Research on Intelligent Text Recognition of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era Based on the BERT Semantic Matching Model

Yanxin Geng
School of Marxism, Jilin Business and Technology College, Changchun 130507, Jilin, China
Corresponding author: Yanxin Geng (e-mail: 13756911306@163.com)

ABSTRACT In the digital age, texts related to Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era have experienced rapid growth, creating higher requirements for accurate semantic identification and intelligent text processing. Traditional manual identification methods face bottlenecks such as low efficiency, high subjectivity, and limited coverage, which restrict precise text recognition in ideological dissemination, policy implementation, and technical document management in advanced engineering sectors. This issue is also relevant to standardized documentation in fields such as electromagnetic engineering, antenna systems, and communication technology, where semantic consistency and terminology accuracy are required. This paper aims to improve the semantic accuracy and intelligent efficiency of ideological text recognition by introducing the BERT semantic matching model into the processing of texts on Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era. First, through literature review, the semantic characteristics of ideological texts, including conceptual systematization, multi-level expression, and contextual interdependence, are clarified, together with the limitations of traditional recognition methods. Second, an ideological text corpus is constructed, including 12,000 annotated samples across three text types: core literature, policy documents, and interpretive articles. On this basis, a four-stage recognition process consisting of preprocessing, semantic encoding, similarity calculation, and result optimization is designed to support accurate semantic matching and intelligent text recognition.

INDEX TERMS bert model, semantic matching, intelligent text recognition, electromagnetic engineering, technical documentation

I. INTRODUCTION

A. RESEARCH BACKGROUND

WITH the continuous dissemination and deepening development of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, its textual carriers have expanded to diverse forms, from traditional documents to online articles and interpretation videos, a breadth necessary to ensure clear communication and implementation across all sectors. According to data from the 2024 China Statistical Yearbook and the Text Statistical Report published by the Communist Party Member Network in October 2025, by 2025, the number of publicly accessible texts related to Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era exceeded 500,000. These include core documents like *Selected Readings from Xi Jinping's Works*, as well as derivative texts such as local government policy interpretations, academic research papers, and media publicity reports. These texts exhibit characteristics of "concept density, semantic rigor, and expressive diversity." The statistical scope of the Communist Party Member Network report aligns closely with the research requirements, covering core literature, policy documents, and authoritative interpretations. For instance, the term "Chinese modernization" may be precisely defined in core literature, described as "a modernization path suited to China's national conditions" in local practice reports, or deconstructed into sub-dimensions such as "modernization for a massive population" and "modernization for common prosperity" in academic papers [1-3].

B. RESEARCH QUESTIONS

1. How to construct a BERT semantic matching model that aligns with the characteristics of ideological texts?

This requires addressing challenges in capturing the model's semantic understanding, such as the "systematic core concepts" (e.g., sub-concepts within the "Five-in-One" overall layout), "multi-level expressions" (e.g., differing formulations of "common prosperity" in core documents versus local practices), and "contextual relevance" (e.g., semantic variations of "comprehensive and strict governance over the Party" in Party building versus anti-corruption contexts).

2. How to establish a "high-quality, comprehensive" corpus of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era?

This requires addressing issues such as "authoritative selection" (distinguishing core documents from nonauthoritative interpretations), "precise annotation" (marking core concepts, semantic relationships, and context types), and "dynamic updating" (incorporating the latest policy formulations).

3. How to design a "layered semantic matching strategy" to accommodate different ideological text recognition needs?

Differentiated matching approaches must be proposed for three scenarios: "precise core concept identification" (e.g., matching the original text of the "Ten Clarifications"), "generalized recognition of derivative expressions" (e.g., local applications of the "New Development Philosophy"), and "ambiguity resolution" (e.g., distinguishing between "comprehensively deepening reform" and "comprehensively advancing law-based governance").

II. THEORETICAL FOUNDATIONS

A. SEMANTIC CHARACTERISTICS OF XI JINPING THOUGHT ON SOCIALISM WITH CHINESE CHARACTERISTICS FOR A NEW ERA TEXTS

As a systematic political theory corpus, the semantic structure of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era exhibits a "three-layer, four-dimensional" characteristic. This core distinction from ordinary texts serves as the fundamental basis for designing semantic matching models [4].

1) Three-Layer Semantic Structure

1. Core Concept Layer: Serves as the "semantic anchor" of the ideological text, encompassing fundamental expressions such as the "Ten Clarifications," "Fourteen Adherences," "Thirteen Major Achievements," and "Six Must-Holds." These expressions exhibit "fixed semantics and rigorous formulation"; for example, "It is clearly stated that the most essential characteristic of socialism with Chinese characteristics is the leadership of the Communist Party of China." Their core meaning remains consistent across all authoritative texts, with minimal variation in phrasing, forming the foundation for identifying the ideological text [5].

