

Application of Natural Language Processing in Intelligent Analysis of Marxist Thought on Agriculture, Rural Areas and Farmers

Weixia Han

School of Marxism, Zhoukou Polytechnic, Zhoukou 466000, Henan, China

Corresponding author: Weixia Han (e-mail: 13903948930@163.com).

ABSTRACT The large scale and decentralized nature of Marxist literature on agriculture, rural areas, and farmers make manual analysis inefficient and limit systematic investigation of theoretical evolution and practical relevance. This study develops an intelligent document-parsing framework based on natural language processing. First, BERT is used for semantic encoding and topic classification of documents concerning land systems, farmers' rights, agricultural modernization, and rural development. Second, named entity recognition is applied to extract key concepts, figures, time nodes, and policy documents, while dependency parsing and graph databases are used to construct a knowledge graph. Finally, TextRank and LDA are combined to generate document summaries and reveal the evolution of themes across historical periods. Experimental results show high topic classification accuracy and strong entity-extraction performance. The proposed framework improves knowledge discovery efficiency and provides an interpretable semantic signal-processing method for intelligent analysis of large-scale ideological and agricultural policy texts.

INDEX TERMS natural language processing, marxist agricultural thought, bert model, knowledge graph, semantic signal processing

I. INTRODUCTION

THE theoretical framework that guides agricultural and rural development in China is Marxist thinking about agriculture, rural areas and farmers. The related literature covers over a century and consists of different forms, including classic works, policy documents, and academic papers. These sources are full of theoretical perspectives of land system reform, peasant classes analysis and the development of methods of agricultural production. The documents are however randomly scattered and their forms of expression are very different and the researcher has to spend a lot of time in retrieving and summarizing the documents manually. The information overload problem has been caused by the fact that the number of documents has been constantly increasing. The evolution of thought can hardly be understood through the traditional process of reading, the logical relationships between important concepts are hardly to be sorted out systematically, and the effectiveness of the cross-period theoretical analysis of decisions is low which limits the level of the theoretical study and reference value of decision-making [1,2].

Document parsing is an intelligent process that has a technical route through natural language processing technol-

ogy. Pre-trained language models develop semantic understanding skills via huge-scale unsupervised learning, named entity recognition technology automatically identifies proper nouns in text, and knowledge graph technology converts unstructured text to a computable semantic network. The paper develops a document parsing system that is built upon deep learning, with the incorporation of BERT semantic encoding, BiLSTM-CRF entity recognition, graph embedding reasoning and others to fulfill functions like automatic topic classification, concept relationship extraction, summary generation and evolution analysis. The system enhances the efficiency of document processing, brings about the patterns of implicit knowledge associations, offers intelligent tool support to the study of Marxist thought agriculture, rural areas and farmers, encourages theoretical innovation and practical implementation [3,4].

II. RELATED WORK

Rule matching techniques have transformed to deep learning techniques in text classification. Convolutional neural networks identify local features with multi-scale convolutional kernels to perform classification at the sentence level, and recurrent neural networks perform classification by gating

to learn long-range interactions. Models can concentrate on the important parts of the information by the introduction of attention mechanisms and the Transformer architecture does not rely on recurrent structures and instead uses self-attention to compute the global semantic associations. Language models are trained on general language representations with large volumes of text, and fine-tuning techniques adjust to the demands of given domain tasks. Method of named entity recognition has changed to statistical learning, as opposed to matching with a dictionary. Conditional random fields are designed to eliminate label conflicts by using global optimization, and neural network models learn character and word features by default [5,6]. Bidirectional long short term memory networks with conditional random fields have since become the standard solution to sequence labeling, and character level embedding solves the out of vocabulary word recognition problem. There are relation extraction tasks where template-based rule methods are used and data-driven learning methods. Remote supervision applies the knowledge bases to create training samples automatically, and attention mechanisms address the issue of multi-instance sentence learning. Construction of knowledge graphs includes entity alignment, attribute fusion, relation completion and so on. Graph embedding techniques map the symbolic knowledge in vectors, enabling similarity computation and reasoning. There are two categories of text summarization namely extractive and generative. Extractive algorithms choose important sentences using sentence importance scores and graph ranking algorithms spread weights over sentence similarity networks. Topic models employ probabilistic graphical models to identify the possible topic distribution of document collections and dynamic topic models monitor the trend of topic change with time, which provides quantitative ways of bibliometric analysis [7,8].

