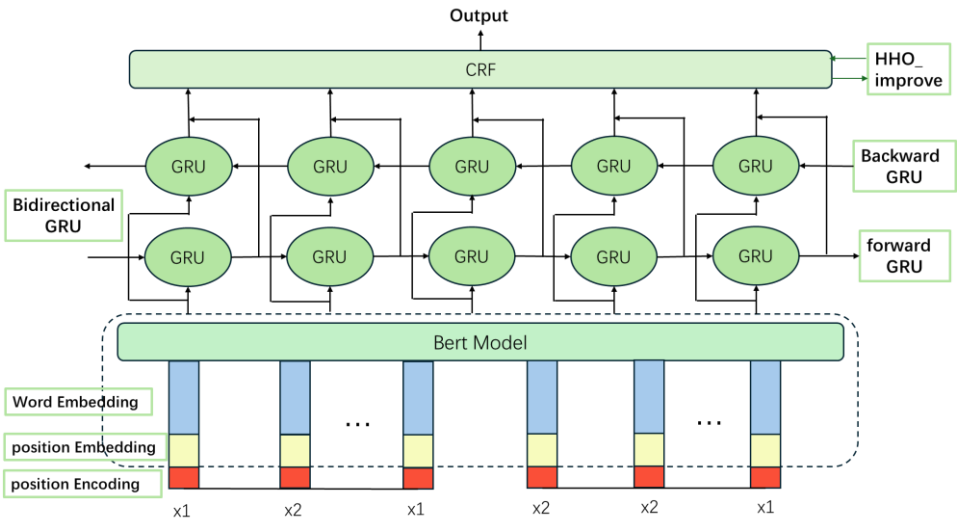

Figure 1. Bert + BiGRU + CRF + HHO_improved model diagram.

In the BERT + BiGRU + CRF + HHO_improved model, the combination of BERT, BiGRU, and CRF is the same as that in the traditional BERT + BiGRU + CRF model. In the CRF section, the HHO_improved algorithm is used to optimize the CRF model parameters to improve the NER performance. Specifically, in the CRF model, the main parameters that must be optimized include the weights W and the transfer matrix T in the function. The HHO_improved algorithm can find the optimal solution in the parameter space through hybrid operations and heuristic searches and improve the performance of the model. These enhancements enable a more balanced trade-off between exploration and exploitation, thereby improving convergence stability and solution diversity.

2.2. BERT-Based Chinese

In recent years, researchers have employed pretrained deep neural networks as language models, achieving strong performance by fine-tuning them for domain-specific tasks [28–30]. A typical probabilistic language model estimates the likelihood of a sentence S as the joint probability of its constituent words, calculated sequentially from left to right, as shown in Equation (1):

$$p(S) = p(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} p(w_i | w_1, w_2, \ldots, w_{i-1}). \quad (1)$$

S denotes the sentence, ω_i represents the i-th word in the sentence, and m is the total number of words. The term $p(w_i|w_1, w_2, \ldots, w_{i-1})$ represents the conditional probability of the next word ω_i given all previous words. Equation (1) indicates that the sentence probability is obtained by multiplying the conditional probabilities of each word given its previous context.

Unlike the traditional left-to-right model, BERT employs a masked language modeling objective to achieve bidirectional contextual understanding.

BERT-base-Chinese is a deep learning model designed for NLP tasks. Specifically, it is a Chinese pretrained language model built upon the BERT architecture originally proposed by Google. The structure of the BERT model is illustrated in Figure 2. The main idea is to use the transformer architecture to pretrain the large-scale text data pretraining model so that the algorithm can quickly and efficiently perform tasks such as semantic understanding and reasoning of Chinese text. The BERT-based Chinese model has been pretrained on large-scale Chinese text data and can handle cross-language text data in both Chinese and English. When the BERT-based Chinese algorithm is used, the pretrained model parameters can be used directly, and further fine-tuning operations can be carried out on this basis. The algorithm performs well in several NLP tasks, such as question-answering systems, text classification, and relation extraction.

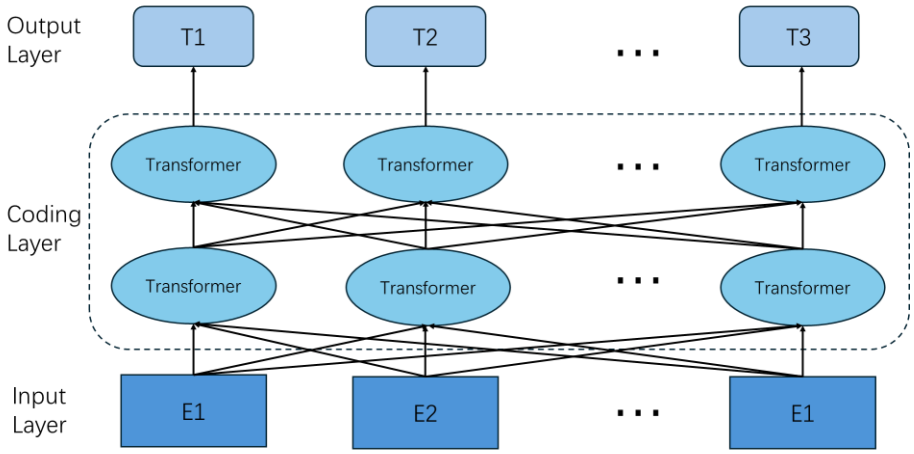

Figure 2. The architecture of the BERT pretrained language model.

However, this unidirectional approach only captures forward dependencies and cannot fully represent bidirectional contextual information, which limits the model's understanding of word semantics.

2.3. BiGRU

A Bidirectional Gated Recurrent Unit (BiGRU) is a bidirectional recurrent neural network that simultaneously models text sequences via context information. This network is commonly used for tasks such as sequence labeling and text classification. The BiGRU

algorithm is applied to the text analysis task of enterprise archive retrieval. First, the dataset is divided into training, validation, and test sets, and the document is preprocessed and cleaned. This process includes word segmentation, removing stop words, and extracting keywords from the document. The BiGRU algorithm is used to train on the training set. Parameter tuning and model selection are performed on the validation set to determine the best algorithm. The optimal algorithm is then applied to the lower set of tests to evaluate the accuracy, recall, and F1 value of the algorithm.

The BiGRU uses forward and backward GRUs for context information feature extraction to weight the output, map the d-dimensional vector to the m-dimensional vector through the linear layer, and obtain the final output label vector list of the BiGRU network, where n is the length of the text sequence and m is the number of entity type labels. The structure of the BiGRU is shown in Figure 3.