2. Derivative Expression Layer: Represents the application and extension of core concepts across different contexts, including policy interpretations, local practices, and academic explanations. These expressions exhibit "semantic association and flexible formulation"; for instance, the core concept "common prosperity" may be articulated in local practice texts as "narrowing the income gap between urban and rural residents to achieve the goal of common prosperity," while in academic texts it might be phrased as "common prosperity is an essential requirement of Chinese-style modernization, embodying a people-centered development philosophy." Though the expressions differ, their meanings all derive from the core concept [6-8].

3. Contextual Explanation Layer: This layer provides supplementary explanations regarding the background, conditions, and scope of core concepts and their derivative expressions, characterized by "semantic support and context dependency." For example, "comprehensively deepening reform" in an "economic domain" context may involve expressions like "the market plays a decisive role in resource allocation"; in a "social domain" context, it may involve expressions like "education and healthcare system reform." Contextual differences lead to variations in emphasis, but the core semantics remain consistent with "comprehensively deepening reform" [9].

2) Four-Dimensional Semantic Attributes

1. Conceptual Interconnectedness: Concepts within ideological texts do not exist in isolation but form a semantic network structured as "core - derivative - associated." For example, the "new development philosophy" (innovation, coordination, green development, openness, and sharing) has direct semantic connections with "high-quality development" and "Chinese-style modernization." "High-quality development" represents the practical goal of the "new development philosophy," while "Chinese-style modernization" serves as its application scenario. Such relationships must be prioritized in semantic matching by the model [10].

2. Context Dependency: The same concept may emphasize different aspects in varying contexts while retaining its core meaning. For instance, "comprehensive and rigorous Party governance" emphasizes "political, ideological, and organizational development" within the context of "Party building," whereas it stresses "deterrence, prevention, and resistance to corruption" in the context of "anti-corruption." Models must integrate contextual information to determine whether a given expression belongs to the ideological domain [11,12].

B. CORE PRINCIPLES OF THE BERT SEMANTIC MATCHING MODEL

The BERT model employs a bidirectional encoder based on the Transformer architecture, achieving deep semantic capture through a "pre-training - fine-tuning" approach. Its core principles encompass three major modules: "bidirectional context embedding," "masked language modeling (MLM)," and "sentence pair prediction (NSP)," providing the technical foundation for semantic matching in ideological texts.

1) Bidirectional Contextual Embedding

Traditional language models (e.g., RNN, LSTM) employ unidirectional contextual encoding (left-to-right or right-to-left), failing to simultaneously utilize semantic information from both sides. This results in inadequate understanding of "ambiguous vocabulary" and "complex sentence structures." The BERT model adopts the Transformer encoder architecture, employing a "multi-head self-attention mechanism" to simultaneously consider words at all positions within the input text. This achieves bidirectional context embedding; for example, for the term 'comprehensive' in "comprehensively and strictly governing the Party," the model can integrate semantic information from both the preceding "adhere to" and the following "govern the Party," accurately grasping its meaning as "encompassing all domains, aspects, and stages of Party building" rather than merely interpreting it as the literal "broad scope" [13]. In ideological text processing, bidirectional contextual embedding effectively captures "logical relationships between concepts." For instance, within the "Five-Sphere Integrated Plan," the parallel relationship between "economic development," "political development," and "cultural development," as well as the causal relationship between "ecological civilization development" and "green development," all require bidirectional context for complete comprehension. The WordPiece tokenizer in BERT prioritizes recognizing high-frequency, complete lexical units. The jieba preprocessing ensures these core concepts appear as such units in the input string, allowing the model to maintain their semantic integrity during its internal tokenization.

2) Masked Language Modeling (MLM)

MLM is one of BERT's core pre-training tasks. Its principle involves randomly masking 15% of words in the input text (e.g., masking "essential requirement" in "Common prosperity is an essential requirement of Chinese-style modernization" as "[MASK]"), then instructing the model to predict the masked words. Through this task, the model learns semantic associations between words in contextual settings. Examples include fixed collocations like "common prosperity" with "essential requirement" and "Chinese-style modernization," as well as semantic correspondences such as "comprehensive rule of law" with "building a socialist country governed by law" [14-16]. For "core concept fixed expressions" in ideological texts, the MLM task enhances the model's ability to recognize "conceptual integrity." For instance, given the input text "The most essential feature of socialism with Chinese characteristics is [MASK] leadership," the model can accurately predict "the Communist Party of China" rather than other terms based on pre-trained semantic associations. This is crucial for precise identification of core concepts [17].

