



A Study on the Influencing Factors of Digital Inclusion of People in Rural Areas Based on Text Mining

Chunlan Dong¹, Wenfeng Fan², Keying Wang¹, Shengxiang Liang^{3,4,*}

¹School of Humanities and Management, Guilin Medical University, Guilin 541004, Guangxi, China.

²School of Game Industry, Fuzhou Software Technology Vocational College, Fuzhou 350213, Fujian, China.

³Health Management Center, Guangxi Clinical Research Center for Diabetes and Metabolic Diseases, The Second Affiliated Hospital of Guilin Medical University, Guilin 541100, Guangxi, China.

⁴Guangxi Key Laboratory of Metabolic Reprogramming and Intelligent Medical Engineering for Chronic Diseases, Guilin Medical University, Guilin 541100, Guangxi, China.

How to cite this paper: Chunlan Dong, Wenfeng Fan, Keying Wang, Shengxiang Liang. (2025) A Study on the Influencing Factors of Digital Inclusion of People in Rural Areas Based on Text Mining. *Journal of Humanities, Arts and Social Science*, 9(6), 1234-1243.

DOI: 10.26855/jhass.2025.06.029

Received: May 27, 2025

Accepted: June 20, 2025

Published: July 15, 2025

***Corresponding author:** Shengxiang Liang, Health Management Center, Guangxi Clinical Research Center for Diabetes and Metabolic Diseases, The Second Affiliated Hospital of Guilin Medical University, Guilin 541100, Guangxi, China; Guangxi Key Laboratory of Metabolic Reprogramming and Intelligent Medical Engineering for Chronic Diseases, Guilin Medical University, Guilin 541100, Guangxi, China.

Abstract

Objective: This study aims to explore the key factors influencing the digital inclusion of rural residents by analyzing social media comments using the LDA topic model. **Design/methodology/approach:** The study employs Python technology to crawl comments related to rural digital inclusion from platform X. After preprocessing, 11,481 valid English texts were selected for analysis using the LDA model to identify potential influencing factors. **Findings:** The results reveal that the main factors influencing digital inclusion include technological infrastructure, social structure, government policies, and individual applications. These factors interact and collectively affect rural residents' digital access and participation. **Originality/value:** This study highlights the multidimensional factors at the technological, social, policy, and individual levels, providing valuable insights for future policy development. It emphasizes the importance of infrastructure investment, digital literacy improvement, policy optimization, bridging the urban-rural digital divide, and promoting cross-sectoral collaboration to achieve comprehensive digital inclusion.

Keywords

Text mining; digital inclusion; rural areas; LDA model

1. Introduction

Digital inclusion refers to ensuring that all groups, especially marginalized populations, have equal access to and the ability to use digital technologies, thereby enabling full participation in modern society. Specifically, digital inclusion encompasses the provision of affordable internet access, appropriate digital devices, digital literacy training, and relevant content in order to bridge the digital divide and promote social equity and inclusion (Sieck et al., 2021).

With the rapid development of information technology, the digitalization of society is driving transformative changes across various sectors worldwide. In rural areas in particular, the widespread adoption of digital technologies not only offers residents greater economic opportunities but also exerts a profound impact on education, healthcare, and public services (Alonso et al., 2024). In the field of education, online learning has provided rural students with access to high-quality educational resources (Oraki et al., 2020). In the healthcare sector, telemedicine platforms have effectively reduced the urban-rural healthcare gap (Potter et al., 2016). Furthermore, e-commerce has also led to

economic growth in China's rural areas, demonstrating the immense potential that digital technologies bring to rural communities (Zhou et al., 2021).

However, digital inclusion in rural areas still faces numerous obstacles. Firstly, weak infrastructure is the main issue, with many rural areas experiencing poor network coverage and unstable signals, affecting the widespread adoption of digital technologies (Meng et al., 2023). Secondly, educational disparities lead to rural residents lacking relevant knowledge and skills in digital technologies, further exacerbating the digital divide (Zhang, 2023). Moreover, although governments worldwide have introduced policies to promote rural digital development, issues such as ineffective policy implementation and insufficient funding still persist (Roberts et al., 2017). Furthermore, the relatively traditional social structures and mindsets in rural areas also present significant challenges to the promotion of digital technologies (Correa & Pavez, 2016).