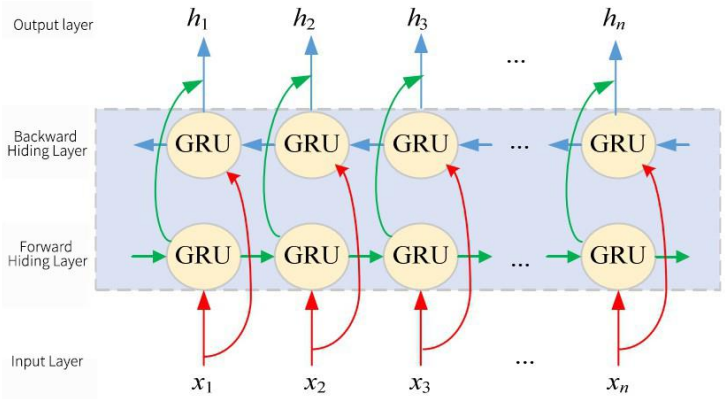

Figure 3. BiGRU structure.

The BiGRU layer is employed to capture contextual information from both past and future sequences. The calculation process is shown in Formulas (2)–(4):

$$\underset{h_t}{\rightarrow} = GRU\left(x_t, \underset{h_{t-1}}{\longrightarrow}\right), \tag{2}$$

$$\underset{h_t}{\leftarrow} = GRU\left(x_t, \underset{h_{t-1}}{\longleftarrow}\right), \tag{3}$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]. \tag{4}$$

Specifically, the forward hidden state $\overrightarrow{h_t}$ is computed by the GRU unit using the current input x_t and the previous forward hidden state $\overrightarrow{h_{t-1}}$, as shown in Equation (2). The backward hidden state $\overrightarrow{h_t}$ is obtained in a similar manner by processing the sequence in the reverse temporal order, as described in Equation (3). Finally, the two directional hidden states are concatenated to form the overall hidden representation h_t, as indicated in Equation (4). This output $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$ integrates information from both directions, thereby capturing complete contextual dependencies within the sequence.

2.4. CRF

The Conditional Random Field (CRF) model is a classical sequence labeling model for labeling a given input sequence. It labels the entire sequence as a whole, taking into account the conditional probabilities between adjacent labels. The BERT + BiGRU + CRF model combines the advantages of BERT, BiGRU, and CRF. It can effectively handle sequence annotations in NLP tasks, such as NER, and performs well. The BERT model can capture the

semantic information of the text, the BiGRU model can capture the sequence information of the text, and the CRF model can model the relationships between tags more accurately. Combining the three can improve the performance of NLP tasks.

The CRF module is mainly used to study the label information of adjacent data, automatically constrains the prediction score of the BERT + BiGRU network output, ensures that the production is as legal as possible, and reduces the probability of illegal sequence output by 50%.

For the input and predicted output sequences, the score can be represented by Equation (5), which is the sum of the transition probability and the state probability.

$$S(X, y) = \sum_{i=0}^{n} (A_{yi,yi+1} + P_{i,yi}). \tag{5}$$

The summation runs from $i = 0$ to n, where n is the length of the input sequence. The terms $A_{yi,yi+1}$ represent the transition scores from the tag at position i to the tag at position $i + 1$. To properly model the transitions at the sequence boundaries, special start y_0 and end (y_{n+1}) tags are introduced. The term $P_{i,yi}$ denotes the state score, which is the output from the BERT+ BiGRU network, for the i-th token being assigned the tag y_i.

Using the Softmax function, the label sequence Y is obtained, and the probability value shown in Equation (6) is obtained:

$$p(y|X) = \frac{e^S(X, y)}{\sum_{\overline{y} \in Y_X} S(X, \overline{y})}. \tag{6}$$

Each node in the CRF network represents a predicted value. According to the prediction sequence of the BERT + BiGRU output, the method finds the most likely path in the network, determines the label of the specified entity, and realizes entity recognition. Therefore, the goal of training is to maximize the probability. This can be achieved via log-likelihood, as shown in Equation (7):

$$logp(y|X) = S(X, y) - \sum_{i=0}^{n} S(x, yi). \tag{7}$$

Finally, the prediction is decoded via the Viterbi algorithm to obtain the best path to solve, as expressed in Formula (8):

$$y^* = \arg\max_{y \in Y_x} S(x, y). \tag{8}$$

2.5. Harris Hawks Optimization-Improved Algorithm

Harris Hawks Optimization (HHO) is a metaheuristic algorithm proposed by Heidari et al. in 2019 [30]. It is designed to solve complex optimization problems by mimicking the cooperative behavior and surprise pounce of Harris' hawks in nature. Distinguished by its dynamic exploration and exploitation phases and its adaptive transition strategy, HHO has demonstrated remarkable efficacy across a wide range of engineering and scientific disciplines. This exposition delineates the core mathematical model and the operational mechanics of the HHO algorithm.

As shown in Figure 4, (a) Conceptual illustration of a Harris hawk adapting its flight path to avoid environmental obstacles during prey pursuit.

(b) Formal workflow of the Harris Hawks Optimization algorithm, the HHO algorithm operates in two primary phases, governed by the escaping energy of the prey, denoted as E. And the transition between soft and hard besiege strategies.

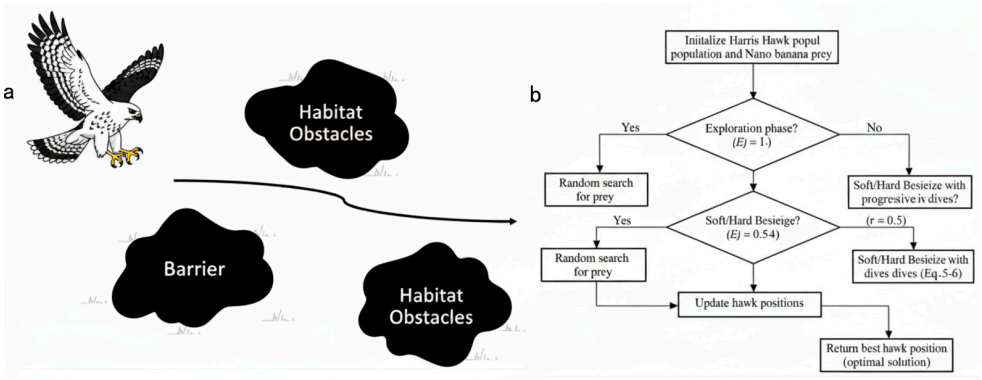

Figure 4. Harris Hawks Optimization algorithm.