3) Sentence Pair Prediction (NSP)

In identifying derivative expressions within ideological texts, training outcomes from the NSP task enable models to determine whether candidate texts are semantically related to core concepts. For instance, given sentence C: "A certain region promotes high-quality economic development through industrial upgrading," and sentence D: "The new development philosophy encompasses innovation, coordination, green development, openness, and sharing." The model can leverage semantic associations learned through the NSP task to determine that Sentence C is semantically related to "new development philosophy" (Sentence D), thereby identifying it as an ideological derivative expression [18].

III. CONSTRUCTION OF IDEOLOGICAL TEXT CORPORA AND MODEL DESIGN

A. CONSTRUCTION OF THE XI JINPING THOUGHT ON SOCIALISM WITH CHINESE CHARACTERISTICS FOR A NEW ERA TEXT CORPUS (XJTS-CORPUS)

The corpus serves as the foundation for BERT model training and testing. The construction of XJTS-Corpus adheres to three principles: "authoritativeness, comprehensiveness, and precision." It encompasses three categories of texts, "core literature, policy documents, and authoritative interpretations," totaling 12,120 annotated samples. The specific construction process is as follows [19].

1) Selection of Corpus Sources

1. Core Literature (120 entries): Sourced from authoritative publishers such as the Central Party Literature Publishing House and People's Publishing House. Includes *Selected Works of Xi Jinping*, *Compilation of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era*, the 19th CPC National Congress Report, and the 20th CPC National Congress Report. These texts represent the "source" of the ideology, featuring the most rigorous and standardized semantics, serving as the "core semantic anchors" of the corpus [20].

2. Policy Documents (5,000 entries): Sourced from the Chinese Government website, the General Office of the State Council, and official websites of various ministries and commissions. These include the 14th Five-Year Plan, rural revitalization strategy documents, and policies on the "dual carbon" goals. Such texts represent the "practical extension" of the ideology at the policy level, exhibiting high semantic consistency with core literature and a degree of normative expression [21].

3. Authoritative Interpretations (7,000 entries): Sourced from authoritative media outlets including the Communist Party of China News Network, People's Daily Online, Xinhua News Agency, People's Daily, and Qiushi Journal. This category encompasses expert commentary articles, reports

on local implementation practices, and academic research papers (published in core journals). These texts represent the "derivative expansion" of the ideology at the interpretation and research level, featuring flexible and diverse expressions. While semantically related to the core literature, they require further matching and verification [22].

A "dual verification mechanism" was applied to these 7,000 authoritative interpretation texts:

(1) Semantic consistency verification using the XJTS-BERT model retained only texts with similarity ≥ 0.65 to core literature, resulting in 6,880 texts;

(2) Expert secondary review by three Marxist theory experts confirmed ideological consistency, achieving a 99.2% approval rate.

To ensure corpus authority, a "three-tier screening mechanism" was established:

Tier 1 Screening (Source Verification) – retained only texts from official channels and core media;

Tier 2 Screening (Content Verification) – excluded texts unrelated to the ideology (e.g., general news reports);

Level 3 Screening (Expert Review) – five Marxist theory experts conducted final review of screened texts to ensure high relevance to Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, achieving a Kappa consistency coefficient of 0.92 [23-25].

2) Corpus Annotation Scheme

Employing a "manual annotation + machine-assisted verification" approach, each corpus item was annotated with three types of tags: "core concept tags," "semantic association tags," and "context tags." The annotation tool used was Brat (a web-based text annotation tool).

Annotation Scheme for the Ideological Text Corpus of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era. As shown in Table 1:

After annotation completion, "cross-validation" and "machine-assisted verification" ensure accuracy: 10% of the corpus is independently annotated by two annotators, calculating annotation consistency (Kappa coefficient of 0.91 for core concept labels, 0.88 for semantic association labels, and 0.93 for contextual labels); a universal BERT model calculates semantic similarity between annotated text and core concepts. If similarity < 0.5 but labeled as "equivalent" or "derived," re-review occurs. Final annotation accuracy reached 98.5%.

3) Corpus Preprocessing and Segmentation

Cleaning: Removes special symbols (e.g., "", "@"), redundant spaces, page numbers, and other irrelevant information;
Word Segmentation: This step uses the jieba segmentation tool combined with a "Core Ideological Vocabulary List" solely for preprocessing to protect the integrity of core concepts (e.g., ensuring "Chinese-style modernization" remains intact). The resulting segmented text is then converted back into a continuous string for input to the BERT model, which performs its native WordPiece subword tokenization. This dual-layer approach ensures that high-frequency, complete core concepts are recognized as coherent semantic units by the model.
Stopword Removal: Eliminate function words like "的," "了," and "在," while retaining "core concept words" and "semantically related words" (e.g., "common prosperity," "high-quality development").

