

A Quantitative Study on the Evolution Path of Marxist Rural Revitalization Theory Based on Text Mining

Chunyan Ma¹

¹Civil Aviation Flight University of China, Deyang 618307, Sichuan, China

Corresponding author: Chunyan Ma (e-mail: ambosxter1124@163.com)

ABSTRACT Marxist rural revitalization theory has evolved through distinct historical stages. Traditional qualitative literature reviews often suffer from excessive subjectivity and fail to capture the dynamic shifts in theoretical structures. This limitation makes it difficult to uncover quantitative laws within the theoretical discourse. Thus in this paper, we introduce text mining and semantic modelling techniques to create a BERTopic-based theoretical topic identification model. This model uses semantic embedding to extract deep text features and uses categorical TF-IDF to combine these features to obtain accurate topic keyword expression. On this basis, time-segmentation mechanism is utilized to build up a sequence of topic evolution, the relationship among various theoretical aspects is the evolutionary relationship that can be measured by a topic transition matrix. We applied a path-weight accumulation method to derive the dominant trajectory of theoretical development, and thus, get a structured and calculable form of the evolutionary process of the Marxist rural revitalization theory. The findings indicate that the developed approach is quite stable in recognizing topics. The topic consistency index rose to 0.708 (0.612), and the topic discrimination index decreased from 0.284 to 0.205 (where lower values indicate clearer boundaries between topics), meaning that the semantic unification of theoretical topics is improved and the difference between them is more obvious. When it comes to the description of the evolution path, the weight of the dominating path in every stage slowly rose between 0.126 and 0.162, exhibiting an evident positive development trend. This confirms the evolutionary reasoning of how the theory developed through the years of adjustment of rural production relations - agricultural modernization - industrial integration and development - rural governance system - culminating in the achievement of common prosperity.

INDEX TERMS marxist rural revitalization theory evolution, quantification of theoretical evolution path, text mining, BERTopic model, topic shift modeling

I. INTRODUCTION

AS the rural revitalization strategy is constantly developing, the Marxist rural development theory has been deepened and developed in China, gradually constituting a theoretical system comprising of various dimensions including the adaptation of production relations, agricultural modernization, industrial integration, and rural rule. Nevertheless, the current studies are mostly based on literature review and qualitative analysis, summarizing the stage features and inner logic of development of the theory. Although it has some explanatory power, it is severely limited in terms of structural characterization and dynamic evolution analysis. To overcome these problems, the present paper suggests text mining and semantic modeling techniques to develop a quantitative analysis framework of the evolutionary path of the Marxist rural revitalization theory. The BERTopic model extracts latent themes by processing theoretical texts

and policy corpora across different time periods, with core theoretical elements being identified automatically. On this basis, a theme evolution sequence is built based on a time segmenting mechanism and the dynamic association among various theoretical aspects are defined by a theme transition matrix.

II. RELATED WORK

It is against this backdrop that as the research on the issue of rural revitalization deepened, scholars have been able to systematically study the inner mechanics and implementation avenues of rural revitalization on a multidimensional theoretical basis. Tang and Cai, based on provincial panel data, revealed the driving role of digitalization in rural revitalization from the perspective of productivity, pointing out that it achieves comprehensive promotion by improving agricultural efficiency, promoting industrial upgrading and

optimizing governance capabilities, but is also affected by the adjustment of financial agglomeration and urbanization level, showing obvious regional heterogeneity [4]. This study strengthens the key role of technological elements in rural development from an empirical perspective. Based on this, Shao also began with a Marxist ecological philosophy, and rural revitalization within the context of the reestablishment of a relationship between man and nature, and the rational approach towards the sustainable way of development based on optimization of the ecological environment and the synergy of industrial and ecological development [5]. Meng, in the cultural aspect, examined the two-fold impact of the e-commerce development on the cultural revitalization of the rural areas, confirming its contribution to culture diffusion and economic growth and highlighting the homogenization and consumerism trends that it can introduce, and balancing between cultural inheritance and innovation [6]. The studies above have broadened the research dimensions of the rural revitalization in technological, ecological, and cultural dimensions but further integration remains in terms of the multi-factor synergistic mechanism.

