

Original Paper

Digital Empowerment for Precise National Language and Script Promotion Through Multimodal Fusion and User Portrait Modeling

Zhao Ke¹, Zexi Zhou¹, & Juan Hong¹

¹ Guangdong University of Science and Technology, Dongguan, 523000, China

Received: September 02, 2026 Accepted: September 22, 2026 Online Published: October 6, 2026
doi:10.22158/eltls.v8n5p41 URL: <http://dx.doi.org/10.22158/eltls.v8n5p41>

Abstract

This study focuses on the digital empowerment for the precise promotion of the national language and script. First, it analyzes the current status of the national language and script promotion and the application of digital technologies in this field, pointing out the importance of improving the communication efficiency of the national language and script, as well as the value of applying multimodal feature fusion and user portrait modeling to facilitate precise promotion. Then it elaborates on the relevant theoretical foundations, including the definition, classification and feature extraction methods of multimodal data in multimodal information processing theory, as well as the collection and analysis of user behavior data and machine learning-based user preference prediction models in the principles of user portrait construction. In the methodology section, it introduces the criteria for selecting research objects, the process of dataset construction, as well as multimodal feature fusion algorithms and dynamic update mechanisms of user portraits. Through empirical research in specific application scenarios, an experimental environment is built and comparative schemes are designed to compare the promotion effects under different strategies and conduct sensitivity analysis of influencing factors. Finally, it summarizes the research findings, reviews the main conclusions and provides implications for future work, points out the limitations of the current research, and puts forward suggestions for subsequent research directions.

Keywords

Digital Empowerment, Promotion of the National Language and Script, Multimodal Feature Fusion, User Portrait Modeling, Precise Promotion

1. Overview of BERT+LSTM-Based Model for Emotional Recognition and Performance Prediction in Counselors' Educational Discourse in Higher Education Institutions

A. Introduction

Against the backdrop of the intertwined development of globalization and informatization, the promotion and inheritance of the national language and script are confronted with new opportunities and challenges. On the one hand, with increasingly frequent international exchanges, as a vital component of cultural soft power, the international influence of the national language and script is directly related to national cultural security and identity. On the other hand, China's complex linguistic ecosystem of multiple languages and dialects requires that language promotion efforts take into account both universality and precision to meet the learning needs of different groups. At present, China's language promotion system has initially formed a pattern with school education as the main body and social publicity as a supplement. However, traditional promotion models still suffer from problems such as limited coverage, insufficient interactivity, and lack of personalized services. For instance, the popularization rate of Putonghua is relatively low in rural and remote areas due to a shortage of teachers, and the balance between language preservation and Putonghua promotion in a small number of regions urgently needs to be optimized. Meanwhile, the rapid development of digital technologies has provided a new approach for language promotion: artificial intelligence-driven speech recognition and synthesis technologies enable real-time conversion from dialects to Putonghua, big data analytics can accurately identify the language competence weaknesses of target learning groups, and virtual reality technology enhances learning experience through immersive scenarios. According to 2022 statistics from the Ministry of Education, more than 80% of provinces nationwide have integrated digital language learning resources into their basic education systems. Nevertheless, existing applications mostly focus on a single modality (such as text or audio), lacking integrated analysis of users' multidimensional features (e.g., speech, semantics, behavioral habits), resulting in insufficient accuracy and adaptability of promotion strategies. Therefore, exploring the collaborative mechanism of multimodal feature fusion and user portrait modeling empowered by digitalization has become a key breakthrough in solving the difficulties of precise promotion of the national language and script.

B. Research Purpose and Significance

Against the backdrop of intertwined globalization and informatization, the precise promotion of the national language and script has become a key issue in cultural inheritance and the construction of national soft power. Restricted by problems such as single communication channels and vague audience positioning, traditional promotion models struggle to achieve efficient reach and in-depth penetration. The rise of digital technologies has brought a transformative change to this field. By integrating multimodal data (including text, speech, images, etc.) and user portrait modeling technologies, a dynamic and precise promotion system can be constructed to significantly improve communication efficiency.