2.5.1. Exploration Phase

In this phase, the hawks perch randomly and await detection of prey based on two strategies. If $q < 0.5$, they perch based on the positions of other family members and the prey. If $q \geq 0.5$, they perch on a random location within the group's home range. The position update for a hawk at iteration $t+1$ is given by Equation (9):

$$\begin{aligned} &\text{if } q \geq 0.5: \\ &\quad X(t+1) = X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 X(t)|, \\ &\text{if } q < 0.5: \\ &\quad X(t+1) = (X_{rabbit}(t) - X_m(t)) - r_3(LB + r_4(UB - LB)), \end{aligned} \tag{9}$$

where $X(t)$ denotes the current position of a hawk, representing a candidate solution in the search space, $X_{rabbit}(t)$ is the position of a randomly selected hawk from the population, $X_{rand}(t)$ denotes the position of the prey -the best current solution, X_m is the average position of the population, and r_1, r_2, r_3, r_4, and q are random numbers within (0,1). LB and UB define the lower and upper bounds of the search space.

2.5.2. Transition from Exploration to Exploitation

The algorithm switches to exploitation based on the escaping energy of the prey, E, which decreases over iterations: The transition between exploration and exploitation in the HHO algorithm is controlled by the escaping energy of the prey, denoted as E. This parameter simulates the prey's decreasing energy over time and determines the hawks' hunting strategy. The escaping energy at iteration t is calculated as

$$E = 2E_0\left(1 - \frac{t}{T}\right). \tag{10}$$

where E_0 is the initial energy, a random number uniformly distributed in the range $(-1, 1)$, t is the current iteration, and T is the maximum number of iterations. If $|E| \geq 1$, the prey still has high energy, and the algorithm focuses on exploration of the search space. If $|E| < 1$, the prey's energy decreases, and the algorithm transitions to the exploitation phase for local refinement around promising regions.

2.5.3. Exploitation Phase: Soft and Hard Besiege

This phase employs four distinct strategies based on the prey's energy and a random chance of escape. The strategies model soft besiege, hard besiege, and each with or without progressive rapid dives. Two main cases are considered: soft besiege and hard besiege, which represent different hunting behavior.

Soft Besiege ($|E| \geq 0.5$ and $r \geq 0.5$): The prey has energy but is softly surrounded. Hawks update their position using

$$X(t+1) = \Delta X(t) - E|JX_{rabbit}(t) - X(t)|. \tag{11}$$

$\Delta X(t) = X_{rabbit}(t) - X(t)$ is the difference vector, and $J = 2(1 - r_5)$ simulates the prey's random jump strength.

Hard Besiege ($|E| < 0.5$ and $r \geq 0.5$): The prey is exhausted and surrounded. Hawks move closer with a simple update:

$$X(t+1) = X_{rabbit}(t) - E|\Delta X(t)|. \tag{12}$$

For cases where $r < 0.5$, the prey has a chance to escape. The algorithm uses a Lévy flight to model the deceptive prey movements and the hawks' dive. A feasibility check between a Lévy flight-based position and a random dive determines the final update, ensuring a more robust and stochastic local search.

2.5.4. Improved Algorithm

As E was defined in Equation (10), we introduce two modifications to improve the adaptive control of the escaping energy and to alleviate premature convergence.

Negative-feedback factor for escaping energy

We replace the original linear escaping energy by a feedback-scaled energy:

$$E_1 = 2(1 - \frac{t}{T})\alpha, \tag{13}$$

where α is a negative-feedback factor dynamically adjusted according to the population fitness change:

$$\Delta f = \left| f_{\text{rabbit}} - \min(f_{\text{population}}) \right|, \ \alpha \leftarrow \alpha \left(1 - \frac{\Delta f}{f_{\text{rabbit}}} + \varepsilon \right). \tag{14}$$

Here f_{rabbit} denotes the current best fitness, $\min(f_{population})$ is the minimum fitness in the population, and ε is a small constant to avoid division by zero. The factor α decreases when the fitness gap narrows, which reduces over-exploitation and helps preserve diversity.

Adaptive initialization of E_0.

To promote escape from local optima when the population has converged prematurely, the initial escaping energy E_0 is sampled from a wider range when population fitness variance is small:

$$E_0 \sim \begin{cases} \mathcal{U}(-1.5, 1.5), & \text{if } \sigma_f < \tau, \\ \mathcal{U}(-1, 1), & \text{otherwise,} \end{cases} \tag{15}$$

where σ_f is the standard deviation of population fitness and τ is a small threshold. Enlarging the sampling range when $\sigma_f > \tau$ increases exploration capability under stagnation.

These simple adaptations improve the balance between exploration and exploitation and empirically reduce premature convergence in enterprise-scale search tasks.

2.6. Bert + BiGRU + CRF + HHO_Improved Algorithm Steps

The complete BERT + BiGRU + CRF + HHO_improved pipeline for enterprise profile named entity recognition consists of six major stages: data preprocessing, contextual encoding, sequence modeling, structured prediction, parameter optimization, and final inference. The overall workflow is illustrated below and Figure 5.

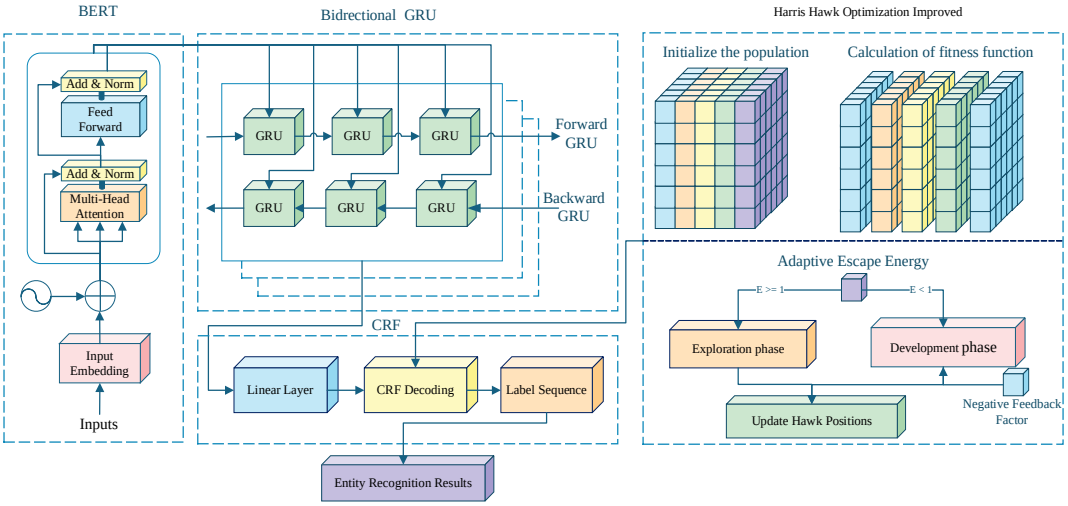

Figure 5. Bert + BiGRU + CRF + HHO_improved Algorithm flowchart.

