Future Digital Technologies and Artificial Intelligence

Open Access Journal

ISSN 3070-0965

May 2026| Volume 02 | Issue 01 | Pages 23-28

https://doi.org/10.55670/fpll.fdtai.2.1.4

Journal homepage: https://fupubco.com/fdtai

Future Publishing LLC

Article

Value co-creation in smart transportation new infrastructure projects: the mechanism of AI-enhanced corporate culture and advertising on strategic niche construction for Shandong SMEs

Hui Yan, Rozaini Binti Rosli*

School of Business & Management, Lincoln University College, 47301, Malaysia

ARTICLE INFO

Article history:
Received 28 December 2025
Received in revised form
19 March 2026
Accepted 24 April 2026

Keywords:
NLP text mining, Machine learning prediction, Value co-creation, Strategic niche construction, Smart transportation infrastructure

*Corresponding author
Email address:
rozaini@lincoln.edu.my

DOI: 10.55670/fpll.fdtai.2.1.4

ABSTRACT

Small and medium-sized enterprises (SMEs) in China's smart-transportation new-infrastructure sector face persistent difficulties in constructing durable strategic niches, yet existing research relies on survey-based methods that overlook the textual signals firms produce through public-facing documents. This study proposes a three-stage analytical framework that integrates NLP feature extraction, machine-learning prediction, and PLS-SEM mediation testing. Drawing on 323 SMEs in Shandong Province and a corpus of 12,864 enterprise text segments, the NLP pipeline extracts culture-sentiment, advertising-sentiment, topic-proportion, and AI-keyword-density features through BERT-wwm-ext, LDA, and TF-IDF. XGBoost achieves the best prediction of strategic niche construction ($R^2 = 0.63$), and SHAP analysis identifies culture-sentiment as the top-ranked feature (mean |SHAP| = 0.42), outperforming all survey-derived variables. PLS-SEM validates that value co-creation partially mediates both paths from AI-enhanced organizational capabilities to niche construction (VAF = 29.3% and 31.7%). The findings indicate that text-derived indicators capture strategic positioning signals that conventional questionnaires miss, offering a replicable mixed-methods paradigm for AI-management crossover research.

1. Introduction

The smart transportation infrastructure in China has witnessed tremendous growth in recent years, driven by investment under the new infrastructure policy regime initiated in 2020. Investment in smart transportation infrastructure exceeds hundreds of billions of yuan across more than 40 pilot cities in China by the end of 2024. Shandong Province in China has three national-level connected vehicle demonstration sites in Jinan, Qingdao, and Yantai. At these sites, local small and medium-sized enterprises provide services such as sensor provision, edge computing modules, data integration, and maintenance platforms to prime contractors and local governments. However, small and medium-sized enterprises in China are structurally weak in the smart transportation infrastructure ecosystem. They are highly dependent on the project cycle and lack direct access to end users. Additionally, local small and medium-sized enterprises in China compete on price rather than on differentiated skills. In a study analyzing value co-creation among enterprises, government organizations, and citizen users in China's smart transportation infrastructure, it was observed that value generation depends on the governance of the data infrastructure as well as on coordination among different government organizations [1]. This leads to another question: how can local small and medium-sized enterprises in China overcome their weak position in the smart transportation infrastructure ecosystem in China. Two organizational resource factors appear promising as facilitators of niche-building activities: corporate culture in AI adoption and AI-based advertising to increase visibility in B2B networks. While the former has garnered most of its research attention in terms of survey methods using Likert scales to measure perceptions without considering textual cues in mission statements of firms, recruitment ads, or WeChat ads, an experimental study of transformer-based language models found that these models were able to classify dimensions of corporate culture according to employee reviews with accuracy rates greater than those of dictionary-based methods [2]. Yet no comparable NLP pipeline has been applied to Chinese-

language enterprise texts in an infrastructure context. On the prediction side, a separate investigation disaggregated digital-intelligence technologies into five domains and tested their effects on innovation performance across 667 Chinese SRDI enterprises using text-mined indicators [3]. However, no similar NLP method has been applied to Chinese language texts in the infrastructure domain. In the area of prediction, another study has segmented digital intelligence technologies into five domains and, using text mining, examined their impact on the innovation performance of 667 Chinese SRDI enterprises.