Nevertheless, the demand for digital inclusion in rural areas is increasingly urgent. Digital technologies can bring economic growth to rural areas, increasing farmers' income through models such as e-commerce and smart agriculture (Deng et al., 2024). In the fields of education and healthcare, digital inclusion can overcome geographical barriers and provide rural residents with the same resources and opportunities as those available in urban areas (Liu et al., 2023). In addition, the widespread adoption of digital technologies can enrich the cultural and spiritual lives of rural residents and enhance their sense of social participation.

In recent years, with the proliferation of social media and the advancement of data mining technologies, an increasing number of scholars have begun to explore the use of text mining techniques to analyze social phenomena reflected on social media and to uncover hidden, valuable information (Diaz-Garcia et al., 2023). Text mining techniques, particularly the Latent Dirichlet Allocation (LDA) topic model, have been widely applied to extract latent themes and patterns from large-scale textual data. For example, Amara et al. utilized the LDA model to conduct topic modeling on multilingual Facebook posts related to COVID-19, covering languages such as English, Arabic, Spanish, Italian, German, French, and Japanese. Their study successfully tracked the evolving trends of pandemic-related discussions across different language communities, demonstrating the effectiveness of LDA in analyzing multilingual social media data (Amara et al., 2021). Based on this, the present study aims to utilize the LDA topic model, in conjunction with relevant comment data from social media platform X, to explore the factors influencing digital inclusion among rural residents and to provide theoretical support for policy formulation. By conducting an in-depth analysis of the influencing factors of digital inclusion within rural communities, this research seeks to offer new perspectives for addressing the digital divide and to provide practical guidance for advancing digital development in rural areas.

2. Methodology

2.1 Data Collection

This study employs Python technology to scrape relevant comment texts under the hashtags #RuralResidentsDigital-Divide, #InternetUseByRuralResidents, and #RuralDigitalResources from the social media platform X. The data collection period spans from January 2014 to January 2025, yielding a total of 31,747 comment texts related to digital inclusion among rural populations. The research strictly adheres to Platform X's data usage policies, ensuring full compliance with platform regulations and relevant legal requirements. To protect user privacy, all collected data underwent anonymization, with no personally identifiable information retained. Furthermore, all data processing and analytical procedures were rigorously followed by ethical guidelines to guarantee the study's fairness and transparency.

Three main reasons support the selection of comment texts from Platform X: First, as a global digital public sphere, it reflects socially significant discussions and attitudes toward rural digital inclusion, offering valuable input for policymaking. Second, its high interactivity and real-time nature allow for the timely capture of emerging trends and needs, informing technological innovation and rural transformation. Third, the platform presents diverse perspectives and rich information, enabling a comprehensive understanding of the real challenges and demands faced by rural communities across different countries.

2.2 Text-mining Framework for Topic Discovery

2.2.1 Data Pre-processing

This study utilizes a multilingual distributed dataset (distribution shown in Figure 1). Given that English texts dominate the dataset while other languages (e.g., Japanese and Spanish) have relatively smaller volumes—potentially

pairs generated by the topic model—to identify the optimal number of topics (Mimno et al., 2011). As shown in Equation (1).

$$\begin{aligned} \text{Cosinesimilarity}(A,B) &= \cos(\theta) \\ &= \frac{A \cdot B}{AB} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \end{aligned} \tag{1}$$

Then, perplexity is used as an evaluation metric to assess the model's fitting degree and generalization capability under different topic numbers. Perplexity is a commonly used indicator for evaluating the quality of language models, reflecting the model's predictive ability for new data (Lv et al., 2023). Generally, a lower perplexity indicates that the model can more accurately predict new data, and thus the model is considered better (Liu et al., 2023). The calculation formula for perplexity is shown in Equation (2).

$$\text{Perplexity} = \exp\left\{-\frac{\sum_{d=1}^M \log p(W_d)}{\sum_{d=1}^M N_d}\right\} \tag{2}$$

Finally, to measure topic interpretability, topic coherence evaluation is also required (Röder et al., 2015). The most fundamental topic of coherence is shown in Formula (3).

$$C_k = \sum_{M=2}^M \sum_{l=1}^m \log \frac{D(v_m^k, v_l^k)+1}{D(v_l^k)} \tag{3}$$

3. Findings

Based on the text mining framework for digital inclusion among rural residents, this study conducts an in-depth exploration of the influencing factors of digital inclusion in rural areas.