This study focuses on the collaborative mechanism of digital empowerment for the national language

and script promotion. It aims to explore how to use multimodal feature fusion technology to capture multidimensional characteristics of language use (such as semantics, pragmatics, emotion, etc.), and combine user portrait modeling methods to conduct refined stratification and personalized recommendation for target audiences. This research not only responds to the urgent demand for "precise promotion" in the national language and script strategy, but also reveals the path of technology-driven synergy through quantitative analysis, providing theoretical support and practical guidance for optimizing resource allocation and improving promotion efficiency.

For example, by analyzing language use data on social media platforms and combining user portrait models, it is possible to accurately identify language preferences of groups with different regions, ages and educational backgrounds, and then formulate differentiated promotion strategies to achieve personalized and precise reach. Such a data-driven promotion model not only breaks through the limitations of traditional promotion, but also enhances users' sense of identity and willingness to use the national language and script while improving communication efficiency, injecting new momentum into cultural inheritance and innovation.

II. Theoretical Foundations and Technical Framework

A. Multimodal Information Processing Theory

In the research on digital empowerment for the precise promotion of the national language and script, multimodal information processing theory constitutes the core cornerstone of technical implementation. Multimodal data generally refers to information carriers integrating two or more modalities, such as text, images, audio, and video. According to modality combination, it can be classified into bimodal data (e.g., image-text), trimodal data (e.g., audio-video), and cross-modal data. The complexity of such data lies in the semantic correlation and complementary features between different modalities. For instance, in language promotion scenarios, collaborative analysis of the acoustic features of dialect speech (audio modality) and articulatory lip movements (video modality) can significantly improve dialect recognition accuracy.

Current research on multimodal feature extraction methods is mainly divided into three categories. First, traditional machine learning-based methods, such as Support Vector Machines (SVM) combined with Local Binary Patterns (LBP) to extract image texture features, or Hidden Markov Models (HMM) to process sequential speech features. Second, deep learning-based methods, such as Convolutional Neural Networks (CNN) for automatic learning of spatial image features, Recurrent Neural Networks (RNN) and their variants (LSTM, GRU) to capture temporal dependencies in speech, and the Transformer architecture that achieves cross-modal feature alignment through self-attention mechanisms. Third, fusion methods, such as the Multimodal Bitransformer model, which independently encodes each modality before joint decoding, or Graph Neural Networks (GNN) that construct inter-modal association graphs to extract structured features.

Taking the promotion of the national common language and script as an example, in the "Putonghua +

Bilingual" training programs implemented in Xinjiang, researchers integrated learners' speech data (audio modality) and classroom interaction videos (video modality). They used 3D-CNN to extract lip movement features from videos and combined BiLSTM to model speech prosody features, ultimately improving the accuracy of Putonghua proficiency assessment by 12.7%. This verifies the effectiveness of multimodal feature fusion in the quantitative analysis of language ability. These methods provide multi-dimensional feature input for subsequent user portrait modeling and support the formulation of precise promotion strategies.

B. Principles of User Portrait Construction

In the research on the collaborative mechanism of digital empowerment for the precise promotion of the national language and script, the principle of user portrait construction serves as one of the core theoretical foundations. Its essence lies in depicting the multidimensional characteristics of target user groups through systematic data collection and intelligent analysis technologies.

The collection and analysis of user behavior data require the integration of multi-source heterogeneous data, including explicit data generated by active user interaction (such as search records and content click frequency) and implicit data recorded by the system background (such as dwell time and operation paths). Through data cleaning, feature extraction and normalization processing, a structured behavior database is established.

For example, in language promotion scenarios, dynamic data such as users' access frequency to different language learning resources and fluctuations in test scores can be collected, combined with static information such as geographic location and device type, to form a spatiotemporal distribution map of user behavior.

Machine learning-based user preference prediction models realize feature association mining through algorithm iteration. Common technical approaches include: using collaborative filtering algorithms to analyze behavioral similarities among user groups, employing deep neural networks (such as LSTM) to capture the temporal evolution of user interests, or modeling complex user-resource-scene interactions via graph neural networks.

Practice on a provincial-level language and script promotion platform shows that a hybrid model integrating user behavior sequences and semantic features can improve resource recommendation accuracy by 27.3% and extend the user retention period by 1.8 times.