The missing link in the chain remains to be identified. If the impact of AI culture and advertising on niches is indirect, then what is the mechanism of this impact? The concept of value co-creation as a potential mediator of this impact seems promising for answering this question. Previous research on the impact of platform empowerment strategies on the niche positions of small and medium-sized enterprises (SMEs) at different stages of the firm life cycle has shown variation in their effects [4]. Yet this is based on conventional regression, not a hybrid approach that can simultaneously draw on features of unstructured text and examine mediation. This current study attempts to address both of those voids by offering three research questions. RQ1: Can NLP successfully quantify AI-enhanced corporate culture and advertising in Chinese enterprise texts? RQ2: How successful is machine learning in forecasting SNC outcomes for SMEs, and what features does machine learning find most important for forecasting SNC outcomes for SMEs? RQ3: How important is value co-creation as a mediator of AI-enhanced organizational capabilities in terms of SMEs' ability to build their own niches? This study is based on a population of 323 SMEs in Shandong Province's smart transportation industry and an equivalent amount of text-based data, comprising 15,000 segments. Its originality is in offering a three-stage approach that can simultaneously draw upon computational text-based analysis and conventional strategic management investigation.

2. Analysis of the problems

2.1 NLP for organizational text mining

Organizational culture has traditionally been assessed using questionnaire-based approaches, which are viewed as indirect indicators of managerial values rather than actual communications. This is where text mining comes in. A review of the methodological aspects of text-mining preprocessing in organizational research found that decisions regarding tokenization, stop-word removal, and feature weighting have a direct bearing on the content and stylistic aspects of language [5].In the cultural measurement domain, the dictionary measure, which has been validated against Hofstede's six-dimensional model of culture, has demonstrated internal consistency reliabilities ranging from 0.81 to 0.95 across employee reviews, mission statements, and corporate websites [6]. For example, in the Chinese language sentiment analysis domain, integrating the BERT architecture with a multi-channel CNN and BiLSTM achieved better results than single-architecture approaches when processing 100,000 sentence-level e-commerce reviews [7]. These features can serve as basic components integrated in this study and applied to the context of enterprise texts developed by SMEs for smart transportation infrastructure. The gap in the literature is domain-specific, as no prior work incorporates all these NLP components in a unified manner for AI-related cultural and advertising signals in Chinese-language B2B infrastructure texts.

2.2 Machine learning in strategic management prediction

The NLP-based feature is used for describing the attributes, and no prediction is generated. To make predictions using the attributes generated by the NLP technique, a prediction layer needs to be added. Ensemble tree-based models can handle non-linear interactions and collinearity among features. SHAP can be used to generate feature importance and explain predictions. In the study on the prediction of ESG greenwashing for Chinese listed companies using the XGBoost-SHAP algorithm, the accuracy of the algorithm was found to be 91.0%, and the AUC value of the algorithm was found to be 0.977 in comparison with other algorithms like SVM, LightGBM, and the neural network [8]. Although the problem under consideration is a binary classification problem for large publicly listed companies, the current study differs from the above study in the consideration of a prediction target that is not fixed along multiple dimensions, including its continuous nature, a reduced sample size, and a feature set consisting of NLP and questionnaire-based variables. So far, no study has been identified that incorporates interpretable machine learning for predicting strategic positioning outcomes for SMEs within the infrastructure ecosystem.