3.1 LDA Topic Modeling and Visualization

We determined the optimal number of topics based on topic similarity, perplexity, and topic coherence.

First, we plotted a curve with topic similarity on the vertical axis, as shown in Figure 2. Generally, lower topic similarity indicates higher topic distinctiveness and potentially better model quality.

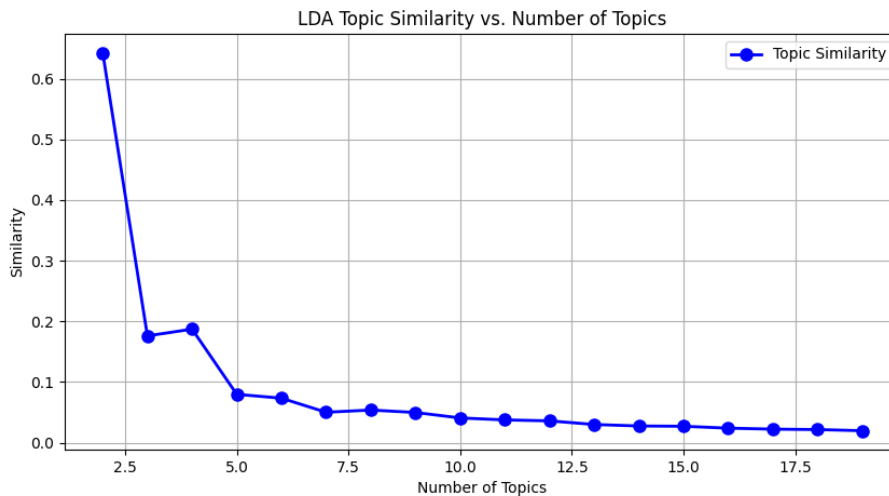


Figure 2. Topic similarity.

Next, we plotted a perplexity curve (Figure 3). Perplexity evaluates language model performance by measuring predictive accuracy on test data—lower values denote better performance.

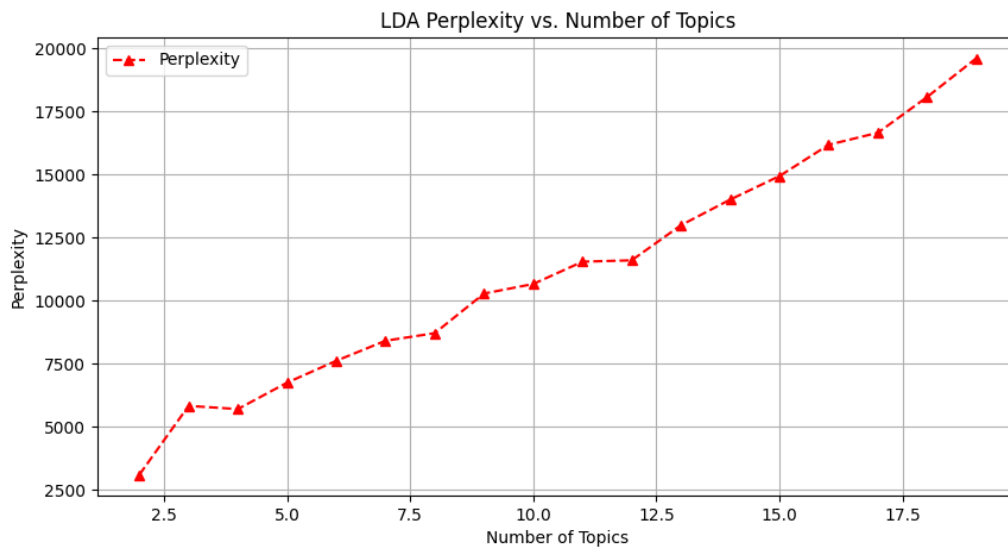


Figure 3. Topic perplexity.

Finally, we generated a topic coherence curve (Figure 4). Higher topic coherence reflects stronger intra-topic relevance and improved interpretability.