III. METHODS

A. BERT-BASED SEMANTIC REPRESENTATION AND TOPIC CLASSIFICATION OF DOCUMENTS

The deep semantic encoding of Marxist agriculture, rural areas, and farmers thought documents was done using a pre-trained BERT model. Following document segmentation into paragraphs, the text was tokenized into a sub-word sequence via a WordPiece tokenizer. The maximum sequence length for each sample was set to 512 tokens, with sequences exceeding this limit managed via a sliding window approach to ensure complete semantic coverage. The BERT encoder was used to obtain a 768-dimensional dense vector representation of semantic features related to context by using a 12-layer Transformer architecture. A multi-label classifier was built on four core latent thematic clusters. To ensure chronological consistency and avoid terminological anachronisms, these clusters (e.g., "Rural Development and Governance") represent broad conceptual domains rather than specific contemporary slogans. For the LDA analysis, these are treated as latent themes rather

than literal keyword matches, allowing for the mapping of evolving terminology—from early "Rural Construction" to modern "Rural Revitalization"—within the same conceptual framework. The BERT-derived [CLS] token embeddings were subsequently projected into a fully connected layer and the probability distribution of each theme was obtained by the use of Sigmoid activation function. The cross-entropy loss function was adopted in the model training, where the learning rate was 2×10^{-5} , a batch size of 32, and 10 training epochs. In order to optimize domain flexibility, 8000 labeled documents were fine-tuned and a contrastive learning mechanism was provided to reduce the distance between texts on the same topic in the vectors [9,10]. Training was performed using the gradient accumulation method to relieve pressure on memory and the weights of the classes were modified dynamically to resolve the sample imbalance issue. Once the model has converged, a threshold of 0.5 is applied in order to make the subject category of the document and therefore have correct subject classification.

B. NAMED ENTITY RECOGNITION AND RELATION EXTRACTION

BiLSTM-CRF model was designed to detect proprietary entities in Marxist literature on agriculture, rural areas and farmers. Word vectors produced by BERT, 768 dimensional, were fed into a bidirectional long short-term memory (LSTM) network. One forward LSTM was used to obtain left-side contextual information and a backward LSTM obtained right-side contextual features. The hidden states on both directions were joined together to make a full sequence representation. The sequence labeling results were globally optimized with a conditional random field (BIO) layer that prevented unreasonable label shifts like that of person-start immediately after time-middle. There were five types of entities: theoretical concepts, historical figures, policy documents, time nodes, and regional names. To minimize subjectivity in tagging "theoretical concepts," annotators followed a rigorous logical criterion: an entity was only classified as a theoretical concept if it appeared in the Marx-Engels Integrated Index or official glossaries of Marxist-Leninist terms. General nouns (e.g., "land" or "labor") were excluded unless they functioned as part of a specific theoretical construct (e.g., "Land Public Ownership"). The BIO labeling system was used to mark entity boundaries.

Relation extraction uses a remote supervision approach to automatically construct training samples. The seed data are triples extracted out of an existent base of knowledge. Entity pairs in sentence fragments are correlated in the literature corpus, and dependency path weights are determined with an attention mechanism. A multi-instance learning model is created to deal with noisy annotations, combining multiple sentence representations of the same entity pair, and choosing the most trustworthy relation type by max pooling. The relation classifier recognizes 12 semantic relations, such as propose-theory, guide-practice and inheritance-development

and provides relation triplets with a confidence value higher than 0.7 that are then stored in the knowledge base.

C. KNOWLEDGE GRAPH CONSTRUCTION AND REASONING

In this paper, the Neo4j graph database is used to store a knowledge graph of Marxist thought of agriculture, rural areas and farmers, where entities are stored as nodes and the results of relation extraction are converted into directed edges. The entity disambiguation module combines the entities that represent the same entity, e.g., farmers cooperative and co-operative organization, using the measurement of string edit distance and semantic similarity. Attribute fusion mechanism is used to combine entity attribute data of various literature sources, including multi-dimensional attribute values, like the time of proposal, scope of application, and historical development, to the land public ownership node.

Knowledge reasoning uses embedding-based method to learn graph structural properties. The TransE model project entities and relations onto a low-dimensional vector space, and representation learning is trained to minimize the distance between the head entity vector and the relation vector and the tail entity vector. The trained embedding vectors are used to predict links to identify implicit connections that are not directly mentioned in the literature. ConvE convolutional neural network model is an entity-relationship multi-hop reasoning model that transforms entity-relation pairs into two-dimensional matrices as the input into the convolutional layer to capture the intricate patterns of interaction. The path ranking algorithm computes the credibility scores of reasoning paths like “Marx-Land Nationalization-Lenin-Agricultural Collectivization”, filling in the missing relation edges in the knowledge graph, growing the size of the graph nodes to 1200 and relation edges to 3500.