Step 1: Data Preprocessing

The original enterprise profile documents are first cleaned and normalized. This process includes sentence segmentation, tokenization, punctuation removal, stop-word filtering, and conversion to the input format required by the BERT tokenizer. Labels in the dataset are aligned with the tokenized sequences to ensure compatibility with the downstream sequence labeling model.

Step 2: BERT Contextual Encoding

Each tokenized sentence is fed into the BERT-base-Chinese model, which produces a contextualized embedding for every token. The BERT encoder captures global bidirectional semantic dependencies through the architecture and masked language modeling, generating the hidden representation:

$$H^{\text{BERT}} = h_1^{\text{BERT}}, h_2^{\text{BERT}}, \ldots, h_n^{\text{BERT}}. \tag{16}$$

Step 3: BiGRU Sequence Feature Extraction

The BERT output sequence is then input into a Bidirectional GRU network to further model sequential dependencies. The forward and backward GRU units compute $\overrightarrow{h}_i, ; \overleftarrow{h}_i$, and the final hidden state representation is obtained via concatenation, $h_i = [\overrightarrow{h}_i; \overleftarrow{h}_i]$. This step enhances the model's ability to capture long-range interactions and entity boundaries.

Step 4: CRF-Based Structured Prediction

The BiGRU output is passed to a Conditional Random Field layer, which models the transition dependencies between adjacent labels. Given emission scores from the BiGRU and learnable transition scores, the CRF computes the sequence-level score and decodes the most probable entity label path using the Viterbi algorithm. This ensures globally consistent and legally constrained label predictions.

Step 5: HHO_improved Optimization of CRF Parameters

During training, the CRF parameters—including emission weights and transition matrix—are further optimized using the HHO_improved algorithm. The improved HHO introduces adaptive escaping energy and feedback-adjusted exploration–exploitation balancing to avoid premature convergence. The population of hawks iteratively updates candidate parameter sets based on the prey (optimal solution) and converges toward a globally optimal CRF parameter configuration.

Step 6: Final Inference and Entity Extraction

After training is complete, the optimized model performs inference on unseen enterprise documents. Input text is encoded by BERT, processed by the BiGRU layer, and decoded by the CRF using the learned transition structure. The resulting label sequence identifies the entity spans, which are then mapped back to the original text to produce the final NER results.

3. Experiments and Discussion

3.1. Experiments Design

To validate the effectiveness of the proposed BERT + BiGRU + CRF + HHO_improved model in the task of semantic retrieval of enterprise archives, this study designs three experiments: a benchmark test experiment, a swarm intelligence optimization performance comparison experiment, and an ablation study.

1. Benchmark Test Experiment

This experiment is designed to evaluate the performance of the proposed model in NER and semantic retrieval tasks, and to compare it with existing mainstream models to verify its advantages in semantic understanding of enterprise archives. The BERT component captures deep semantic information of the text, the BiGRU effectively models the contextual dependencies of text sequences, the CRF addresses label dependency issues, and the HHO_improved algorithm enhances the global optimality of model parameter selection. Theoretically, this combination fully exploits the contextual information and semantic correlations of the text, thereby improving entity recognition and retrieval performance. Seven baseline models are selected for comparison, including BiGRU, vanilla BERT, BERT–BiGRU, BERT–BiGRU–CRF, T5, XLNet, and DeBERTa-v3-Base. Each model is trained, validated, and tested on the same dataset, with consistent data partitioning to ensure comparability.

2. Swarm Intelligence Performance Comparison Experiment

This experiment aims to verify whether the HHO_improved algorithm demonstrates superior efficiency and stability in optimizing CRF parameters compared to traditional optimization methods. Key parameters of the CRF model are optimized using HHO_improved and other traditional swarm intelligence algorithms, respectively. The convergence speed, final performance, and parameter stability of different optimization algorithms under the same number of training epochs and computational resources are recorded. Convergence curves and error stability curves are plotted to visually illustrate optimization efficiency and stability.

3. Ablation Study

The purpose of this experiment is to analyze the individual contributions and synergistic effects of each module (BERT, BiGRU, CRF, HHO_improved) on the overall performance of the model. Different modules of the deep learning model contribute differently to feature representation and sequence modeling capabilities. By systematically removing or replacing modules, their impact on NER and semantic retrieval performance can be quantified, thereby validating the rationality and necessity of the design. The full model with all four modules—BERT + BiGRU + CRF + HHO_improved—is constructed. Multiple variant models are formed by removing or replacing key modules, such as the following:

- Removing HHO_improved and using default CRF parameter optimization;
- Removing CRF and directly using BiGRU output for sequence labeling;
- Removing BiGRU and using only BERT for encoding;
- Removing BERT and using random embedding initialization.

Each variant model is trained under the same dataset and training strategy. The independent contributions and interactions of each module are analyzed to reveal the importance and synergistic effects of the components in the model design.

3.2. Datasets and Data Preprocessing

The dataset contains 13.331 logs, which are large in scale, representative, and can reflect the production and operation activities of enterprises in different business scenarios. The dataset is classified according to the content and attributes of the log, including normal and abnormal samples, domestic and foreign businesses, enterprise development strategies, market competition strategies, daily operation management, and other dimensions. The dataset is based mainly on the daily records of a single coal enterprise and covers the operations of the enterprise in different periods. The knowledge graph can bridge the semantic gap and improve the accuracy of model retrieval.

The proportion of normal samples is 57.96%, and the proportion of abnormal samples is 27.04%, which provides enough data for evaluating the accuracy of the model in anomaly detection tasks. In addition, 60.02% of the records in the dataset relate to domestic operations, and 39.98% relate to foreign operations, helping to test the model's adaptability to different business areas. In terms of content, the sample proportions of the enterprise development strategy, market competition strategy, and daily operation management are 44.38, 22.74, and 32.88, respectively, covering the key elements of enterprise operation and providing diverse samples for semantic analysis and relationship extraction.