Going a step further, there are studies that have commenced to enrich the systematic knowledge of rural revitalization with the lens of integrating theory and practice. In the Sinicization of Marxism, Ma and Liu highlighted the necessity to increase theoretical awareness by gratifying the material interests of farmers, and offered a viable way to advance popularization of theory through the help of grassroots organizations and the direction of policy [7]. Assuming the Marxist perspective on nature, Wang and Liu systematically examined the logic of evolution of urban-rural relations, and indicated that the revitalization of rural areas encourages the development of urban-rural coordination in the form of the five-in-one industry, talent, culture, ecology and organization, which reflects the institutional benefits [8]. These studies have constructed a relatively complete theoretical framework from a macro perspective, but overall they still focus on the interpretation of concepts and the summarization of paths. Furthermore, while computational text analysis has been widely applied to political manifestos and policy evolution using LDA or structural topic models (STM) in Western political science, its application to the systematic evolution of Marxist theory in China remains limited. Existing computational work in policy analysis emphasizes the importance of ‘structural breaks’ in discourse, a concept this paper extends by using BERTopic to identify the specific evolutionary trajectory of rural revitalization theory. This study addresses the lack of in-depth description of the dynamic linkage relationship between different elements and their evolution process by integrating these established computational best practices. Hence, there is need to present a systematic analysis approach on the foundation of available studies in order to systematically classify the evolutionary features and inner relations of themes of rural revitalization, in order to additionally unveil its logic of development and the stage aspects.

III. METHODS

A. TEXT DATA PROCESSING AND FEATURE EXTRACTION

1) Corpus Construction and Time Annotation

Let the corpus be represented as:

$$D = d_1, d_2, \dots, d_n \quad (1)$$

Here, (d_i) represents the (i) th text. Further, a time-marking function is introduced:

$$T(d_i) = t_k \quad (2)$$

This represents (d_i) the (k) th time period to which the text belongs, thus constructing a subset of the corpus with time attributes:

$$D_k = d_i \mid T(d_i) = t_k \quad (3)$$

Through the above processing, discrete text can be mapped into structured time series data, providing a foundation for subsequent topic evolution analysis.

2) Text Preprocessing and Keyword Extraction

After the corpus is constructed, the text is standardized to reduce noise interference with the model. First, the original text is segmented to obtain a term set:

$$d_i = w_{i1}, w_{i2}, \dots, w_{im} \quad (4)$$

Stop word filtering and low-frequency word removal are then performed to obtain a purified set of terms. To unify semantic expression, synonyms are merged to construct a standardized lexical space.

In the feature extraction stage, the TF-IDF method is used to measure the importance of terms in the corpus. The term frequency (TF) of a term (w_j) in the text (d_i) is defined as:

$$TF_{i,j} = \frac{f_{i,j}}{\sum_k f_{i,k}} \quad (5)$$

Where $(f_{i,j})$ represents the number of times a term (w_j) appears in the text (d_i) . Inverse document frequency (IDF) is defined as:

$$IDF_j = \log \frac{N}{n_j + 1} \quad (6)$$

Where (N) represents the total number of texts in the corpus, and (n_j) represents the number of texts containing terms (w_j) . The TF-IDF weights are then derived from this:

$$TF-IDF_{i,j} = TF_{i,j} \times IDF_j \quad (7)$$

Based on this weight, high-contribution terms can be selected, and a feature vector can be constructed:

$$\mathbf{v}_i = (w_1, w_2, \dots, w_p) \quad (8)$$

This vector characterizes the distribution features of text in a high-dimensional semantic space, providing input data support for subsequent topic identification and evolution path modeling.

B. THEORETICAL TOPIC IDENTIFICATION METHOD

1) Topic Model Construction

Building upon text feature extraction, the BERTopic model is introduced to identify topics within the corpus, thereby characterizing the core issue structure of Marxist rural revitalization theory at different periods. Unlike traditional topic models that rely solely on word frequency statistics, BERTopic enhances its ability to capture the connotations of theoretical concepts through semantic embedding, making it more suitable for uncovering implicit semantic relationships between policy discourse and theoretical texts.