B. DESIGN OF THE IDEOLOGICAL TEXT-ADAPTED BERT SEMANTIC MATCHING MODEL

Based on the general BERT model (bert-base-chinese), this model integrates the "three-layer semantic structure" and "four-dimensional semantic attributes" of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era. Optimization occurs across three dimensions, "pre-training fine-tuning, semantic encoding enhancement, and matching strategy design," to construct the "Ideological Text-Adapted BERT Model (XJTS-BERT)."

1) Pre-training Fine-tuning: Enhancing Semantic Capture of Ideological Texts

The pre-training data for the general BERT model primarily consists of general texts (e.g., news, fiction), which inadequately captures the core concepts and semantic associations of ideological texts. Therefore, "secondary pre-training" is required for fine-tuning:

1. Pre-training Data: Utilize core literature (120 entries) and policy documents (5,000 entries) from the XJTS Corpus to construct an "Ideological Text Pre-training Dataset," totaling approximately 10 million characters;

2. Pre-training Tasks: Retain MLM and NSP tasks while adjusting parameters for ideological text characteristics:

MLM Task: Increase masking ratio for core concept vocabulary (from 15% to 25%). This boosts masking probability for key terms like "Chinese-style modernization" and "common prosperity," enhancing the model's semantic memory of core concepts; To validate this adjustment, an ablation study was conducted with five masking ratios (15%, 18%, 21%, 25%, 28%) under identical training conditions. Results showed that the 25% ratio yielded optimal performance: core concept recognition accuracy peaked at 92.8% (an 18% improvement over 15%), semantic association recognition accuracy reached 87.2% (a 15% improvement), and validation loss remained stable at 0.23, indicating no overfitting. A ratio of 28% led to increased validation loss (0.31) and decreased accuracy, confirming an overfitting trend. Furthermore, a "concept variation expression recognition test" demonstrated that the model trained with a 25% mask rate achieved 88.3% accuracy in identifying variant expressions (e.g., "modernization road that suits China's national conditions" for "Chinese-style modernization"), a 12.7% improvement over the 15% rate, proving enhanced semantic understanding beyond mere memorization.

TABLE 1. Annotation Scheme for the Ideological Text Corpus of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era

Label Type	Label Content	Example	Label Description
Core Concept Label	Specific categories corresponding to core expressions like "Ten Clarifications" and "Fourteen Persistences"	If the text involves "upholding the Party's leadership over all work," label it as "Fourteen Persistences - Upholding the Party's leadership over all work"	1. Core documents need to label all involved core concepts; 2. Policy documents and interpretation texts label 1-2 most relevant core concepts; 3. Non-core ideological texts label "None".
Semantic Association Label	Semantic association types between the text and core concepts (Equivalent, Derived, Associated)	If the text is a local practice case of "common prosperity," label it as "Common Prosperity - Derived"	1. Equivalent: The text expression is completely consistent with the core concept (e.g., original core documents); 2. Derived: The text is an application or extension of the core concept (e.g., policy documents, practice cases); 3. Associated: The text has a semantic connection with the core concept but is not directly derived (e.g., relevant analysis in academic research).
Context Label	Release subject, application scenario, and text type of the text	If the text is a rural revitalization policy issued by a provincial government, label it as "Release Subject: Provincial Government; Application Scenario: Rural Revitalization; Text Type: Policy Document"	1. Release Subject: Divided into "Central Government, Local Government, Authoritative Media, Academic Institution"; 2. Application Scenario: Divided into "Economic Construction, Political Construction, Cultural Construction, Social Construction, Ecological Civilization Construction, Party Building"; 3. Text Type: Divided into "Core Document, Policy Document, Interpretation Article, Academic Paper".

NSP Task: Input sentence pairs restructured as "core concept sentence + derivative expression sentence" (e.g., "Chinese-style modernization is modernization on a massive scale" + "a region advancing modernization practices on a massive scale"), enabling the model to learn semantic associations between core concepts and their derivative expressions;

3. Training Parameters: Employed the Adam optimizer with a learning rate of 2e-5, batch size of 32, and 5 training epochs to prevent overfitting (monitored via validation set loss; training halted when validation loss failed to decrease for two consecutive epochs). Through secondary pre-training, the model achieved an 18% improvement in core concept recognition accuracy within ideological texts and a 15% increase in semantic association recognition accuracy for derivative expressions, laying the foundation for subsequent semantic matching [26]. An adaptation verification experiment compared three methods: "jieba-only + model," "model-only subword," and the proposed "dual-layer segmentation." The dual-layer method achieved 92.8% core concept recognition accuracy, outperforming the model-only method (82.5%) by 10.3%, confirming its effectiveness.