In terms of time distribution, 37.18% of the data are from the previous year, 43.5% from 1–3 years, and 19.27% from more than 3 years of historical data, a time span that provides rich support for the model's long-term forecasting and trend analysis. Overall, the dataset structure design can be useful for covering enterprise production, the operation market strategy, and other dimensions of crux, so that it has a strong representation in the enterprise semantic retrieval task.

It is important to note that this study leverages a single, comprehensive proprietary dataset. This approach was necessitated by the highly specialized and sensitive nature of enterprise archive data, which often contains proprietary operational and strategic information. Consequently, publicly available datasets suitable for this specific vertical task are extremely scarce. We posit that the depth, diversity, and re-al-world representativeness of our chosen dataset, as detailed above, provide a robust foundation for validating our model. The primary contribution of this work lies in the novel model architecture and its optimization process, which is effectively demonstrated through rigorous evaluation on this substantial internal dataset.

Table 1 shows in the process of semantic retrieval and experiments of enterprise archives, the development and construction of enterprises are the focus.

Table 1. Dataset Statistics and Distribution.

Dataset Statistics	Count	Proportion (%)	Average Text Length (Characters)	Remarks
Total Logs	13,331	100	141	Total size of dataset, including daily operational records of coal companies
Category distribution				
normal sample	7725	57.96	135	Normal production operation status record
abnormal sample	3606	27.04	155	Logs of abnormal events, production failures, or unexpected events
Business area distribution				
Domestic business	7999	60.02	115	Mainly focus on the operation and production of the domestic market

Table 1. *Cont.*

Dataset Statistics	Count	Proportion (%)	Average Text Length (Characters)	Remarks
Overseas business	5332	39.98	125	Focus on foreign market expansion and related strategies
Content category distribution				
Enterprise development strategy related	5917	44.38	130	Topics including corporate development, expansion planning and long-term goals
Market competition strategy related.	3031	22.74	140	Covers information on market competition, customer strategy and product positioning
Daily operations and management-related time distribution	4383	32.88	100	Covers daily production processes, resource management and staffing
Last 1 year	4957	37.18	122	Data from the past year, suitable for observing current business trends
1 to 3 years of data	5805	43.55	118	Includes data from the past 1 to 3 years, suitable for long-term trend analysis
More than 3 years of data	2569	19.27	115	more than three years of data, useful for model generalization
Log type distribution				
Production records	6793	50.94	115	Production-related logs reflecting the operation of the production line
Running records	4706	35.32	118	Logs related to operations management, including resource and efficiency management
Marketing and Sales Records	1832	13.74	128	Covers content related to marketing, customer relations and sales
Text length distribution				
Short Text (& le;100 characters)	3478	26.08	80	Short records, mainly used to describe daily operations
Medium Text (100 200 characters)	7586	56.89	150	Common text lengths, including more detailed descriptions of events
Long text (& ge;200 characters)	2267	17.03	220	Long records, including detailed descriptions of complex events or problems
Keywords Frequency				
“Extension”	1259	9.44		Logs related to company development and business expansion
“Competition”	988	7.41		content covering market competition and strategic adjustment
“Management”	1731	12.99		Keywords related to management process and resource allocation
“Security”	2056	15.42		High-frequency items related to safety production and management
“Performance”	1397	10.48		Performance measurement and productivity improvement related

The logs are semi-structured text entries derived from enterprise business systems.

3.3. Environment of Software and Hardware

The hardware environment used for this experiment is a computer with an Intel Core i7 9700K CPU, 16 GB of memory, and an NVIDIA RTX 2080 Ti GPU. The proposed deep learning model uses Python 3.12.10 and a TensorFlow framework and is evaluated on a dataset containing thousands of enterprise profiles.

3.4. Evaluation Index

In this paper, the proposed deep learning model is compared with traditional retrieval methods based on keywords and rules through experiments. The 4 methods, which are based on indicators such as accuracy, precision, recall, and F1-score, were tested on thousands of company profiles. The experimental results show that our proposed deep learning model significantly improves retrieval accuracy and recall and is more effective than traditional and rule-based retrieval methods are.

Accuracy represents the percentage of semantic searches that match correctly and incorrectly across all searches. The precision rate indicates the proportion of true abnormal semantic data among all the matched abnormal semantic data in the matching process. The

recall rate represents the percentage of abnormal semantic data detected in the constructed dataset out of the total abnormal semantic data. The comprehensive evaluation index of the F1-score is based on the accuracy rate and recall rate. These indicators are calculated from (17) to (20) as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{17}$$

$$Precision = \frac{TP}{TP + FN}, \tag{18}$$

$$Recall = \frac{TP}{TP + FN}, \tag{19}$$

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{20}$$

3.5. Parameter Setting

Considering the large number of data samples in this experiment, the BERT model is used for pretraining in text processing, and the BERT parameter is set to 101,677,056. For model training, the BiGRU model is used for processing, and the parameter settings of the BiGRU are 2,332,296. The model dense layer parameters of the algorithm are determined by the num and hidden labels. The size is typically determined according to the amount of sample data processed. The detailed parameter settings of the model in this paper are shown in Table 2:

Table 2. Bert + BiGRU + CRF + Harris Hawk Algorithm Parameter Setting.

Parameter Name	Parameter Value
Enter ID	0
Pay attention to the mask.	0
Bert.	101,677,056
Bidirectional Cyclic Unit	2,332,296
Dropout	0
Intensive	Number of tags × Hidden size

3.6. Experiments Results

3.6.1. Benchmark Test Experiment Result

In this study, we compare the functions of five different models and three additional reference models in the semantic retrieval of enterprise archives. The five main models are the BERT basic model (Bert_Base_Chinese), BERT combined with the bidirectional gated cyclic unit model (Bert + BiGRU), BERT combined with the bidirectional gated cyclic unit and conditional random field model (Bert + BiGRU + CRF), BERT combined with the bidirectional gated cyclic unit, conditional random field and HHO_improved algorithm model (Bert + BiGRU + CRF + HHO_improved), and only the bidirectional gated cyclic unit model (BiGRU), with the XLNet, T5 and DeBERTa_v3_Base models as additional references, representing diverse and influential pre-trained language model approaches. as shown in Table 3.

