Analysis of China's Green Computing Power Policy Texts based on LDA Model

Lingwei Zhu*

Portland Institute, Nanjing University of Posts and Telecommunications, Nanjing, China

*Corresponding author: p22000401@njupt.edu.cn

Abstract

In recent years, the contradiction between the exponential growth of global computing power demand and energy environmental constraints has become increasingly prominent. It is urgent to build a new efficient and low-carbon computing power system through technological innovation and model transformation. China's green computing power development path not only provides important support for the high-quality development of the digital economy but also contributes a Chinese solution to global ecological governance. This paper systematically analyzes China's green computing power-related policies, In this study, we collect a total of 104 Chinese government policy documents related to "green computing power" issued between January 2007 and May 2025. Using an LDA-based topic modeling approach to extract 13 sub-themes and subsequently perform hierarchical clustering to aggregate them into five major themes. The results provide a comprehensive, data-driven perspective on policy evolution and inform future strategies for advancing green computing power.

Keywords

Green Computing Power, LDA Model, Policy Text Analysis, Topic Identification, Sustainable Development.

1. Introduction

With the rapid development of technologies such as artificial intelligence and big data, the global demand for computing power has grown exponentially (examples of exponential growth) [1]. How to meet the high computing power demand while effectively controlling carbon emissions has become a research focus in global policy and academic circles [2]. In 2020, China explicitly proposed the goals of "carbon peaking" by 2030 and "carbon neutrality" by 2060 [3]. Under the parallel goals of "double carbon" and the high-quality development of the digital economy, China has successively issued a series of documents to guide the green and low-carbon development of computing power infrastructure and the computing power industry, which fully reflects the Chinese government's attention to green computing power [4].

2. Definition and Discussion of Green Computing Power

Domestic scholars define "green computing power" mainly from three dimensions: technical elements, lifecycle, and value objectives. The White Paper on Green Computing Power defines green computing power as evaluating the load and business output per unit carbon emission taking the whole machine as the object [5]. This view has limitations, only focusing on the energy efficiency of hardware devices. However, with the continuous deepening development of green computing power, its definition has become more perfect. The Research Report on the Development of China's Green Computing Power released for two consecutive years defines it from the full cycle of computing power production, supply, operation, management, and application.

3. LDA Model-Based Green Computing Power Policy Text Analysis

3.1. Data Collection and Preprocessing

To ensure the systematicness and comprehensiveness of the green computing power policy literature analyzed in this study and the quality of the literature data to more accurately reflect the development status of China's green computing power policy field, the policy texts analyzed in this study are sourced from five official websites: the State Council Policy Document Library, the Chinese Government Network, the Ministry of Industry and Information Technology Policy Document Library, the National Development and Reform Commission, and the National Energy Administration. The retrieval strategy was to search for the keywords "green computing power" and combinations of "computing power" and "green" on each official website, with the retrieval time range from January 2007 to May 20, 2025. After manual screening to eliminate incomplete information or documents insufficiently relevant to the theme of green computing power, a total of 104 policy documents were finally obtained as research samples.

Subsequently, to facilitate subsequent LDA topic model training and analysis, Python scripts were used to preprocess the literature data: first, the title, publication date, and body content of the documents were extracted and saved in CSV format; then, regular expressions were used in the Python environment to delete blank lines, special symbols, and redundant spaces; next, the Jieba word segmentation library's paddle mode was called to segment the text; finally, a proprietary stop word list was constructed based on the open-source Chinese stop word list (cn.stopwords) and manually added high-frequency non-topic words (such as "MIIT", "monthly", "not yet", "see", "announcement", "order", etc., 341 phrases), and stop words were deleted to obtain a phrase vector set that can provide effective information for topic analysis.

3.2. Determination of the Optimal Number of Topics

Before conducting LDA topic analysis, the perplexity index is used to measure the generalization ability of the model under different numbers of topics to determine the optimal number of topics. Perplexity is defined as:

$$\text{Perplexity}(D) = \exp\left(-\frac{\sum_m \log p(w_m)}{\sum_m N_m}\right) \tag{1}$$

where $p(w_m)$ is the generation probability of document m, and N_m is the number of words in document m. Then, the sklearn.decomposition. LatentDirichletAllocation was called in the Python environment to train the model one by one for the tokenized corpus in the range of K = 2 to K = 15 (step size of 1), and the corresponding perplexity was calculated. The optimal number of topics was determined by:

$$K^* = \arg\min_K \text{Perplexity}(D|K) \tag{2}$$

The results are shown in the Figure 1. When K = 13, the corresponding perplexity reached the minimum value of 1726.52, indicating that the model had the best fitting and generalization ability on new texts at this time. Therefore, the optimal number of topics in this study was determined to be 13. All subsequent topic training, keyword extraction, and topic score calculation based on LDA were carried out in the K = 13 environment to ensure the stability and reliability of the topic recognition results.