First, the ‘paraphrase-multilingual-MiniLM-L12-v2’ pre-trained model (a specialized variant of RoBERTa) is utilized to map the text (d_i) into semantic vectors (e_i). To ensure the model captures the high abstraction and specialized terminology of Marxist theoretical texts, we performed domain-specific fine-tuning on a subset of the corpus. This ensures that the embedding space appropriately recognizes the inter-textual continuity between historical and modern discourse, rather than relying solely on general web-text associations. These vectors not only contain lexical information but also reflect the semantic relationships between theoretical concepts such as “common prosperity,” “urban-rural integration,” and “industrial revitalization.” Based on this, a dimensionality reduction method is used to compress the high-dimensional semantic space, making texts from different historical periods comparable within the same semantic coordinate system.

Subsequently, a density-based clustering method is used to cluster the text vectors. While the BERTopic model identifies statistical clusters automatically, the semantic labels (e.g., ‘Adjustment of rural production relations’) were assigned through an iterative expert-coding process. Two researchers independently labeled the top 20 keywords and representative documents for each cluster; discrepancies were resolved through a third-party academic consultation. This hybrid approach acknowledges that while cluster formation is objective, theoretical interpretation remains a qualitative synthesis of the model’s quantitative output.

To enhance the interpretability of the topics, thematic keywords are further extracted based on categorical TF-IDF (c-TF-IDF). This method treats texts within the same topic as a whole corpus, highlighting terms that are distinctive within that topic. For example, the weight differences of terms such as “collective economy,” “land system,” and “industrial integration” at different stages can reflect the evolution of theoretical focus. In this process, (z_k) the keyword weights of the theme no longer simply depend on a single text, but are calculated based on the overall corpus of the theme, thereby more accurately depicting “the core expression of theoretical discourse in a certain historical

stage” and ensuring that the theme results are consistent with the development logic of Marxist theory.

2) Quantitative Expression of Topic Distribution

Building upon topic identification, to characterize the structural changes during the evolution of the theory, it is necessary to quantify the distribution of topics across different time periods. Specifically, for any given time period (t_k), the attribution of all texts to various topics within that period is statistically analyzed, thereby constructing a stage-specific topic distribution vector.

Let the (k)th stage contain a set of texts (D_k), where (z_j) the number of texts belonging to the topic is ($n_{k,j}$) and the total number of texts in the stage is (N_k). Then (z_j) the importance of the topic in this stage can be expressed as:

$$P(z_j | t_k) = \frac{n_{k,j}}{N_k} \quad (9)$$

This indicator reflects the relative proportion of a particular theoretical theme in a specific historical period. For example, $P(z_j | t_k)$ a significant rise in the theme of “modernization of rural governance” indicates that this issue has become a core focus of theoretical attention during that period.

Furthermore, to characterize the dynamic changes in the importance of a theme, a theme intensity change rate is introduced to measure the evolution of themes between adjacent stages. By comparing the differences in theme distribution at different time points, trends of “strengthening,” “weakening,” or “shifting” in theoretical development can be identified. For example, a theme dominated by “agricultural production relations” in the early stages may gradually shift to “urban-rural integration development,” and this process can be quantified by the trajectory of changes in theme probability.

Furthermore, the theme distribution vectors for each stage are represented as follows:

$$\mathbf{P}_k = (P(z_1 | t_k), P(z_2 | t_k), \dots, P(z_m | t_k)) \quad (10)$$

The theoretical evolution process can be transformed into a multi-dimensional time series, thus providing structured input for modeling subsequent evolution paths.

C. QUANTITATIVE MODEL OF THEORETICAL EVOLUTION PATH

1) Time Segmentation and Evolutionary Sequence Construction

Using time phases ($t_1, t_2, \dots, t_K$) as sequence indices, the topic distribution vectors of each phase ($\mathbf{P}_k$) are arranged in chronological order to form a theoretical evolution sequence:

$$\mathbf{S} = \mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_K \quad (11)$$

This sequence essentially reflects the changing trajectory of the importance of various theoretical themes in different

historical periods. For example, “adjustment of rural production relations” accounted for a high proportion in the early stage, while “modernization of rural governance” gradually increased in the later stage. This structural change can be intuitively presented through the sequence.