Table 3. Algorithm comparison experiment.

Word Embedding	Model	Accuracy (%)	Precision (%)	Recall Rate (%)	F1-Score (%)
T5	T5	35.62	31.01	23.80	26.93
XLNet	XLNet	47.49	47.37	47.49	46.91
DeBERTa-v3-Base	DeBERTa-v3-Base	50.12	75.01	50.12	33.52
BiGRU	BiGRU	49.17	49.06	49.17	44.84
Bert	Base_Chinese	56.72	62.08	56.72	51.15
	BiGRU	77.81	77.91	77.81	77.80
	BiGRU + CRF	80.64	81.79	80.64	80.47
	BiGRU + CRF + HHO _improved	93.05	93.05	93.05	93.05

In this study, we conducted a series of controlled experiments on the task of semantic retrieval in enterprise archives to evaluate the impact of various pre-trained models and their hybrid configurations on retrieval performance. We began with the BERT base model (Bert_Base_Chinese) as our baseline, then incrementally augmented it with bidirectional gated recurrent units (BERT + BiGRU), conditional random fields (BERT + BiGRU + CRF), and finally with hyperparameter optimization via the HHO_improved algorithm (BERT + BiGRU + CRF + HHO_improved), in order to assess how sequential modeling and structured tagging enhance semantic representation. Concurrently, we trained a standalone BiGRU model as a lightweight comparator. Additionally, we included three state-of-the-art architectures—T5, XLNet, and DeBERTa-v3-Base—as auxiliary references, aiming to compare their generative capabilities, deep bidirectional encoding, and refined self-attention mechanisms in the context of archival semantic matching. All models were trained and evaluated on the same annotated corpus, employing accuracy, precision, recall, and F1-score as evaluation metrics, thereby systematically elucidating the strengths and limitations of each network architecture and optimization strategy in complex text retrieval scenarios.

3.6.2. Benchmark Test Experiment Result Discussion

The BERT + BiGRU + CRF + HHO_improved model achieves the highest accuracy, recall, and F1 value, making it the best model. The BERT and BiGRU models are relatively popular, showing high accuracy, recall, and F1 values and similar performance. The BERT + BiGRU, BERT + BiGRU + CRF + HHO_improved model contains BiGRU, CRF, and other technologies and achieves good results, as shown in Figure 6.

The BERT + BiGRU + CRF + HHO_improved algorithm inherits the advantages of BERT, the BiGRU and CRF model, and the HHO_improved algorithm. BERT provides powerful text representation capabilities, two-way gated loop units can learn sequence information and contextual language structures, and conditional random fields can capture relationships between labels. The HHO_improved algorithm can find the optimal parameter solution more efficiently. This multi-model combination improves the representation learning and sequence labeling capabilities of the BERT + BiGRU + CRF + HHO_improved algorithm. NER tasks require identifying specific entities that appear in the text and modeling and labeling the sequence relationships of different labels in the text. In the BERT + BiGRU + CRF + HHO_improved model, the CRF model can effectively model the annotation label as a recursive structure, fully consider the relationships between entities, and ensure consistency. In addition, under the optimized HHO_improved algorithm, the parameters of the conditional random field have a better optimization effect and improve the model's accuracy. The BERT model can adapt to different text data scenarios and achieve better representation through data preprocessing and model building. In addition, under

the optimized HHO_improved algorithm, the relative optimal solution can be quickly found in the search space, thus improving the model's adaptability to different data.

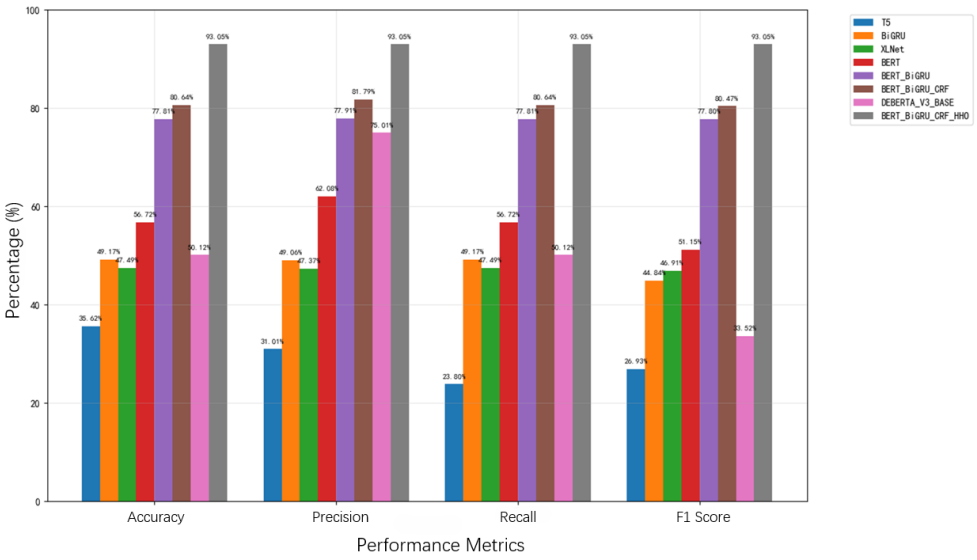

Figure 6. Comparison of experimental results.

BERT + BiGRU + CRF + HHO_improved performs well in named entity recognition tasks. By integrating various excellent technologies, sequence labeling problems can be effectively solved.

3.6.3. Swarm Intelligence Performance Comparison

The primary objective of this experiment is to validate the superiority of HHO_improve d over other swarm intelligence algorithms when applied to the BERT + BiGRU + CRF model. Specifically, we evaluate the comparative advantages of HHO_improved in terms of convergence speed and accuracy by conducting experiments under the same dataset and computational environment. Five swarm intelligence algorithms are considered in this study: PSO, WOA, GWO, HHO, and HHO_improved. The experimental data are identical to those used in Experiment 1, and the parameter configurations remain consistent. The results of the comparative analysis are presented as follows and Table 4:

Table 4. Swarm Intelligence Performance Comparison Experiment Result.