2) Semantic Encoding Optimization: Incorporating Contextual and Hierarchical Information

The semantics of ideological texts depend on "context" and "hierarchy." The semantic encoding of the general BERT model only considers text content and does not incorporate this type of information. The following optimizations are required:

1. Contextual Information Embedding: During text encoding, convert "context labels" (issuing authority, application scenario, text type) into "context vectors." Concatenate these with the text's word vectors before inputting into the model. For example: "Issuing Authority: Central Government" → vector [1,0,0,0] (Central = 1, Local Government = 0, Authoritative Media = 0, Academic Institution = 0). "Application Scenario: Economic Development" is converted to vector [1,0,0,0,0,0] (Economic Development = 1, Other Scenarios = 0). A linear layer maps the context vector to a dimension matching the word vector (768-dimensional), which is then added to the word vector to achieve joint encoding of "content + context" [27].

2. Multi-level Attention Mechanism: For the "Core Concept Layer - Derived Expression Layer - Contextual Explanation Layer" of ideological texts, "Hierarchical Attention Weights" are designed:

Core Concept Layer: Attention weight set to 0.5, ensuring the model prioritizes core semantics;

Derived Expression Layer: Attention weight set to 0.3, focusing on semantically related information;

Contextual Explanation Layer: Attention weight set to 0.2, focusing on supplementary semantic information.

Through weight adjustments in the self-attention mechanism, the model allocates varying degrees of attention to semantic information across different layers during encoding. For instance, when identifying derivative expressions of "common prosperity," the model prioritizes the core term "common prosperity" (0.5 weight), followed by expressions like "local practices" (0.3 weight), and finally contextual information such as "issuing entities" (0.2 weight) [28-30]. Following semantic encoding optimization, the model achieved a 12% improvement in recognition accuracy for "context-dependent expressions" (e.g., "comprehensively deepening reform" in diverse scenarios) and a 10% enhancement in semantic capture completeness for "multi-level expressions" (e.g., long texts containing core concepts and derivative explanations).

3) Design of Hierarchical Semantic Matching Strategy

To address the three core scenarios in ideological text recognition (precise core concept identification, generalized recognition of derivative expressions, and resolution of ambiguous expressions), differentiated semantic matching strategies were designed to ensure accurate recognition across all contexts.

IV. EXPERIMENTAL VALIDATION AND RESULTS ANALYSIS

A. EXPERIMENTAL DESIGN

To validate the effectiveness of the XJTS-BERT model in intelligent ideological text recognition, two types of experiments were designed: "comparative experiments" and "scenario tests." Comparative experiments validated the model's superiority by contrasting its performance with traditional methods and generic BERT models. Scenario tests verified the model's practicality through real-world application scenarios. The hierarchical semantic matching strategies for different ideological text recognition scenarios are shown in Table 2.

B. EXPERIMENTAL DATA AND ENVIRONMENT

1. Experimental Data: The test set from XJTS-Corpus (1,212 entries) was used, comprising 12 core documents, 500 policy documents, 600 interpretive articles, and 100 academic papers. Among these, 1,000 entries were ideology-related texts and 212 were non-ideology-related texts, ensuring balanced data distribution.

2. Comparison Methods:

Traditional Method 1: TF-IDF + Cosine Similarity (semantic matching based on word frequency, unable to capture deep semantics);

Traditional Method 2: Word2Vec + K-Means (clustering matching based on static word vectors, unable to handle contextual dependencies);

General BERT Model: bert-base-chinese model without pre-training on ideological texts (only performs basic semantic encoding);

3. Experimental Environment: Hardware: Intel Xeon Gold 6248 CPU, NVIDIA Tesla V100 GPU (32GB); Software: Python 3.8, PyTorch 1.10, Transformers 4.18, Scikit-learn 1.0.

C. EXPERIMENTAL TASKS AND EVALUATION METRICS

1. Experimental Tasks:

Task 1: Core Concept Identification – Determine whether test-set texts contain core concepts such as "Ten Clear-Cut Principles" and "Fourteen Perseverances," comprising 300 test samples (12 core documents, 188 policy documents, 100 interpretive articles);

Task 2: Derivative Expression Identification – Determine whether test set texts are derivative expressions of core ideological concepts, comprising 600 test samples (300 policy documents, 200 interpretive articles, 100 academic papers);

Task 3: Ambiguity Resolution – Precisely distinguish easily confused conceptual expressions (e.g., "comprehensive and rigorous Party governance," "comprehensive rule of law," "comprehensive deepening of reforms"). Total of 312 test samples (12 policy documents, 300 interpretive articles);

2. Evaluation Metrics: Employed four key metrics – "Precision (P)," "Recall (R)," "F1 Score (F1)," and "Semantic Consistency (SC)." Semantic consistency is calculated as the mean cosine similarity between candidate text and core concept sentence vectors.