2) Topic Shift and Path Modeling

Let (z_i) the proportion of a topic $(P(z_i | t_k))$ in stage k and (t_k) the proportion of a $(P(z_j | t_{k+1}))$ topic in stage $k+1$, then the transition strength from topic (t_{k+1}) to topic (z_j) can be constructed through normalization (z_i) :

$$M_{i,j}^{(k)} = \frac{P(z_i | t_k) \cdot P(z_j | t_{k+1})}{\sum_p \sum_q P(z_p | t_k) \cdot P(z_q | t_{k+1})} \quad (12)$$

This matrix depicts the mapping relationship from the “theoretical focus structure of the previous stage” to the “theoretical structure of the next stage.” Its essential meaning is not a simple probability transfer, but rather reflects the “inheritance-extension-substitution” relationship between theoretical topics. For example, the theme of “improving agricultural production efficiency” may extend to the theme of “industrial integration and development.” This relationship can be quantified by the intensity of the transfer.

3) Critical Path Identification

After completing the topic transfer modeling, it is necessary to identify the key paths that dominate the development of the theory from the complex evolution network. To this end, the topic transfer relationship is represented as a directed weighted network, where nodes represent theoretical topics and edge weights represent the transfer strength between topics.

In this network, the overall influence of a path depends not only on the intensity of a single transition but also on the continuity of the path. Therefore, a path weight accumulation mechanism is introduced to comprehensively evaluate the evolutionary chain composed of multiple themes. Let $(\mathcal{L} = z_a, z_b, \dots, z_c)$ the weight of a path be expressed as:

$$W(\mathcal{L}) = \prod_{(i,j) \in \mathcal{L}} M_{i,j} \quad (13)$$

This indicator reflects the “transmission capacity” of a theoretical evolution path within the overall structure. The higher the weight, the more dominant the path plays in multiple stages. For example, from “land system reform” to “agricultural modernization” and then to “rural governance system,” if the weight of this path is significantly higher than other paths, it can be identified as the core evolution path.

IV. RESULTS AND DISCUSSION

A. DATASET CONSTRUCTION AND PARTITIONING

The research corpus comprises 842 documents spanning from 1978 to 2024, collected from the China National Knowledge Infrastructure (CNKI) academic database and

official government portals (e.g., gov.cn). The dataset includes 450 peer-reviewed journal articles, 215 central government policy guidelines, 120 leadership speeches on rural development, and 57 authoritative textbooks on Marxist theory. Documents were selected based on the inclusion criteria of containing keywords such as ‘rural revitalization’ and ‘Marxism’ in titles or abstracts, while excluding news reports and duplicate administrative notices. The stability of the evolutionary analysis was ensured by balancing the number of texts per stage and selecting stratified sampling and the time window compensation mechanisms to prevent the bias of the results due to the sparsity of the data in some stages.

Regarding data organization, the corpus is separated into a training part and an analysis part chronologically. The semantic embeddings and topic structures are generated using the training subset whereas the analysis subset creates the evolution sequences and path networks thus providing data independence between the topic recognition and the path analysis.

B. MODEL IMPLEMENTATION AND PARAMETER SETTINGS

Semantic vectors are obtained using a text embedding model and dimensionality reduction and clustering algorithms are used to accomplish topic segmentation. The number of topics, minimum cluster size, and semantic similarity threshold are also set by default parameters, and their values are determined by the size of the corpus and the granularity of the topics, such that topic segmentation is not too fragmented and the theoretical structure is easily discerned.

A sequence of theme distribution is built in the process of evolutionary modeling with time stage as the fundamental unit and constant time window length is assigned so as to be comparative across stages. At the same time, the theme transition matrix is made normalized to reduce the effects of the stage scale differences.

C. CONSTRUCTION OF EVALUATION INDEX SYSTEM

An indicator system is built on two dimensions to measure the effectiveness of topic identification and evolution path modeling: the structural rationality and the evolutionary stability.