Swarm Intelligence Fitness	5	15	25	35
PSO	0.56	0.56	0.56	0.56
WOA	0.56	0.55	0.52	0.52
GWO	0.56	0.56	0.56	80.56
HHO	0.56	0.56	0.54	0.54
HHO_improved	0.56	0.56	0.51	0.51

3.6.4. Swarm Intelligence Performance Comparison Experiment Result Discussion

The experimental results demonstrate that the HHO_improved algorithm generally outperforms several traditional swarm intelligence methods in terms of optimization effectiveness and stability. Notably, the HHO_improved variant achieves a performance that is approximately 5% better than those of the other algorithms under comparison, particularly as the problem complexity increases. In contrast, while PSO maintains consistent results across different parameter levels, it shows no improvement, suggesting possible convergence stagnation. WOA exhibits a gradual decline in performance with increasing

complexity, and GWO displays a significant outlier under more challenging conditions, which may indicate instability. Overall, the HHO_improved not only achieves superior optimization accuracy but also demonstrates more consistent and reliable behavior, as further illustrated in Figure 7.

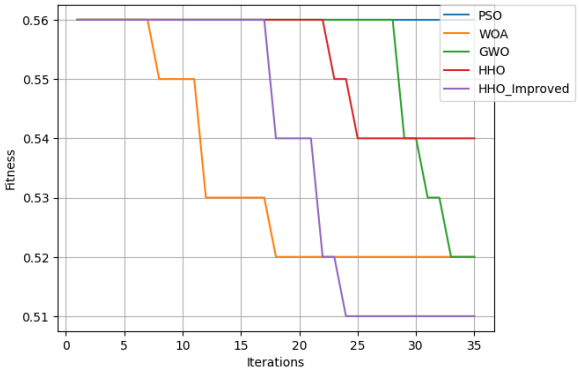

Figure 7. Comparison of fitness values of various swarm intelligence algorithms.

3.6.5. Ablation Experiment Result

To fully evaluate the independent contribution and synergy of each module in the model to the semantic retrieval performance, a series of ablation experiments were designed to analyze the impact on the overall model performance by gradually removing or replacing key modules. The complete model BERT + BiGRU + CRF + HHO_improved includes four core modules: BERT semantic embedding, BiGRU sequence modeling, CRF sequence annotation, the HHO_improved algorithm. Ablation experiments progressively remove individual modules and analyze the performance of each variant model.

The experiment is carried out on the enterprise archive semantic retrieval dataset, and the evaluation indicators include the accuracy rate, precision rate, recall rate, and F1-score. The specific model variants are as follows:

- Complete model (BERT + BiGRU + CRF + HHO_improved): All modules are included as benchmarks for performance comparison.
- CRF removal (BERT + BiGRU + HHO_improved): The CRF was replaced with Softmax to verify the role of the CRF in sequence tagging.
- Delete Harris Hawk (BERT + BiGRU + CRF): Delete the HHO_improved module and verify its contribution to parameter optimization.

The experimental results are shown in the Table 5:

Table 5. Ablation study of HARRIS HAWK and CRF validity test data.

Model Variants	Accuracy (%)	Precision (%)	Recall Rate (%)	F1-Score (%)
Complete model	93.05	93.05	93.05	93.05
Remove HHO_improved (BERT + BiGRU + CRF)	80.64	81.79	80.64	80.47
Remove CRF (BERT + BiGRU + HHO_improved)	76.45	88.34	78.09	80.56

3.6.6. Ablation Experiment Result Discussion

The experimental data and results are shown in Table 5. This finding shows that adding the CRF model to the BERT + BiGRU model significantly improves the accuracy of the performance index. In addition, after adding the HHO_improved module on this basis, the 4 performance indicators of the model are significantly improved, and the degree of improvement is different, indicating that the module has played a certain role in improving the performance of the model. The CRF module can use rich internal and contextual feature information in the annotation process and has excellent feature fusion ability.

The HHO_improved module provides a label smoothing strategy, which can effectively reduce the impact of label noise on model performance, thereby improving the robustness of the model.

1. Module independent contribution analysis

When the CRF module was removed, the F1-score decreased from 93.05% to 80.56%, a decrease of 12.49%. The CRF module improves the accuracy and context consistency of named entity recognition by modeling the dependencies between tags in sequence labeling tasks. The CRF has obvious advantages over simple Softmax layers, especially when dealing with semantic retrieval tasks with complex label distributions.

To remove the HHO_improved module, and the F1-score decreased to 80.47%, which was a decrease of 12.58%. Through the two-stage exploration and development mechanism, HHO_improved effectively optimizes the parameters of the CRF layer, enabling the model to adapt to dynamic and complex input contexts. In addition, HHO_improved has excellent search efficiency in high-dimensional optimization problems, which significantly improves the generalization ability of the model.

2. Module Synergy

Co-optimization of HHO_improved and CRF: The experimental results show that combining the HHO_improved and CRF modules is key in optimizing the sequence annotation task. HHO_improved dynamically optimizes the transfer matrix and weight parameters of the CRF so that the model performs better than the control model with fixed parameters in long-sequence tasks.

4. Conclusions

In this paper, we propose and validate an innovative enterprise archive semantic retrieval model that integrates BERT, BiGRU, CRF, and an HHO_improved algorithm. This hybrid architecture effectively addresses the limitations of traditional methods in complex semantic understanding and retrieval efficiency by synergistically combining deep semantic representation, sequential pattern learning, and intelligent parameter optimization.

The proposed model demonstrates significant performance improvements, achieving an F1-score of 93.05% and a precision of 93.05% on our test dataset, as detailed in Table 3. These results substantially outperform traditional retrieval methods and other deep learning baselines. The key to this success lies in the complementary roles of each component: BERT provides deep contextual embeddings, BiGRU captures bidirectional sequential dependencies, CRF ensures globally optimal label sequences, and the HHO_improved algorithm plays a critical role in enhancing model generalization through efficient hyperparameter optimization, as evidenced by its faster convergence speed and contribution to the 12.58% F1-score improvement observed in our ablation study.

While this study presents a robust framework for enterprise archive retrieval, we acknowledge its current limitation of being validated primarily on a single, albeit comprehensive, proprietary dataset. This limitation, however, directly motivates our immediate future work, which includes validating the model's performance on public benchmarks and extending its application to multilingual enterprise environments. The immediate next step stemming from this research is the development of a software prototype that integrates this validated algorithmic core into a practical intelligent archive management system.

This work establishes a solid foundation for next-generation enterprise knowledge management systems by demonstrating a quantifiably effective approach to semantic retrieval. The immediate next step is to develop a software prototype that encapsulates this validated algorithmic core into a deployable module for enterprise archive management systems. We believe our research opens promising avenues for developing more

adaptive, efficient, and intelligent archival management solutions in increasingly complex business environments.