D. COMPARATIVE EXPERIMENTAL RESULTS AND ANALYSIS

1) Task 1: Core Concept Identification Results

Core concept identification requires "exact matching." The performance comparison of different methods in the core concept recognition task is shown in Table 3.

2) Task 2: Derived Expression Recognition Results

Derived expression recognition requires achieving "generalized matching." The performance comparison of different methods in the derivative expression recognition task is shown in Table 4.

Result Analysis:

1. Weak Generalization Capability of Traditional Methods: TF-IDF fails to recognize synonymous but differently phrased expressions (e.g., "common prosperity" and "promoting common prosperity in urban and rural areas"), resulting in low recall rates. Word2Vec clustering is significantly influenced by phrasing variations (e.g., "common prosperity practices" and "common prosperity cases" grouped into separate clusters), leading to low precision rates.

2. Limited General BERT Model Generalization: While capturing partial semantic associations, it fails to integrate contextual information. Recognition accuracy for "derivative expressions across different scenarios" (e.g., "high-quality development" in economic vs. social contexts) is insufficient, with semantic consistency at only 0.72.

3) Task 3: Ambiguity Resolution Results

Ambiguity resolution requires achieving "precise differentiation." Experimental results are shown in Table 5 below (using three categories of ambiguous expressions: "comprehensive and rigorous Party governance," "comprehensive rule of law," and "comprehensive deepening of reforms").

Result Analysis:

1. Traditional methods exhibit poor ambiguity resolution: TF-IDF and Word2Vec rely solely on surface-level lexical features, failing to distinguish core attribute differences (e.g., subject, objective) between phrases like "comprehensive and rigorous Party governance," "comprehensive rule of law," and "comprehensive deepening of reforms." Ambiguity resolution accuracy falls below 60%, e.g., misclassifying "a locality advancing judicial system reform" as "comprehensive deepening of reforms" (which is actually a derivative expression of "comprehensive rule of law").

2. General BERT model shows moderate ambiguity resolution capability: While it captures some attribute differences through context, it lacks attribute comparison optimization.

TABLE 2. Hierarchical Semantic Matching Strategies for Different Ideological Text Recognition Scenarios

Recognition Scenario	Matching Strategy	Specific Implementation Steps	Applicable Text Types
Core Concept Accurate Recognition	Keyword Anchoring + Context Cosine Similarity Calculation	1. Extract core concept keywords from the candidate text (e.g., "Ten Clarifications," "Chinese-style Modernization"); 2. Retrieve the core definition text corresponding to the keyword from the XJTS-Corpus; 3. Encode the candidate text and core definition text separately using the XJTS-BERT model to obtain vectors; 4. Calculate the cosine similarity of the two vectors: if similarity ≥ 0.85, mark as "core concept matched".	Core documents, policy documents (texts requiring precise core concept matching).
Derived Expression Generalization Recognition	Sentence Vector Clustering + Dynamic Threshold Adjustment	1. Encode derived expression texts in the training set using the XJTS-BERT model to obtain sentence vectors; 2. Use K-Means clustering to group semantically similar sentence vectors into clusters (e.g., "common prosperity" derived expression cluster, "new development concept" derived expression cluster); 3. Calculate the distance between the candidate text's sentence vector and each cluster center: if distance $\leq$ dynamic threshold (adjusted by text type: 0.7 for policy documents, 0.65 for interpretation articles, 0.6 for academic papers), mark as "derived expression matched".	Policy documents, interpretation articles, academic papers (texts with flexible expressions).
Ambiguous Expression Disambiguation	Concept Attribute Comparison + Semantic Difference Calculation	1. Extract the core attributes (subject, object, target, scope) of ambiguous expressions (e.g., attributes of "comprehensively governing the Party with strict discipline": Subject: Communist Party of China; Object: Party building; Target: Maintain the Party's advanced nature; Scope: All fields of the Party); 2. Encode the candidate text and attribute descriptions of each ambiguous expression using the XJTS-BERT model to obtain attribute vectors; 3. Calculate the attribute vector similarity between the candidate text and each ambiguous expression, select the expression with the highest similarity (≥ 0.7) as the matching result.	Easily confused concept expressions (e.g., "comprehensively deepening reform" vs. "comprehensively governing the country according to law").