2.3 Value co-creation in platform ecosystems

Although quantification and prediction of this process are feasible using natural language processing and machine learning techniques, the mechanism remains unspecified. Service-dominant logic suggests that value is co-created by integrating resources in collaboration among many parties. In a qualitative study of mobility integrators in Germany, some inhibitors to service integration were identified as a lack of willingness to share data, differences in revenue models, and ownership of the value co-created in service bundling [9]. The actor configuration in China's smart transportation sector has a specific structure. In the co-creation process, the SMEs are in touch not only with the platform operators but also with government procurement offices and state-owned prime contractors. The nature of the co-creation process is not only based on business negotiations but also on the public sector's policy and data governance. Whether the co-creation process in this specific context leads to the transfer of the effects of AI-enhanced organizational capabilities to the niche has not been tested.

2.4 Conceptual and technical framework

The three identified gaps correspond to a three-stage analytical pipeline. In Stage 1, natural language processing methods are used to analyze enterprise-level text data and produce structured feature vectors for each firm's culture and AI-related advertising. In Stage 2, features from Stage 1, along with survey and demographic information, are used to train XGBoost and Random Forest models to predict strategic niche construction scores, and SHAP is used to rank features by importance. In Stage 3, partial least squares structural equation modeling is used to test whether value co-creation acts as a mediator in the relationship between organizational capabilities and strategic niche

construction. The theory for the dependent variable is based on a longitudinal case study of a Chinese EV-charging SME, which showed that continuous development and deployment of capabilities within a dynamic ecosystem allowed the firm to develop from a peripheral participant to a niche leader [10]. The three-stage design allows each method to address the question it handles best.

3. Materials and methods

3.1 Research design and data sources

The study adopts a three-stage sequential design that progresses through feature extraction using NLP techniques, prediction using machine learning techniques, and, finally, mediation analysis using SEM techniques. The survey dataset used in this study comprises 323 valid questionnaires collected from owners and senior managers of SMEs engaged in smart transportation new-infrastructure development projects in the cities of Jinan, Qingdao, and Yantai, from March to August 2025. The text dataset used in this study comprises 15,000 text segments collected from the official websites of companies, WeChat public account articles, and tender document company profiles. As shown in Table 1, the dual role of the data sets is clearly presented.

Table 1. Data sources and their roles in the three-stage analytical pipeline

Data Source	Volume	Content Type	Role in Pipeline
Survey questionnaire	323 valid responses	Likert-scale ratings on four constructs and demographics	Target variable for ML prediction (Stage 2) and latent constructs for SEM (Stage 3)
Enterprise text corpus	~15,000 segments	Mission and vision statements, WeChat marketing posts, recruitment pages, tender-document profiles	Input for NLP feature extraction (Stage 1)

3.2 Stage 1: NLP-based feature extraction pipeline

The raw text segments are processed with the Jieba library (version 0.42) for Chinese word segmentation. Next, stop words are eliminated based on the set of stop words created using Baidu Dictionary and 86 industry-related words. In addition, HTML tags and URLs are eliminated. Text segments containing fewer than ten characters are also eliminated. The processed corpus is further processed in three parallel channels. In the sentiment channel, the BERT-wwm-ext model from the Harbin Institute of Technology is used to score the text segments. This model has 12 layers, 768 units, and uses Chinese whole-word masking. The scores range from -1 to +1. The scores are further averaged for each firm to produce the culture sentiment and advertising sentiment. In the second channel, the LDA model is used to analyze the text segments. Topics are determined by maximizing the coherence score (Cv), with counts ranging from 4 to 15. In the keyword channel, TF-IDF scores are computed using the AI terminology dictionary, which contains 47 terms. In addition, AI keyword density for each firm is computed. The features are summarized in Table 2. The output is a feature matrix with 323 rows and p columns, where the rows correspond to enterprise IDs.