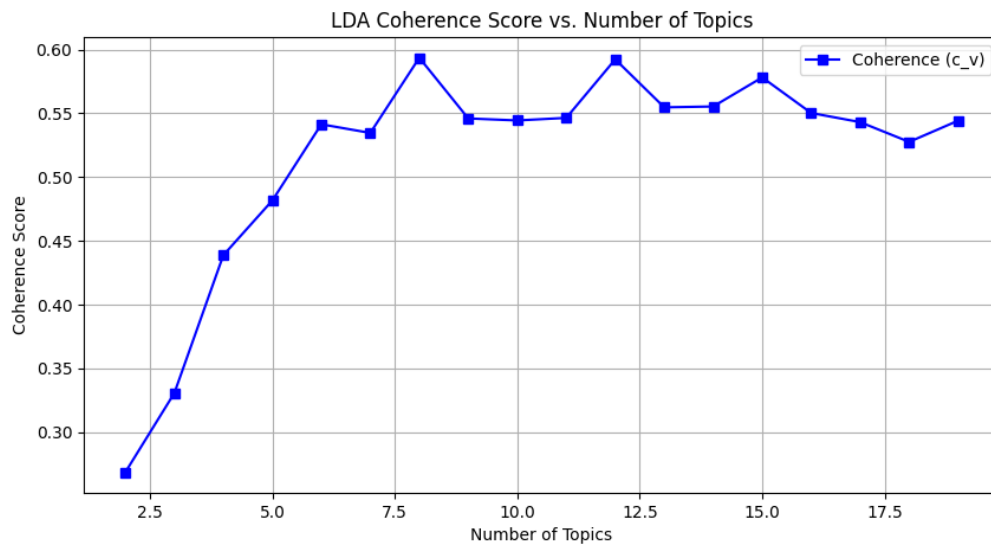


Figure 4. Topic coherence.

By comprehensively considering these three metrics, we ultimately set the number of topics in the model to 8 and reran the model to obtain the most probable topic category for each document.

The analysis results of the LDA topic model for rural residents' digital inclusion are presented in Table 1. Based on the text modeling results, this study manually summarized general titles related to each topic, generating a "topic-vocabulary" table for each topic. From Table 2, it can be observed that the themes of digital inclusion among rural populations encompass multiple aspects, such as rural residents' livelihoods, mobile data, remote areas, and the digital divide. The keywords under these topics highlight critical barriers to digital inclusion for rural populations. For example: Rural livelihoods, Remote areas, Digital divide, Mobile data costs. These factors significantly impede rural populations' integration into the digital society, thus constituting key influencing factors for digital inclusion.

Table 1. Theme-Glossary

No.	Thematic	Main Keywords
1	Mobile Connectivity	Connectivity, mobile, digital, state, connect, provide, network, south, project, speed
2	Mobile Phone	Mobile, month, phone, people, available, pay, shutdown, connection, follow, place
3	Village Network	Village, network, phone, school, people, power, smart, first, service, connection
4	Rural People	People, medium, rural, social, population, world, small, country, high, information
5	Digital Access	Access, rural, digital, infrastructure, divide, expand, reliable, learn, education, economy
6	Rural and urban	Rural, urban, core, people, health, consumption, report, covid, population, well
7	Billion Connect	Million, billion, connect, rural, program, affordable, people, connectivity, single, president
8	Government Speed	Public, government, speed, local, service, mobile, connectivity, national, night, confirm

3.2 Analysis of Affecting Factors

To ensure the accuracy of influencing factors, this study integrates the thematic distribution characteristics of rural digital inclusion and consolidates the eight topics through inductive analysis. For example: Technical infrastructure factors, Social structural factors. The categorization aligns with prior research by relevant scholars to maintain methodological rigor (Ordonez et al., 2011; Cheng et al., 2024). From technical, social, policy, and individual dimensions, we further identified and summarized eight influencing factors of digital inclusion among rural populations for subsequent analysis. The results are presented in Table 2.