TABLE 3. Performance Comparison of Different Methods in Core Concept Identification Task

Method	P	R	F1	SC
TF-IDF + Cosine Similarity	78.3%	65.7%	71.5%	0.62
Word2Vec + K-Means	80.5%	70.2%	75.0%	0.65
General BERT Model	88.2%	82.5%	85.3%	0.78
XJTS-BERT Model	95.6%	92.8%	94.2%	0.91

TABLE 4. Performance Comparison of Different Methods in Derivative Expression Recognition Task

Method	P	R	F1	SC
TF-IDF + Cosine Similarity	65.2%	58.3%	61.6%	0.55
Word2Vec + K-Means	70.1%	65.8%	67.9%	0.58
General BERT Model	82.3%	78.5%	80.4%	0.72
XJTS-BERT Model	90.5%	87.2%	88.8%	0.83

TABLE 5. Performance Comparison of Different Methods in Ambiguous Expression Disambiguation Task

Method	Average P	Average R	Average F1	Accuracy
TF-IDF + Cosine Similarity	58.7%	52.3%	55.3%	48.5%
Word2Vec + K-Means	62.1%	58.9%	60.5%	55.2%
General BERT Model	78.5%	75.2%	76.8%	72.3%
XJTS-BERT Model	92.3%	89.7%	91.0%	90.5%

It struggles to distinguish "ambiguous expressions with similar attributes" (e.g., "comprehensive deepening of reforms" and "comprehensive rule of law" both involve "institutional development"), achieving only 72.3% accuracy in ambiguity resolution.

3. XJTS-BERT model demonstrates strong ambiguity resolution: Through "conceptual attribute comparison + semantic difference calculation," the model precisely distinguishes ambiguous expressions; for instance, given the candidate text "A certain province strengthens the construction of intra-Party regulations and systems," the model extracts its attributes and calculates similarities of 0.92 with "comprehensive and rigorous Party governance," 0.35 with "comprehensive rule of law," and 0.28 with "comprehensive deepening of reforms."

E. SCENARIO TEST RESULTS AND ANALYSIS

To validate the model's practicality in real-world applications, three typical scenarios were selected for testing: "Intelligent Ideological Text Retrieval," "Academic Paper Topic Identification," and "Policy Text Consistency Verification." Each scenario utilized 100 actual texts. Test results are as follows.

1) Scenario 1: Intelligent Ideological Text Retrieval

Test Task: Users input the search keyword "common prosperity." The model retrieves relevant texts from 1,000 mixed texts (500 common prosperity-related texts and 500 unrelated texts), evaluating retrieval accuracy and efficiency.

Test Results:

- Retrieval Accuracy: XJTS-BERT model achieved 93.2% precision, 90.5% recall, and an F1 score of 91.8%; traditional TF-IDF method yielded 75.8% precision, 68.3% recall, and an F1 score of 71.9%.

- Retrieval Efficiency: XJTS-BERT model took 0.3 seconds per text retrieval, while the traditional TF-IDF method took 0.1 seconds. Although the model's retrieval time was slightly longer, it demonstrated significantly improved accuracy and supported "semantic association retrieval" (e.g., searching for "common prosperity" could simultaneously retrieve related texts such as "narrowing urban-rural income gaps" and "rural revitalization for prosperity").

Scenario Value: Addresses "missed hits and false positives" in traditional keyword retrieval, enabling rapid access to ideological texts. Researchers can swiftly locate local implementation cases and academic studies on "common prosperity," boosting research efficiency.

2) Scenario 2: Academic Paper Topic Identification

Test Task: Determine whether 100 academic papers (including 60 papers related to Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era and 40 unrelated papers) focus on ideological research, and label their specific research directions (e.g., "Chinese-style modernization," "common prosperity").

Test Results:

- Topic identification accuracy: XJTS-BERT model achieved 92.0% accuracy, with 95.0% accuracy for core concept research papers and 88.3% for derivative research papers; the general BERT model achieved 80.0% accuracy, with 85.0% for core concept research papers and 73.3% for derivative research papers.

- Research Direction Labeling Accuracy: XJTS-BERT model achieved 89.0% accuracy, accurately labeling specific directions like "Chinese-style modernization," "common prosperity," and "new development philosophy"; the general BERT model achieved 75.0% accuracy, often confusing research directions between "new development philosophy" and "high-quality development."

V. CONCLUSIONS

The model's practicality has been validated in diverse real-world scenarios, demonstrating its robust utility across various high-stakes industrial applications, including relevant operational contexts. Through testing in three application areas, "intelligent retrieval of ideological texts," "academic paper topic identification," and "policy text consistency verification," the model demonstrated high accuracy and strong applicability. It provides intelligent tools for ideological dissemination, academic research, and policy implementation, advancing the transformation of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era from "manual-dominated" to "intelligence-driven" text processing.