Author Contributions: Conceptualization, X.S.; methodology, X.S. and J.C.; software, J.C. and Y.W.; validation, J.C. and Y.W.; formal analysis, X.S.; investigation, J.C.; resources, X.S.; data curation, X.S. and Y.W.; writing—original draft preparation, X.S. and Y.W.; writing—review and editing, X.S., J.C. and Y.W.; visualization, Y.W.; supervision, X.S. and J.C.; project administration, X.S.; funding acquisition, L.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during the current study are openly available at [https://github.com/Changer-He/BBCHCdata.git] (accessed on 10 October 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
BiGRU	Bidirectional Gated Recursive Unit
CRF	Conditional Random Field
HHO	Harris Hawks Optimizer
HHO_improved	Harris Hawks Optimizer Improved
CNN	Convolutional neural network
NLP	Natural Language Processing
GTE	General Text Embedding model
LLM	Large Language Model
MTEB	Massive Text Embedding Benchmark
AIR-Bench	Audio Instruction Benchmark
NER	Named Entity Recognition

References

1. Meng, T.; Hui, L. Application of Information Technology in Digital Archives Management. In Proceedings of the 2017 International Conference on Education and E-Learning, Bangkok, Thailand, 2–4 November 2017; pp. 112–116.
2. Yu, L.; Li, J. Semantic Retrieval—A New Direction of Intelligent Archive Retrieval. In Proceedings of the 2019 National Youth Archives Academic Forum, Chongqing, China, 13–16 November 2019; China Literature and History Publishing House: Beijing, China, 2019; pp. 154–162.
3. Ye, W. Research on Intelligent Archive Retrieval Technology Based on Semantic Analysis. *Off. Bus.* **2014**, *200*, 69–71.
4. Yin, B.; Wang, W.; Wang, L. A Review of Deep Learning Research. *J. Beijing Univ. Technol.* **2015**, *41*, 48–59.
5. Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Ahmad, J.; Fime, A.A. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access* **2024**, *12*, 26839–26874. [CrossRef]
6. Liu, J.; Liu, Y.; Luo, X. Progress in Deep Learning Research. *Appl. Res. Comput.* **2014**, *31*, 1921–1930, 1942.
7. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
8. Zhang, Y.; Wallace, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv* **2015**, arXiv:1510.03820.
9. Wang, Y.; Wu, Z.; Xie, J. A Review of Semantic Retrieval Systems for Scientific and Technological Literature. *New Technol. Libr. Inf. Serv.* **2015**, *258*, 1–7.
10. Lv, Y. Research on the Implementation Framework and Key Issues of Cross-media Semantic Retrieval of Digital Archive Resources. *Arch. Sci. Study* **2014**, *137*, 65–70.
11. Zhou, J. Research on ontology based local document analysis query extension method. *Sci. Technol. Commun.* **2011**, *36*, 47, 54.

12. Qi, B.; Cao, C.; Zheng, Y.; Li, Y. Research on Semantic Retrieval Methods for Domain Knowledge Documents. *Comput. Eng. Appl.* **2012**, *48*, 146–150.
13. Jin, B.; Guo, J.; Xu, X. Research on Optimizing Document Retrieval Using Domain Ontology: Design and Implementation Based on KIM Platform. *New Technol. Libr. Inf. Serv.* **2013**, *240*, 27–33.
14. Lv, Y. Research on the Implementation Strategy of Cross Media Semantic Association Aggregation of Digital Archive Resources. *Arch. Sci. Study* **2015**, *146*, 60–65.
15. Zhang, Y.; Ye, L.; Fang, B.; Li, C.; Wang, X. Similar case retrieval algorithm based on word frequency inverse document frequency and legal ontology. *Intell. Comput. Appl.* **2021**, *11*, 229–235.
16. Li, Z.; Zhang, X.; Zhang, Y.; Wu, D.; Sun, R. Towards general text embeddings with multi-stage contrastive learning. *arXiv* **2023**, arXiv:2308.03281. [CrossRef]
17. Lee, C.; Roy, R.; Xu, M.; Wang, Z.; De Sa, C. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv* **2024**, arXiv:2405.17428. [CrossRef]
18. Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; Wei, F. Improving text embeddings with large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand, 11–16 August 2024; pp. 1–15.
19. Luo, K.; Qin, M.; Liu, Z.; Xiao, S.; Zhao, J.; Liu, K. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024; pp. 101–115.
20. Dai, L.; Liu, H.; Xiong, H. Improve dense passage retrieval with entailment tuning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024; pp. 88–100.
21. Shi, P. Optimization of library resource retrieval based on vector space model. *Inf. Rec. Mater.* **2025**, *26*, 218–220. [CrossRef]
22. Huang, T. Research on Deep Medical Retrieval Model Based on Medical Knowledge Base Extension. Master's Thesis, Shandong University, Qingdao, China, 2019.
23. Huang, S.; Chen, X. Latent semantic analysis based on power iteration-randomized singular value decomposition. *J. Xiamen Univ.* **2023**, *62*, 679–686.
24. Arias Hernández, R.; Rockembach, M. Building trustworthy AI solutions: Integrating artificial intelligence literacy into records management and archival systems. *AI SOCIETY* **2025**, *40*, 4265–4282. [CrossRef]
25. Yi, Y. Research on innovative strategies for archive management in the "Internet+" era. *China Manag. Informationization* **2017**, *20*, 166–167.
26. Haki, K.; Safaei, D.; Magan, A.; Griffiths, M. Integrating Generative AI Into Enterprise Platforms: Insights From Salesforce. *Inf. Syst. J.* **2025**, *35*, 1497–1512. [CrossRef]
27. Tinn, R.; Cheng, H.; Gu, Y.; Usuyama, N.; Liu, X.; Naumann, T. Fine-tuning large neural language models for biomedical natural language processing. *Patterns* **2023**, *4*, 100729. [CrossRef] [PubMed]
28. Wei, F.; Li, C.; Xu, Q. Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 2786–2792. [CrossRef]
29. Fatemi, S.; Hu, Y.; Mousavi, M. A Comparative Analysis of Instruction Fine-Tuning LLMs for Financial Text Classification. *arXiv* **2024**, arXiv:2411.02476. [CrossRef]
30. Heidari, A.A.; Mirjalili, S.; Faris, H.; Aljarah, I.; Mafarja, M.; Chen, H. Harris hawks optimization: Algorithm and applications. *Future Gener. Comput. Syst.* **2019**, *97*, 849–872. [CrossRef]