This study, based on the eight topics mined by the LDA topic model, systematically reveals the multidimensional influencing factors of rural residents' digital inclusion and their interaction mechanisms. First, mobile network infrastructure (Topic 1) and village-level network coverage (Topic 3) serve as foundational conditions, directly affecting the physical feasibility of digital access in rural areas. This aligns with Van Dijk's "layered digital divide theory," which posits that inadequate infrastructure leads to a "primary access divide," limiting residents' basic exposure to digital services (Huisman & Dijk, 2021). The study found that network latency and signal dead zones in remote rural areas may exacerbate residents' distrust of digital tools, creating a "technology rejection" mentality (e.g., "shutdown," "available" in Topic 2).

Secondly, user behavior (Topic 2) and digital literacy (Topic 5) highlight the importance of subjective initiative in digital inclusion. Although some rural residents own smartphones, keywords such as "pay" and "learn" indicate that non-standardized payment scenarios and a lack of skill training limit the depth of their digital participation. This finding aligns with Warschauer's "social inclusion technology adaptation framework"—technology use must match users' cultural habits and cognitive abilities (Warschauer, 2004a). For example, elderly groups tend to rely on traditional services due to operational barriers, while younger groups may migrate to urban areas due to rural-urban disparities, further weakening rural digital human capital.

Thirdly, inclusive connectivity policies (Topic 7) and network guarantee services (Topic 8) play a crucial regulatory role by reducing access costs ("affordable," "program") and improving service responsiveness ("speed," "night"). For example, India's BharatNet project significantly increased rural broadband coverage through fiber-optic network subsidy policies, but its inadequate implementation efficiency exposed the limitations of single-policy approaches (BL New Delhi Bureau, 2025; IANS, 2025). This study reveals that population outflow (Topic 4) and urban-rural resource allocation imbalances (Topic 6) may lead to a "policy implementation attenuation effect"—where infrastructure investments struggle to sustain due to labor loss, thereby diminishing policy effectiveness.

Table 2. Table of factors influencing the digital inclusion of rural groups

Dimension	Influencing factors	No.	Explain
Technical facilities	Mobile network infrastructure	1	Whether rural areas have poor mobile signal stability and network delays due to network infrastructure issues.
	Village network coverage	3	Whether there are network coverage issues in rural areas that do not meet the needs of rural schools and residents.
Social structure	Characteristics of the rural population	4	Whether rural residents are affected by population size, social structure, access to information, etc., affects digital adaptation.
	Urban/rural differences	6	Whether there are differences between urban and rural residents in terms of digital resources, health, consumption, and access to information.
Government policy	Network assurance services	8	Whether the government is boosting the digital engagement of rural residents by improving the speed of rural mobile network connections and guaranteeing network nightly services.
	Inclusive connectivity policy	7	Whether the Government is pursuing an inclusive policy to ensure universal access and affordable prices for rural residents.
Individual applications	Digital literacy for users	5	Whether rural residents have access to education and training with the help of digital infrastructure to improve the digital literacy of the rural population and reduce the digital divide.
	User behaviour	2	Whether rural residents' willingness to use digital products is affected by mobile phone payments, internet connection habits, device availability and frequency of use.

4. Conclusions

This study employs text mining methods to identify the influencing factors of digital inclusion among rural residents. The results effectively achieve the preliminary objectives of this research and provide empirical support for policy formulation, technological advancement, and educational needs related to digital inclusion in rural areas. More specifically, the main contributions and findings of this study can be summarized into the following two aspects:

- (1) Analyzing textual content characteristics to identify fundamental factors influencing rural residents' digital inclusion: The study found that keywords such as "connectivity" and "infrastructure" ranked prominently among terms like "rural," "digital," and "mobile," indicating an urgent demand for rural digitization and mobile network infrastructure and accessibility. This highlights network infrastructure as the core driver of rural digital inclusion. Additionally, the research revealed the level of network coverage in rural areas, residents' ability to acquire digital devices, disparities in digital resource allocation between urban and rural areas, and regional heterogeneity—namely, that digital inclusion progresses more slowly among rural residents in developing regions compared to developed countries.
- (2) Identify and categorize the influencing factors of digital inclusion among rural residents: According to this study's analysis, the influencing factors of rural digital inclusion exhibit multidimensional and multilevel characteristics. Through the application of the LDA topic model, the research reveals the key roles of various levels—from technological infrastructure, social structure, government policies, to individual applications—in promoting digital inclusion among rural populations. Specifically, the improvement of technological infrastructure, such as mobile network construction and village-level network coverage, serves as a foundational condition for digital inclusion, directly affecting rural areas' digital access capabilities. Meanwhile, urban-rural disparities and rural population characteristics within the social structure constrain rural residents' digital adaptability and participation to some extent. On this basis, effective government policies, particularly the implementation of network guarantee services and inclusive connectivity policies, can significantly enhance rural residents' digital participation. Simultaneously, individual digital literacy and user behavior are also critical factors determining the depth of digital inclusion. By enhancing education and training, optimizing payment scenarios, and improving device usage support, the digital divide can be effectively narrowed, promoting comprehensive digital inclusion for rural populations.