D. ABSTRACT GENERATION AND TOPIC EVOLUTION ANALYSIS

In order to produce summaries of Marxist-Leninist reports on agriculture, rural territories and farmers. The documents are broken into sentence units and a sentence similarity matrix is generated. Cosine distance is used to determine the strength of the semantic association between BERT vectors. To integrate thematic relevance into the extractive process, the sentence importance scores calculated by the TextRank propagation algorithm were weighted by their correlation with the dominant topics identified by the LDA model. Specifically, sentences that exhibited high cosine similarity to the primary LDA topic vectors received a priority boost, ensuring the final summary reflects both structural importance and thematic core. In long documents, hierarchical strategy of summarization is used. The paragraph-level summaries are compressed at paragraph level to create a full-text summary with compression ratio of 15 percent of the original text.

The LDA model was applied to perform thematic analysis of evolution to reveal the trend in ideological evolution. A

corpus of literature was separated into five historical periods and appropriate corpus was used to train an LDA model. The size was to be eight topics and 1000 rounds was to be the number of iteration. The extent of the development of the theme was measured by the calculation of KL divergence of the theme distribution between the different periods. The weight of the major themes in their lexical representation in each period is presented in Table 1. The theme Land Reform had a weight of 0.32 percent in 1950-1960, Agricultural Modernization 0.28 percent in 1980-2000 and Rural Revitalization in 2010-2020, 0.35 percent. The dynamic thematic model monitored the developments of keywords. The incidence of the term collectivization was highest in the 1970s, and it has steadily decreased since then, whereas terms like property rights and market have greatly increased since the reform and opening up.

TABLE 1. Distribution of Marxist Thought on Agriculture, Rural Areas, and Farmers in Different Historical Periods

Period	Land reform	Collective economy	Agricultural modernization	Farmers' rights	Rural revitalization
1950-1960	0.32	0.25	0.12	0.18	0.05
1960-1980	0.18	0.38	0.15	0.20	0.03
1980-2000	0.12	0.22	0.28	0.25	0.08
2000-2010	0.08	0.15	0.25	0.30	0.18
2010-2020	0.05	0.10	0.20	0.22	0.35

IV. RESULTS AND DISCUSSION

A. DATASET CONSTRUCTION AND EXPERIMENT SETUP

A total of 127,800 Marxist documents regarding agriculture, rural areas, and farmers published between 1949 and 2025 were retrieved in CNKI and Wanfang databases. Following deduplication and quality screening, 8520 valid documents remained. The topic categories of 5,000 documents and the named entities and relation triples of 1,200 documents were annotated by a manual annotation team of five graduate students majoring in Marxist theory with a Kappa consistent coefficient of 0.86. The dataset was separated into training, validation and test sets in a 8:1:1 ratio. The experimental setup comprised of an NVIDIA Tesla V100 with 32GB of VRAM, and the deep learning was performed with PyTorch, version 1.12. The BERT model was loaded with bert-base-chinese pre-trained weights, the BiLSTM hidden layer dimension was 256 and the learning rate was slowly increased to 2×10^{-5} through a warmup strategy. The alpha parameter of the LDA topic model was 0.1, beta was 0.01 and Gibbs sampling was 1000 iterations. Every model training uses an early stopping mechanism; the training is terminated when the loss of the validation set does not decrease in five consecutive rounds.

B. PERFORMANCE EVALUATION OF TOPIC CLASSIFICATION AND ENTITY EXTRACTION

The evaluation metrics in this paper are accuracy, recall, and F1 score to determine the difference in performance of the BERT classifier compared to traditional machine learning

techniques. The baseline models consist of three classifiers: Support Vector Machine (SVM), Naive Bayes and Random Forest. The TF-IDF technique is employed in extracting features. The topic classification task performance data are detailed in table 2. BERT model showed the highest results in all the four topic categories at 92.7%, which is 8.3% point higher than SVM. The land system topic classification ranked highest with the F1 score of 94.2% with a relatively lower score of 89.8% with regards to the rural governance topic, which was established to be as a result of overlapping boundaries between the two topic of rural governance and the farmers rights topic. A comparison of entity extraction performance is made in Table 3. BiLSTM-CRF model had a total F1 of 89.3%, and the theoretical concept entity recognition accuracy of 92.5%. Time-related entities such as the rate of recall was relatively low at 85.7%, probably because there are various forms of representing time in literature. Name homonyms led to a policy document entity extraction F1 score of 88.6%, the highest historical figure entity recognition performance of 91.8% and the lowest regional name entity recognition F1 score of 86.4%.