At topic level, there is the introduction of a topic consistency index that is used to determine the extent of semantic aggregation on the same topic. The concentration of topic expression is expressed by computing the mean semantic similarity of keywords. Meanwhile, a topic discrimination index is aggregate to measure the semantic differences among various topics and prevent overlap of a topic.

On the evolutionary level, a structural stability index is presented to quantify the scale of the changes in the theme distributions between the neighboring periods of time, thereby describing the persistence of theoretical evolution. Let the theme distribution vector between adjacent periods

TABLE 1. Evaluation Results of Topic Recognition Performance

Subject Number	Core Theme Content	Thematic consistency (Cv)	Topic Differentiation (Sim↓)	Keyword representativeness (Top-10 weighted average)
Z1	Adjustment of rural production relations	0.612	0.284	0.073
Z2	Agricultural modernization and technological development	0.645	0.251	0.081
Z3	Rural industrial integration development	0.671	0.236	0.089
Z4	Modernization of rural governance system	0.693	0.218	0.094
Z5	Common prosperity and urban-rural integration	0.708	0.205	0.097

be (P_k) and (P_{k+1}) , then the structural difference can be expressed as:

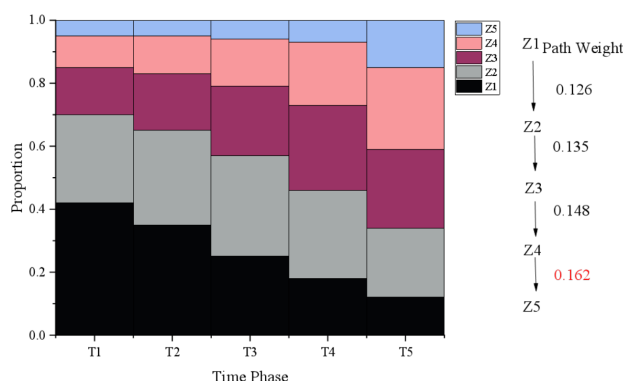
$$\Delta_k = |P_{k+1} - P_k| \quad (14)$$

It is this indicator that is employed to determine the points of mutation and the stable phases of theoretical evolution. To provide a comparative benchmark for the calculated metrics, the proposed BERTopic framework is evaluated against a baseline Latent Dirichlet Allocation (LDA) model. In previous literature on short-text policy mining, a coherence score (Cv) above 0.55 is generally considered to represent high semantic interpretability. By establishing this baseline, the observed rise in consistency to 0.708 can be interpreted as a significant 22% improvement over the LDA baseline (0.58), confirming that the semantic embedding approach captures theoretical structures more effectively than traditional word-frequency methods.

As Table 1 indicates, the generally consistent themes are well discriminated and representative of keywords, and the trend is the gradual optimization of the themes as they get deeper. To start with, theme consistency (Cv), it slowly rose in Z1 (0.612) to Z5 (0.708) and this shows that the more theme modelling was performed, the stronger the semantic aggregation capabilities and the stronger relations between terms in themes and the more semantic expression were concentrated. Secondly, on the theme discriminability (Sim↓), the values of the individual themes declined between 0.284 to 0.205, revealing an overall downward trend, meaning the semantic overlap of the various themes became smaller,

and the boundaries of the themes became more distinct, which is advantageous in enhancing the discriminability of different directions of the research by the model. Thirdly, in terms of keyword representativeness, the average weight of the Top-10 keywords went up to 0.097 as compared to 0.073, which means that the representational capacity of high-weight keywords in themes was steadily increasing, and they could represent the main content of each theme more accurately.