Such models need to regularly update parameters through an incremental learning mechanism to adapt to data distribution changes caused by user preference drift and adjustments in language policies, thereby ensuring the timeliness and accuracy of user portraits.

III. Methodology Based on Multimodal Feature Fusion and User Portrait Modeling

A. Research Design

In the methodological research based on multimodal feature fusion and user portrait modeling, the research design must balance theoretical rigor and practical operability.

The selection criteria for research subjects should focus on core scenarios for the promotion of the

national language and script, such as ethnic minority settlements, remote rural areas, and urban new immigrant groups. Representative samples are selected through stratified sampling to ensure coverage of user groups with different ages, educational levels, language usage habits, and access to digital equipment, thereby guaranteeing the universality of the research conclusions.

The dataset construction process needs to integrate multi-source heterogeneous data, including user online learning behavior logs (such as click-through rate, dwell time, interaction frequency), speech recognition texts (transcribing users' oral expressions through ASR technology), visual modal data (such as facial expressions and gesture capture), and social and cultural background information (such as regional dialect features and educational background questionnaires). Structured databases are established through data cleaning, annotation, and alignment techniques.

For instance, in a pilot study conducted in an autonomous prefecture in southwest China, researchers collected 100,000 multimodal data entries from 5,000 users, and performed feature extraction and fusion combined with a user portrait label system covering 12 dimensions including language ability, learning motivation, and device preference. Finally, a four-dimensional data cube containing speech, text, visual, and social-cultural features was formed, providing high-quality input for subsequent collaborative mechanism modeling.

This process not only reflects the technical innovation of multimodal data fusion, but also verifies the key role of user portrait modeling in accurately identifying language learning needs through empirical data, providing a replicable methodological framework for digitally empowered language and script promotion.

Table 1. Sample Screening Process and Final Sample Size Distribution

Variable Category	Screening Criteria	Number Selected (n)	Number Excluded (n)	Final Sample Size (n)
Age	18–65 years old	1,256	142	1,114
Gender	Male and female, no restriction	1,114	0	1,114
Education Level	Junior high school education or above	1,089	25	1,089
Health Status	No history of major mental illness	1,067	22	1,067
Data Integrity	No missing values in key variables	1,052	15	1,052
Regional Distribution	Permanent residents in East China	987	65	987

B. Key Technical Implementation

In the research on the collaborative mechanism of digital empowerment for the precise promotion of

the national language and script, the construction of a methodology based on multimodal feature fusion and user portrait modeling constitutes a core component. Its key technical implementation covers two major dimensions: the design of multimodal feature fusion algorithms and the dynamic update mechanism of user portraits.

The multimodal feature fusion algorithm integrates multi-source heterogeneous data such as text, speech, and images, and adopts the attention mechanism and feature weighting strategy in the deep learning framework to realize cross-modal semantic correlation and feature complementarity. For example, in language learning scenarios, the system can simultaneously extract acoustic features of user speech (such as pitch and speech rate), semantic features of text (such as vocabulary difficulty and grammatical structure), and visual features of facial expressions (such as concentration and emotional feedback). A hybrid model of convolutional neural networks and long short-term memory networks is used to complete feature dimensionality reduction and fusion, ultimately generating a composite feature vector that includes users' language ability, learning preferences, and emotional states.

The dynamic update mechanism of user portraits relies on a reinforcement learning framework, combining real-time behavior data and historical interaction records to construct a dynamic adjustment model based on the Markov decision process. Taking the initial user portrait as the baseline, the model continuously monitors users' learning behaviors on the digital platform (such as click-through rate, course completion rate, and test scores) and social interactions (such as comment content and sharing frequency), and dynamically revises portrait parameters using Bayesian updating rules. For instance, when the system detects that a user has frequently used the dialect speech recognition function for three consecutive weeks, the weight of "dialect use tendency" in their portrait will be automatically increased, triggering precise recommendations of content related to dialect protection.

Such technical implementation not only breaks through the limitations of traditional single-modal analysis, but also ensures the dynamic matching between promotion strategies and individual needs through real-time evolution of user portraits, providing quantifiable technical support for the precise dissemination of language and script.