Table 2. NLP-extracted features by extraction channel

Channel	Method	Output Variables	Scale
Sentiment	BERT-wwm-ext	Culture-sentiment score and Advertising-sentiment score	Continuous, −1 to +1
Topic	LDA, optimal k selected by Cv	Topic-proportion vector per firm	Proportions summing to 1
Keyword	TF-IDF with AI-term dictionary, 47 terms	AI-keyword density for culture texts and advertising texts	Continuous, 0 to 1

3.3 Stage 2: Machine learning prediction and interpretation

The target variable is strategic niche construction (SNC), calculated as the mean score across 10 questions on a 5-point Likert scale. The feature set includes NLP features extracted from Stage 1, along with the meanings of the remaining three features for the three constructs. The features are also augmented with demographic features. Standardization is performed for features to have a mean of 0 and a variance of 1. If the correlation between any pair of features exceeds 0.85, one feature is dropped. Three different models are trained: an ensemble and a linear regression. The hyperparameters for all three models are listed in Table 3. 5-fold cross-validation is used to evaluate model performance. Grid search is performed for hyperparameters to avoid information leaks in the inner loop of a nested cross-validation method. To better understand the model, SHAP values for the best-performing model using TreeExplainer are calculated. Global feature importance is calculated using the mean absolute SHAP values across all 323 firms. Local explanations for high- and low-SNC firms are also calculated.

Table 3. Machine learning model configurations

Model	Key Hyperparameters	Evaluation Metrics	Validation Strategy
XGBoost	lr = 0.05, depth = 6, n = 500, $\lambda = 1.0$	R^2, RMSE, MAE	5-fold nested CV with grid search
Random Forest	trees = 500, max features = $\sqrt{p}$, min leaf = 5	R^2, RMSE, MAE	5-fold nested CV with grid search
Linear Regression, baseline	Default OLS	R^2, RMSE, MAE	5-fold CV

3.4 Stage 3: SEM auxiliary validation

The SEM stage tests the role of value co-creation as a mediator in the relationship between AI-enhanced corporate culture and AI-enhanced advertising and strategic niche construction. PLS-SEM is used instead of Covariance-Based SEM (CB-SEM) for the following reasons: a medium sample size (n=323) relative to the model's complexity, a focus on predictive validity rather than global model fit, and the presence of formative and reflective indicators for some constructs. All analyses will be conducted using SmartPLS 4. The evaluation of the measurement model will be conducted according to the following criteria: factor loadings (>0.708), Cronbach's alpha and composite reliability (>0.70), average variance extracted (AVE > 0.50), and heterotrait-monotrait ratio (HTMT < 0.85). For the structural model, bootstrapping will be done with 5,000 subsamples for generating bias-corrected confidence intervals and a two-tailed test atα= 0.05 for a one-tailed test.

4. Results and discussion

4.1 NLP Feature Extraction Results

After the preprocessing step, 2,136 segments were removed as they were either below the ten-character limit or contained only boilerplate language. The final dataset contained 12,864 valid text segments from 323 firms. The latent Dirichlet allocation coherence score was maximized for k=7. This means that the seven topics were deemed to have clear meaning. The topic "Innovation orientation" was the highest at 0.19, while the lowest was "cost efficiency" at 0.07. The culture sentiment score was 0.34 on average (SD=0.18), and the intelligent equipment sub-sector performed better than the maintenance services sub-sector by a score of 0.41 compared to 0.27. The advertising sentiment score was slightly more positive and had a mean of 0.41 (SD= 0.22). The density of AI keywords in the culture texts was 0.086 on average (SD = 0.041), while the advertising texts had a higher score of 0.112 (SD = 0.053). As shown in Table 4, the three channels have distinct but complementary feature profiles.