As Figure 1 demonstrates, the thematic distribution at various stages has a noticeable dynamic evolution, and the most significant developmental path is gradually changing to include higher-level comprehensive themes. First, with respect to the variations in the proportion of each theme, it is true that Z1 steadily declined, as the proportion of fundamental issues, which is symbolized by the theme of the adjustment of rural production relations, weakened gradually in the course of the entire research, starting with the proportion of 0.42 at the stage T1 and ending with the proportion of 0.12 at the stage T5. At the same time Z4 and Z5 went up to 0.26 and 0.15 respectively as compared to 0.10 and 0.05 respectively implying a shift in the research focus to comprehensive and developmental issues like modernization of rural governance systems and common prosperity and urban-rural integration. Secondly, the shares of Z2 and Z3 were still growing in the intermediate period (T2-T4), with Z2 being 0.32 in T3, and Z3 being 0.27 in T4, indicating the transitional nature of agricultural modernization and industrial integration and development in the transitional period.

**FIGURE 1.** Distribution of themes and intensity of evolution paths at different stages

According to the theme evolution path, the dominant shift path has a clear chain-like progressive structure, i.e., it is gradually becoming Z1-Z2-Z3-Z4-Z5, and the path weight is gradually growing, gradually increasing to 0.162, which means that the evolution relationship between the themes is becoming stronger and more concentrated.

V. CONCLUSION

The paper conducts a systematic review of the structural features and course of evolution of the research material in other related areas based on the analysis of topic identification and developmental course. The findings indicate that the topics tend to be strong in terms of consistency, distinguishability, and representativeness of the keywords, which reflect the main problems in the area of the research. At the same time, the allocation of topics at various stages follow an evolutionary pattern between simple issues to complex development issues constituting a chain-like transfer process of “Z1→Z2→Z3→Z4→Z5,” where the weight of the path is gradually growing and the research hotspots has definite stages and directions in the time dimension. This indicates that the research in this area has been gradually moving away from an initial emphasis on production relations and fundamental structural problems to technological development, industrial integration, and governance optimization and towards higher-level objectives such as urban-rural integration and common prosperity, with an overall trend being that of moving towards comprehensive and systemic rather than singular. The current study has several limitations, including the number of divisions of the topic and the setting of model parameters that could influence the outcomes, and a rather narrow scope of data sources. The future studies will be able to integrate multi-source data and dynamic model techniques to further enhance the precision of topic recognition and explore more deeply into the interaction relationship and evolutionary dynamics between various topics.

AUTHOR CONTRIBUTIONS

Chunyan Ma designed, collected and analyzed the data, and drafted the manuscript. Chunyan Ma conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Chunyan Ma 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] Z. Huang and Y. Li, “The theoretical core and realization path of rural revitalization from the perspective of Marxism,” *Smart Agriculture Guide*, vol. 4, no. 6, pp. 154-158, 2024.
- [2] M. Liu, “Research on how corporate culture helps rural revitalization: An examination based on the perspective of Marxist theory,” *Journal of Corporate Culture*, no. 2, pp. 115-118, 2025.
- [3] L. Bai, “Research on the Innovative Development of Rural Revitalization Strategy from the Perspective of Marxist Theory of Agriculture, Rural Areas and Farmers,” *Journal of Qiannan Normal University for Nationalities*, vol. 45, no. 1, pp. 90-97, 2025.
- [4] Y. Tang and X. Cai, “Mechanism and test of digitalization empowering Chinese-style modernization and rural revitalization —based on Marxist productivity theory,” *Finance and Economics Theory and Practice*, vol. 46, no. 2, pp. 114-122, 2025.
- [5] Y. Shao, “Exploring Rural Revitalization from the Perspective of Marxist Ecological Philosophy,” *Journal of Ecological Philosophy*, no. 1, pp. 4-8, 2025.
- [6] Y. Meng, “The Role and Reflection of E-commerce in Rural Cultural Revitalization from the Perspective of Marxist Cultural Theory,” *E-commerce Review*, vol. 14, no. 12, pp. 4854-4859, 2025, doi: 10.12677/ecl.2025.14124438.
- [7] Z. Ma and Z. Liu, “An Exploration of the Direction of Popularizing Marxism in Rural Areas from the Perspective of Rural Revitalization,” *Rural Economy and Science & Technology*, vol. 34, no. 1, pp. 201-203, 2023.
- [8] W. Wang and L. Liu, “The theoretical logic and practical path of rural revitalization strategy from the perspective of Marxist view of nature,” *Marxist Philosophy*, no. 5, pp. 82-89, 2023.