Based on our analysis and findings, government agencies can further refine inclusive policies, enterprises can adjust strategies to enhance internet access in rural areas, and social organizations can participate more actively in digital literacy training for rural residents, collectively advancing the process of rural digital inclusion.

As an exploratory study, this research inevitably has certain limitations. First, the social media data samples selected for analysis were sourced exclusively from Platform X. While these data possess high social influence and interactivity, potential biases may remain. Additionally, focusing solely on English texts may overlook the unique needs of non-English-speaking populations, as user requirements could vary significantly across different languages and cultural contexts. Furthermore, although this study examined multidimensional influencing factors, its conclusions remain dependent on existing data and model analysis results, without further validation or in-depth exploration of these factors' actual impact and feasibility in policy implementation. More empirical research is needed to verify and support these findings.

In summary, this study provides a multidimensional analysis of the influencing factors of digital inclusion among rural residents, offering a theoretical foundation for future policy formulation. To achieve equitable development in the digital society, policymakers should prioritize sustained investment in technological infrastructure while addressing urban-rural disparities, strengthening digital literacy education, and promoting cross-sector collaboration to enable more comprehensive and sustainable rural digital inclusion.

Author Contributions

Conceptualization: C.D. and S.L.; methodology: C.D. and W.F.; formal analysis: C.D.; writing—review and editing: C.D., S.L., and K.W. All authors have read and agreed to the published version of the manuscript.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- Alonso, N., Vicent, L., & Trillo, D. (2024). Digitalisation and rural tourism development in Europe. *Journal of Rural Studies*.
- Amara, A., Hadj Taieb, M. A., & Ben Aouicha, M. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Applied Intelligence*, 51, 3052-3073.
- An, Y., & Yan, Y. (2022). Intelligent retrieval method of library document information based on hidden topic mining. *Web Intelligence*, 20(2), 93-102. London, England: Sage UK.
- BL New Delhi Bureau. (2025, March 21). Standing Committee on DoT recommends immediate resolution of bottlenecks in spectrum allocation, BharatNet execution. *The Hindu BusinessLine*.
<https://www.thehindubusinessline.com/news/standing-committee-on-dot-recommends-immediate-resolution-of-bottlenecks-in-spectrum-allocation-bharatnet-execution/article69358982.ece>
- Cheng, W., Yang, J., Wu, X., Zhang, T., & Yin, Z. (2024). A quantitative study on factors influencing user satisfaction of