Table 4. Summary statistics of NLP-extracted features (n=323)

Feature	Mean	SD	Min	Max
Culture-sentiment score	0.34	0.18	−0.21	0.82
Advertising-sentiment score	0.41	0.22	−0.15	0.91
AI-keyword density, culture	0.086	0.041	0.00	0.23
AI-keyword density, advertising	0.112	0.053	0.00	0.31
Topic proportion, innovation orientation	0.19	0.09	0.02	0.44
Topic proportion, data-driven management	0.17	0.08	0.01	0.39
Topic proportion, collaborative partnership	0.14	0.07	0.01	0.35

4.2 ML prediction performance and SHAP analysis

As shown in Table 5, the best predictive results were obtained with the XGBoost model, which yielded a mean cross-validated R2 of 0.63, an RMSE of 0.43, and an MAE of 0.33. The second-best results were obtained with the Random Forest method, yielding an R^2 of 0.58. The linear regression model yielded only an R2 of 0.41. The large discrepancy between ensemble methods and linear regression suggests that the relationship between NLP-based features and strategic niche construction cannot be effectively modeled with linear regression. One reason for the better results achieved by the models is the use of heterogeneous features. The NLP-based features have non-linear interaction patterns.

Table 5. Prediction performance comparison across three models (5-fold cross-validation)

Model	R^2	RMSE	MAE
XGBoost	0.63	0.43	0.33
Random Forest	0.58	0.46	0.35
Linear Regression	0.41	0.54	0.42

The SHAP analysis for the XGBoost model provides information about global feature importance ranking, as illustrated in Figure 1. The culture sentiment score is identified as the most important feature in the model with the highest mean absolute SHAP value at 0.42, followed by AI keyword density in advertisements at 0.35. Value co-creation, based on survey scale measurements, is identified as the sixth feature importance ranking at 0.16, behind firm size at 0.19. The top three features identified in Figure 1 are all from the natural language processing analysis, thereby supporting the idea that text-based measurements provide information not available through survey measurements. Local SHAP values for the model are also illustrated in Figure 1 for two different companies. One firm has a high SNC score at 4.52 and is boosted by culture sentiment and innovation orientation topics, while the second firm has a low SNC score at 2.13 and is depressed by low levels of AI keyword density and advertisement sentiment.