TABLE 2. Comparison of topic classification performance of different models

Model	Land system	Farmers' rights	Agricultural modernization	Rural governance	Overall accuracy
Support Vector Machine	0.852	0.838	0.845	0.829	0.844
Naive Bayes	0.796	0.781	0.803	0.775	0.792
Random Forest	0.868	0.855	0.862	0.847	0.859
BERT	0.942	0.925	0.938	0.898	0.927

TABLE 3. Performance Evaluation of Named Entity Recognition

Entity type	Accuracy	Recall rate	F1 value
Theoretical concepts	0.925	0.918	0.921
historical figures	0.928	0.908	0.918
Policy documents	0.892	0.880	0.886
Time Node	0.901	0.857	0.878
Territory name	0.875	0.853	0.864
Overall	0.904	0.883	0.893

C. KNOWLEDGE GRAPH VISUALIZATION AND ANALYSIS

The representation of the created knowledge graph in this paper is done with the help of Gephi. The size of nodes indicates how many times an entity appears in the literature and edge thickness indicates strength of relationships. The Force Atlas 2 layout algorithm groups semantically related ideas into community forms, and four general knowledge clusters are found, namely land system, production relations, distribution mechanism, and organizational form. Centrality analysis reveals that "land public ownership" exhibits the highest degree centrality, serving as a core conceptual hub, which is linked to other concepts by 87 edges. Meanwhile, the "farmers' cooperative" entity demonstrates significant betweenness centrality, acting as a critical bridge between disparate theoretical modules. Path analysis demonstrates the chain of reasoning, including things like "Marx -

theory of surplus value - agricultural surplus - industrial-agricultural scissors difference" and illustrates the pedigree of thought. The time dimension layer shows the changes in the graph structure in different historical periods. Following the reform and opening up, the node of the market economy quickly turned into a connecting hub and the volume of the subgraph of the reform of property rights increased three times. The subgraph mining shows that the idea cluster of targeted poverty alleviation - rural revitalization - common prosperity is an emerging concept, which represents the trend of modern theoretical innovation.

V. CONCLUSION

The paper applies deep learning technology to the automation of the process of analyzing Marxist texts on rural problems, breaking the efficiency bottleneck of manual reading. BERT semantic encoding in combination with BiLSTM-CRF sequence annotation is successful in capturing deep semantic features of the texts. Knowledge graph construction converts the disjointed information into a graph of knowledge and LDA topic modeling quantitatively displays the development of ideas. The framework offers smart solutions to theoretical researchers to speed up the retrieval and discovery of knowledge in documents. Nevertheless, the system continues to have the problem of information loss during long document summarization and the cross-document viewpoint conflict detection system should be enhanced. Future research will focus on multimodal learning with graph data, creating more comprehensible algorithms, and applying the analysis to other areas of Marxist theory, creating a more comprehensive knowledge service platform.

AUTHOR CONTRIBUTIONS

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

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] S. R. Ghosh, et al., "Relevance of Gandhi's thought in the context of rural development," Towards Excellence, 2023, doi: 10.37867/te150109.

- [2] N. Cebing, "For rural old-age support security situation of legal thought to TongShan County as respondents," *Elder Law*, 2020, doi: 10.47297/wspelwsp2515-824402.20200401.
- [3] A. Le, "Thought and practice on rural domestic sewage treatment in China," *Water Pollution and Treatment*, 2022, doi: 10.12677/wpt.2022.104024.
- [4] Yun Luo, "The enlightenment of Yan Yangchu's civilian education thought on contemporary rural education," *Advances in Education*, 2022, doi: 10.12677/ae.2022.1211749.
- [5] J. Beattie, M. Binder, H. Beks, et al., "'I hadn't ever really thought it was something that I could do': Rural background medical graduates' pathways to medicine," *Australian Journal of Rural Health*, vol. 33, no. 2, 2025, doi: 10.1111/ajr.70047.
- [6] C. McKay, T. N. Kirk, and M. Leake, "'I thought it was going to be trash': Rural high school students' disability-related perception change following Paralympic school day," *Journal of Teaching in Physical Education*, vol. 42, no. 4, 2023, doi: 10.1123/jtpe.2022-0163.
- [7] C. Reid, L. M. Kelly, L. Stanbrook, et al., "'We thought we would just be following the district nurse around': Innovation in Australian student nursing placements in rural primary health care," *Journal of Practical Nurse Education and Practice*, 2021, doi: 10.29173/jpnep16.
- [8] E. Warner, D. Azizoddin, S. G. Frank-Pearce, et al., "'It frees your body from that pain thought': A mixed-methods exploration of patterns, contexts, and experiences of cannabis use for pain in rural communities," *The Journal of Pain*, vol. 25, no. 11, Art. no. 104636, 2024, doi: 10.1016/j.jpain.2024.104636.
- [9] T. Goremusandu, "Analysis of Neshani Andreas's *The Purple Violet of Oshaantu* as a presentation of the writer's liberal African feminist thought to the inequalities of rural Namibian women," *Research in African Literatures*, vol. 54, no. 2, 2024, doi: 10.2979/ral.00009.
- [10] L. Zheng, "The Clustering Effect Analysis of College Marxist Theory Network Propagation Based on Ant Colony Intelligent Algorithm," *Applied Mathematics and Nonlinear Sciences*, 2023, doi: 10.2478/amns.2023.1.00415.