- micro-mobility in China in the post-sharing era. *Sustainability*, 16(4), 1637.
<https://doi.org/10.3390/su16041637>
- Correa, T., & Pavez, I. (2016). Digital inclusion in rural areas: A qualitative exploration of challenges faced by people from isolated communities. *Journal of Computer-Mediated Communication*, 21(3), 247-263.
- Deng, J., Li, X., & Zhang, N. (2024). The impact of digital rural construction on rural revitalization—Empirical evidence from Chinese county panel data. *Agriculture*, 14(11), 1903.
- Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2023). A survey on the use of association rules mining techniques in textual social media. *Artificial Intelligence Review*, 56(2), 1175-1200.
- Guo, S., & Zhang, G. (2023). Comparisons of the Economist topics on three countries from 1991 through 2016. *Libri*, 73(1), 37-50.
- Huisman, M., & van Dijk, J. (2021). The digital divide. *Communications*, 46(4), 611-612. <https://doi.org/10.1515/commun-2020-0026>
- IANS. (2025, March 27). 2.18 lakh gram panchayats have BharatNet link for high-speed Internet services: Minister. *Indiaglitz*. <https://www.indiaglitz.com>
- Javed, R. T., Nasir, O., Borit, M., Vanhée, L., Zea, E., Gupta, S., et al. (2022). Get out of the BAG! Silos in AI ethics education: Unsupervised topic modeling analysis of global AI curricula. *Journal of Artificial Intelligence Research*, 73, 933-965.
- Kim, M., Park, Y., & Yoon, J. (2016). Generating patent development maps for technology monitoring using semantic patent-topic analysis. *Computers & Industrial Engineering*, 98, 289-299.
- Liu, S., Zhu, S., Hou, Z., et al. (2023). Digital village construction, human capital, and the development of the rural older adult care service industry. *Frontiers in Public Health*, 11, 1190757.
- Liu, Z., Shan, S., & Shao, B. (2023). Research on public information needs and public library information service strategies in the post epidemic era. *Intelligence Science*, 41(7), 179-188.
<https://doi.org/10.13833/j.issn.1007-7634.2023.07.021>
- Lv, K., Xiang, M., & Jing, J. (2023). A new species of the genus Lepidoptera (Coleoptera, Staphylinidae) from China. *Library and Intelligence Work*, 67(12), 89-102.
<https://doi.org/10.13266/j.issn.0252-3116.2023.12.009>
- Meng, X., Wang, X., Nisar, U., et al. (2023). Mechanisms and heterogeneity in the construction of network infrastructure to help rural households bridge the "digital divide." *Scientific Reports*, 13(1), 19283.
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262-272). ACL: Stroudsburg, PA, USA.
- Oraki, A. H., Fadavi, M. S., & Saeidian Khorasgani, N. (2020). Identifying the components of distance education in rural areas to provide a distance education model for secondary schools in villages of Iran. *Iranian Journal of Educational Sociology*, 3(3), 145-157.
- Ordonez, T. N., Yassuda, M. S., & Cachioni, M. (2011). Elderly online: Effects of a digital inclusion program in cognitive performance. *Archives of Gerontology and Geriatrics*, 53(2), 216-219.
<https://doi.org/10.1016/j.archger.2010.11.007>
- Paek, S., Um, T., & Kim, N. (2021). Exploring latent topics and international research trends in competency-based education using topic modeling. *Education Sciences*, 11(6), 303.
- Pan, J. (2019). How Chinese officials use the Internet to construct their public image. *Political Science Research and Methods*, 7(2), 197-213.
- Potter, A. J. M., Ward, M. M. P., Natafqi, N. M., et al. (2016). Perceptions of the benefits of telemedicine in rural communities. *Perspectives in Health Information Management*, 13(1), 1-13.
- Roberts, E., Anderson, B. A., Skerratt, S., et al. (2017). A review of the rural-digital policy agenda from a community resilience perspective. *Journal of Rural Studies*, 54, 372-385.
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the*

- eighth ACM International Conference on Web Search and Data Mining* (pp. 399-408). ACM: New York, NY, USA.
- Sieck, C. J., Sheon, A., Ancker, J. S., et al. (2021). Digital inclusion as a social determinant of health. *NPJ Digital Medicine*, 4(1), 1-7.
- Warschauer, M. (2003). *Technology and social inclusion: Rethinking the digital divide*. Cambridge, USA: MIT Press.
<https://doi.org/10.7551/mitpress/6699.001.0001>
- Xie, X., Li, D., Zhu, W., Zhang, L., Du, X., & Wang, H. (2022). Drug efficacy prediction in tumors based on LDA model. In *2022 41st Chinese Control Conference (CCC)* (pp. 5747-5752). IEEE: New York, NY, USA.
- Zhang, D., & Zhang, M. (2022). A review of research progress in the application of LDA topic modelling in graphical intelligence domain. *Library Intelligence Knowledge*, 6, 143-157.
<https://doi.org/10.13366/j.dik.2022.06.143>
- Zhang, Y. (2023). Measuring and applying digital literacy: Implications for access for the elderly in rural China. *Education and Information Technologies*, 28(8), 9509-9528.
- Zhou, J., Yu, L., & Choguill, C. L. (2021). Co-evolution of technology and rural society: The blossoming of Taobao villages in the information era, China. *Journal of Rural Studies*, 83, 81-87.