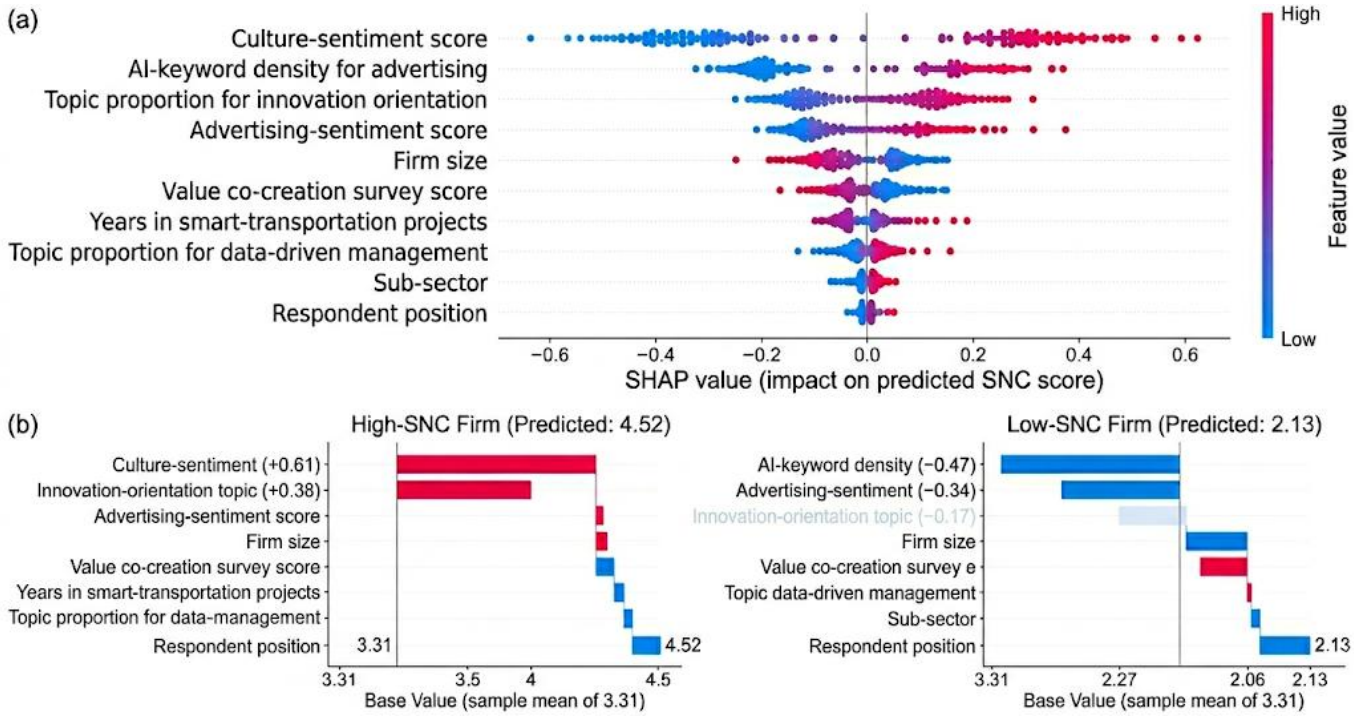

Figure 1. SHAP beeswarm plot of global feature importance and local explanation comparison for the XGBoost model

4.3 SEM auxiliary validation results

The measurement model satisfied all the evaluation criteria. The factor loadings ranged from 0.72 to 0.89. The values of Cronbach's alpha ranged from 0.83 to 0.91. In addition, the composite reliability values ranged from 0.87 to 0.93. The average variance extracted values ranged from 0.56 to 0.68. The results of the structural model are presented in Table 6. All the direct paths were significant at $p<0.001$. The path coefficients for AI-enhanced corporate culture were 0.35 for value co-creation and 0.28 for strategic niche construction. In addition, the path coefficients for AI-enhanced advertising were 0.31 for value co-creation and 0.22 for strategic niche construction. The path coefficient for value co-creation was 0.33 for strategic niche construction. The R^2 value for strategic niche construction was 0.54. In addition, the Q2 value for strategic niche construction was 0.41. Both indirect paths mediated by value co-creation were significant. This is illustrated in Table 6. The VAF values were 29.3% and 31.7%. This implies that both paths are mediated to some extent.

Table 6. Structural model path coefficients and mediation results

Path	β	t-value	p	95% CI
AI-enhanced corporate culture→Value co-creation	0.35	6.14	<0.001	0.24 to 0.46
AI-enhanced advertising→Value co-creation	0.31	5.43	<0.001	0.20 to 0.42
AI-enhanced corporate culture→Strategic niche construction	0.28	4.82	<0.001	0.17 to 0.39
AI-enhanced advertising→Strategic niche construction	0.22	3.67	<0.001	0.10 to 0.34
Value co-creation→Strategic niche construction	0.33	5.91	<0.001	0.22 to 0.44
Indirect effect, corporate culture→VCC→SNC	0.116	4.27	<0.001	0.072 to 0.168
Indirect effect, advertising→VCC→SNC	0.102	3.89	<0.001	0.58 0.154

4.4 Integrated discussion

The findings of the ML and SEM studies complement each other. SHAP results indicate that the culture sentiment score has the highest impact in predicting strategic niche construction. In addition, the PLS-SEM results showed that AI-based corporate culture has the highest total effect (0.28 + 0.116 = 0.396). Advertising features have the second-highest impact in both studies. This convergence of prediction-based and causal inference-based studies provides further support for the absence of method-based artifacts in the observed associations. A notable finding that requires further investigation. Value co-creation ranked only sixth in SHAP importance (0.16), even though it has a significant mediation effect in SEM (VAF=~30%). This is because the SEM study tested the theoretically informed mediation effect of value co-creation in the presence of organizational capabilities. SHAP results, however, estimated the marginal contribution of value co-creation to the prediction, holding all other features constant. In this case, NLP-based features have already captured a significant portion of the variance that value co-creation must explain. What does this imply for SME managers in Shandong Province’ s smart transportation industry? It implies that textual cues have more diagnostic power in predicting niche construction than subjective assessments of value co-creation.