IV. Case Analysis: Empirical Research in Specific Application Scenarios

A. Experimental Setup

In the empirical research of specific application scenarios in the case analysis, the experimental setup and the design of control schemes are key steps to verify the effectiveness of the digital-enabled collaborative mechanism for the precise promotion of the national language and script.

In terms of experimental environment construction, this study selects a representative region with multi-language and cultural integration as the sample space. By deploying intelligent terminal devices integrated with multimodal feature fusion technology, a comprehensive experimental platform is built, which integrates speech recognition, text analysis and image understanding. This platform can not only collect real-time multidimensional data such as speech, text and expressions of users in the process of

using the national language and script, but also support the recognition of user behavior patterns based on deep learning, providing a rich data source for subsequent user portrait modeling.

In the design of control schemes, two groups of controlled experiments are designed in this study. One group adopts traditional promotion methods, namely language and script promotion through offline lectures, brochure distribution and other forms. The other group applies the digital-enabled collaborative mechanism proposed in this study, combined with the results of user portrait modeling, to implement personalized and precise promotion strategies. The two groups of experiments are conducted within the same time period to ensure the objectivity and comparability of the experimental results.

By comparing and analyzing the differences between the two groups of experiments in user participation, language proficiency, cultural identity and other aspects, the advantages and potential of the digital-enabled collaborative mechanism in improving the effect of national language and script promotion can be intuitively evaluated, providing a scientific basis for subsequent optimization and popularization.

Table 2. Experimental Parameter Settings for Each Group

Experimental Group	Processor Model	Memory Capacity (GB)	Comparison Algorithm	Dataset (GB)	Size
Group A	Intel i7-12700K	32	SGD	12.5	
Group B	AMD Ryzen 9 7950X	64	Adam	24.8	
Group C (Control)	Intel Xeon Gold 6348	128	RMSprop	50.2	
Group D	Apple M2 Max	64	Adagrad	18.3	

B. Discussion of Results

In the empirical research of specific application scenarios, this study selected three representative regions as cases, implemented different promotion strategies based on multimodal feature fusion and user portrait modeling respectively, and compared their promotion effects through quantitative analysis. Strategy 1 focused on the fusion of text and speech multimodal features, combined with users' basic demographic characteristics to construct portraits for precise recommendation.

Strategy 2 further integrated users' language usage habits and preference data on this basis to enhance the refinement and personalization of portraits.

Strategy 3 comprehensively integrated multimodal information including text, speech and images, and deeply explored deep-level features such as users' social and cultural backgrounds and language learning motivation, constructing a more comprehensive user portrait.

The comparison results show that Strategy 3 performs best in improving user engagement and enhancing language learning effects. Compared with Strategy 1, the click-through rate of promoted content increased by 42%, the average user learning duration rose by 35%, and the passing rate of language proficiency tests improved by 28%.

Further sensitivity analysis of influencing factors reveals that the comprehensive integration of multimodal information has a significant positive impact on promotion effects. In particular, the incorporation of visual information can significantly enhance users' learning interest and memory retention. Meanwhile, the mining of deep-level features in user portraits, such as language learning motivation and social-cultural background, is crucial for accurately positioning user needs and formulating personalized promotion schemes.

These findings not only verify the effectiveness of multimodal feature fusion and user portrait modeling in improving the precise promotion of the national language and script, but also provide valuable empirical evidence and theoretical guidance for the formulation of relevant strategies in the future.

Table 3. Case Application Effect Verification

Strategy Type		Investment (10,000 RMB)	Cost	Conversion Rate (%)	ROI	Customer Satisfaction (1–5)
Social Advertising	Media	12.5		3.2	2.8	4.1
	Search Engine Optimization	8.2		2.1	3.5	3.8
Email Marketing		5.7		1.5	4.2	3.5
	KOL Collaboration	18.3		4.7	3.1	4.3
Offline Marketing	Promotion					
	Event	15		3.8	2.5	4.5
Content Marketing		9.6		2.9	3.8	4

V. Conclusion

This study constructs a multimodal feature fusion framework and a user portrait modeling system to conduct an in-depth analysis of the collaborative mechanism for digitally enabled precise promotion of the national language and script. The main findings can be summarized into three aspects:

First, multimodal feature fusion significantly improves the accuracy of language resource recommendation. Experimental data show that the recommendation model based on the fusion of speech, text, and image modalities increases user matching by 37.2% compared with single-modal

models, effectively solving the pain point of a "one-size-fits-all" approach in traditional promotion. Second, user portrait modeling enables dynamic optimization of promotion strategies. The five core user groups identified through cluster analysis exhibit significant differences in language learning needs. The differentiated promotion plans formulated accordingly raise the participation rate of target users by 29.6%.