The corpus coverage requires expansion: The current corpus primarily consists of Chinese texts, lacking multilingual materials (such as foreign-language translations of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era and international research papers), which limits its ability to support cross-lingual ideological text recognition. Additionally, the proportion of "social media texts" (e.g., Weibo and WeChat Official Account articles) in the corpus is insufficient, making it difficult to adapt to these "short-form, colloquial" vehicles for ideological dissemination.

AUTHOR CONTRIBUTIONS

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

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, Jilin Provincial Social Science Fund Project (Special Project of Marxist Engineering) "Research on the Historical Context and Advancement Path of Chinese Modernization" Project No.: 2024M11.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese and Triplet Network Designs," arXiv preprint arXiv:1908.10084, 2019.
[3] Research Office of the Central Committee of the Communist Party of China, *Selected Excerpts on Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era*. Beijing, China: People's Publishing House, 2023.
[4] Communist Party of China News Network, *Statistical Report on Texts Related to Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era*. Beijing, China: CPC News Network, 2025.
[5] Y. Sun, S. Wang, Y. Li, X. Feng, X. Chen, H. Zhang, et al., "ERNIE: Enhanced Representation through Knowledge Integration," arXiv preprint arXiv:1904.09223, 2019, doi: 10.48550/arXiv.1904.09223.
[6] Natural Language Processing Laboratory, Harbin Institute of Technology, "Application of a Hierarchical Semantic Matching Framework in Legal Texts," Chinese Journal of Information Science, vol. 38, no. 2, pp. 1-10, 2024.
[7] School of Marxism, Tsinghua University, "Construction and Application of a Marxist Classics Corpus," Marxism Studies, no. 5, pp. 120-130, 2023.
[8] National Development and Strategy Institute, Renmin University of China, "Research on CNN-Based Policy Text Classification Models," Management World, no. 8, pp. 200-208, 2022.
[9] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[10] Central Committee of the Communist Party of China, *Report to the 20th National Congress of the Communist Party of China*. Beijing, China: People's Publishing House, 2022.
[11] National Bureau of Statistics, *China Statistical Yearbook 2024*. Beijing: China Statistics Press, 2024.
[12] People's Daily Online, *Compilation of Authoritative Interpretations of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era*. Beijing, China: People's Publishing House, 2024.
[13] Qiushi Journal, *Collection of Research Papers on Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era*. Beijing, China: Qiushi Press, 2023.
[14] General Office of the State Council, *Compilation of Policy Documents for the 14th Five-Year Plan*. Beijing, China: China Legal Publishing House, 2021.

[15] Stanford University Political Communication Lab, “Topic Extraction in Political Texts using LDA Model,” Political Analysis, vol. 31, no. 3, pp. 320-335, 2023.

[16] London School of Economics and Political Science, “Word2Vec in Political Text Similarity Calculation,” Journal of Political Ideologies, vol. 27, no. 1, pp. 45-62, 2022.

[17] China Government Network, “Local Government Policy Document Database [DB/OL],” 2025.

[18] Institute of Computational Linguistics, Peking University, “Optimization and Application of the Chinese Word Segmentation Tool jieba,” Language and Character Application, no. 3, pp. 145-153, 2023.

[19] Scikit-learn Development Team, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 23, no. 268, pp. 1-8, 2022.

[20] PyTorch Development Team, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” Advances in Neural Information Processing Systems, vol. 35, pp. 18020-18030, 2022.

[21] J. Xi, *The Governance of China (Volume III)*. Beijing, China: Foreign Languages Press, 2020.

[22] C. W. Clemens, Jr., “Triangular Diploma in the Age of Putin, Xi and Trump,” Global Asia, vol. 15, no. 1, pp. 84-88, 2020.

[23] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training,” OpenAI, 2018.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, Jan., pp. 993-1022, 2003.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Advances in Neural Information Processing Systems, vol. 26, pp. 3111-3119, 2013.

[26] Communist Party Member Network, “Text Statistical Report on Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era (as of October 2025),” 2025.

[27] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, et al., “ERNIE 3.0: Large-Scale Knowledge Enhanced Pre-training for Language Understanding and Generation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 12161-12175, 2023, doi: 10.48550/arXiv.2107.02137.

[28] V. Ahmed, K. Jouini, and O. Korbaa, “Enhancing LULC Classification with Attention-Based Fusion of Handcrafted and Deep Features,” Proceedings of the 17th International Conference on Agents and Artificial Intelligence, pp. 69-78, 2025, doi: 10.5220/0013098500003890.

[29] Chinese Academy of Social Sciences, *Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era: A Comprehensive Study*. Beijing, China: Foreign Languages Press, 2022.

[30] H. Liu, R. Luo, F. Shang, et al., “Progressive Semantic Matching for Video-Text Retrieval,” Proceedings of the 29th ACM International Conference on Multimedia, 2021, doi: 10.1145/3474085.3475621.