5. Conclusion

The study proposes and tests an analytical framework in three stages that incorporates NLP-based feature extraction, machine-learning prediction, and SEM mediation testing to investigate the role of AI-assisted corporate culture and advertising in SMEs' strategic niche construction within Shandong's new infrastructure for smart transportation. NLP feature extraction successfully quantified cultural and advertising features from 12,864 segments of Chinese enterprise text retrieved from three different sources. XGBoost was found to be better at predicting niche construction scores than Random Forest and linear regression (R^2=0.63). SHAP analysis showed that text features have greater predictive power than survey features. PLS-SEM successfully tested partial mediation of value co-creation for both paths from organizational capabilities to niche construction (VAFs = 29.3% and 31.7%). The major contribution of this study is its methodology, as it is one of the first studies to combine computational text analysis and strategic management research to show that enterprise content is as predictive of enterprise positioning as survey responses from managers. The limitations of this study are its cross-sectional design and the inclusion of only one province in China. Only three sources of text data were considered for this study, and only 12,864 segments were used for analysis, which may not be representative of enterprises with less online presence. Future research can incorporate additional sources, such as annual reports and patents, for analysis; use longitudinal research to study how these enterprises' niches evolve over time; and test this framework on other infrastructure segments, such as 5G and energy networks.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, figure manipulation, competing interests, and compliance with research ethics policies. The author adheres to publication requirements that the submitted work is original and has not been published elsewhere in any language.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the corresponding author.

Conflict of interest

The author declares no potential conflict of interest.

References

[1] J. Zhang, S. Li, and Y. Wang, "Shaping a smart transportation system for sustainable value co-creation," Information Systems Frontiers, vol. 25, no. 1, pp. 365-380, 2023.

[2] S. Koch and S. Pasch, "CultureBERT: Measuring corporate culture with transformer-based language models," in 2023 IEEE International Conference on Big Data (BigData), pp. 3176-3184, 2023.

[3] T. Zhao, F. Zhang, G. Zhuo, Q. Zhang, and Q. Yuan, "The impact of digital intelligence technologies on innovation performance: Evidence from specialized, refined, differential and innovative enterprises," PLoS ONE, vol. 21, no. 2, e0339567, 2026.

[4] Y. Han, "Platform empowerment and SMEs niches base on different life cycles," Technology in Society, vol. 81, p. 102848, 2025.

[5] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," Organizational Research Methods, vol. 25, no. 1, pp. 114-146, 2022.

[6] M. Schachner, M. M. Ardag, P. Holtz, J. Großer, C. Hartz, H. van Herk, and H. Dobewall, "Extracting organizational culture from text: The development and validation of a theory-driven tool for digital data," European Journal of Work and Organizational Psychology, vol. 33, no. 5, pp. 571-582, 2024.

[7] H. Fang, G. Jiang, and D. Li, "Sentiment analysis based on Chinese BERT and fused deep neural networks for sentence-level Chinese e-commerce product reviews," Systems Science and Control Engineering, vol. 10, no. 1, pp. 802-810, 2022.

[8] Z. Jianfeng and Q. Tiantian, "Interpretable predictive model for listed companies ESG greenwashing based on XGBoost and SHAP," Scientific Reports, 2026.

[9] T. Schulz, H. Gewald, M. Böhm, and H. Krcmar, "Smart mobility: Contradictions in value co-creation," Information Systems Frontiers, vol. 25, no. 3, pp. 1125-1145, 2023.

[10] L. Shao, S. Ren, J. Wang, and J. You, "Scaling up into a niche leader in an emerging economy through continuous business model innovation: An effectuation perspective," Asia Pacific Journal of Management, pp. 1-30, 2025.