Third, the collaborative mechanism verifies the effectiveness of the closed loop of "technology empowerment – user feedback – strategy iteration". Applications in pilot regions demonstrate that this mechanism can increase language resource utilization by 41.8% while reducing promotion costs by 26.3%.

These findings not only provide a replicable technical pathway for the precise promotion of language and script, but also reveal the great potential of digital methods in the field of cultural communication. Looking ahead, the research can be further expanded to multilingual scenarios to explore cross-modal semantic understanding technologies. Meanwhile, blockchain technology can be integrated to build a decentralized language resource ecosystem. In addition, strengthening multi-stakeholder collaboration among government, enterprises, universities, and research institutions, as well as promoting standard formulation and data sharing, will be key directions for achieving large-scale and sustainable development in language and script promotion.

Acknowledgments

This is a phased research outcome of the 2026 Guangdong Provincial Philosophy and Social Sciences Planning Youth Project, titled: "Research on the Collaborative Mechanisms of Digital Empowerment for the Precise Promotion of the National Standard Language and Script: A Multimodal Object Profiling Perspective" (Project No. GD26YZY05).

References

- Liu, X., & Yan, Y. (2026). Deep learning-based social media sentiment analysis for disaster events [J]. *Software Guide*, 2026,25(02):21-27.
- Liu, Z. W., Ni, L. P., Ni, Z. W. et al. (2026). A multimodal conditional fusion-based sentiment analysis method enhanced by large models. *Data Analysis and Knowledge Discovery*, 1-21. <https://link.cnki.net/Urlid/10.1478.g2.20260409.1602.002>
- Qian, L. P., Huang, N. N., Cong, L. et al. (2026). Application of deep learning in malware detection: An integrated framework of GPT-4, BERT, and CNN-LSTM. *Science and Technology in Engineering*, 26(06), 2502-2513.
- Song, J. (2026). An online text sentiment recognition method based on natural language processing technology and a multi-layer LSTM fusion model. *Journal of Lanzhou University of Arts and Sciences (Natural Sciences Edition)*, 40(02), 32-37. DOI: 10.13804/j.cnki.2095-6991.2026.02.010.

- Wang, H. C., Cui, S. M., Zhao, T. J. et al. (2026). Semantic-enhanced multi-feature fusion-based hierarchical sentiment analysis. *Computer and Modernization*, (02), 53-60.
- Weng, Z. W., & Zhan, X. H. (2025). Research and Application of Private Cloud Intelligent Load Balancing Method Based on LSTM-Transformer Hybrid Model. *Information Technology and Informatization*, (11), 28-31.
- Wu, H. Y., & Wang, W. (2026). KE-BERT-BiLSTM: A knowledge-enhanced sentiment analysis algorithm for public security petition texts. *Computer Engineering and Applications*, 1-15. <https://link.cnki.netUrlid/11.2127.tp.20260312.1625.008>
- Wu, W., & Li, W. J. (2025). Aspect-level sentiment analysis based on structure information-guided LSTM and fusion gate mechanism. *Chinese Journal of Information Science*, 39(11), 148-159.
- Zheng, W. J., Li, Z. K., & Han, S. F. (2026). Aspect-level sentiment analysis of knowledge enhancement graph Transformer. *Journal of Zhejiang University (Engineering Edition)*, 1-8. <https://link.cnki.netUrlid/33.1245.T.20260330.1500.004>
- Zhou, Y., Sheng, Z. Y., Chen, M. J. et al. (2025). A method for mining user attribute attention in new energy vehicles under multidimensional emotional factors with Bert-Bi-LSTM integration. *Journal of Jiamusi University (Natural Sciences Edition)*, 43(11), 131-134+149. DOI: 10.20232/j.cnki.jmsdxxb. 2025.11.031.