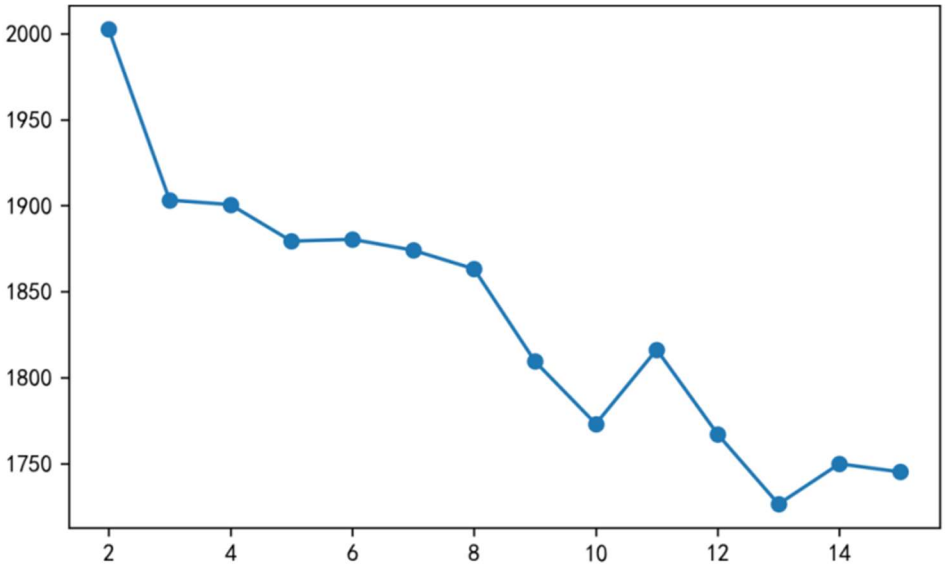

Figure 1. LDA Confusion Curve

3.3. Implementation Process of the LDA Model

After text segmentation and cleaning, this study performed topic mining on the green computing power policy texts (a total of M = 104 documents, vocabulary size V = 11284) based on the LDA (Latent Dirichlet Allocation) model. The model assumes that the document-topic distribution $\theta_m \epsilon \mathbb{R}^K$ and the topic-word distribution $\phi_k \epsilon \mathbb{R}^V$ follow Dirichlet priors respectively:

$$\theta_m \sim \mathrm{Dir}(\alpha) \tag{3}$$

$$\phi_k \sim \mathrm{Dir}(\beta) \tag{4}$$

where the hyperparameters $\alpha = \beta = 1/K$, and the number of topics K=13. Subsequently, for the n -th word of each document m, the topic assignment $z_{m,n} \sim \mathrm{Multinomial}(\theta_m)$ was first sampled from θ_m, and then the word term $\omega_{m,n} \sim \mathrm{Multinomial}(\phi_{z_{m,n}})$ was sampled according to $\phi_{z_{m,n}}$. To achieve parameter estimation, this study used the Variational Bayes algorithm in scikit-learn for approximate inference. During training, the document-word matrix $X \epsilon \mathbb{N}^{M \times N}$ was input, the maximum number of iterations was fixed at 2000, and convergence was determined when the change in ELBO between adjacent iterations was less than 10^{-4}. The algorithm alternately performed the following two steps until convergence: first, calculating the posterior probability of the topic for each document-word pair:

$$q(z_{m,n} = k) \propto \theta_{m,k} \phi_{k,\omega_{m,n}} \tag{5}$$

Second, under the condition of fixing $q(z)$, updating the distribution parameters by maximizing the Evidence Lower Bound:

$$\theta_{m,k} \propto \alpha_k + \sum_n q(z_{m,n} = k), \phi_{k,v} \propto \beta_v + \sum_{m,n:\omega_{m,n}=v} q(z_{m,n} = k) \tag{6}$$

After training, the document-topic distribution matrix $\Theta = [\theta_{m,k}]_{M \times K}$ and the topic-word distribution matrix $\Phi = [\phi_{k,v}]_{K \times V}$ could be extracted from the model object. Based on each row of Φ, the top 15 high-probability words were selected as topic keywords; based on the column-

wise accumulated scores of Θ, the importance of each topic in the overall corpus could be measured.

3.4. Topic Clustering Analysis

After extracting 13 small topics via LDA, this study further employed hierarchical clustering to aggregate the small topics, aiming to reveal the internal correlations between topics and summarize them into more concise major themes.

Following the normalization of small topics and distance calculation, the Ward minimum variance method was applied to the normalized topic-word distribution matrix for hierarchical clustering. A linkage matrix $\mathcal{Z}$ was constructed based on the Euclidean distances of the topic vectors, which recorded the clustering process with the cluster indices and distance values d for each merge. Unlike traditional methods that "hard-code" the number of clusters, this study identified the natural segmentation threshold τ by observing the "jump" points in the merge distances within $\mathcal{Z}$ -i.e., the optimal pruning position occurred when a merge distance significantly increased.

The clustering results are shown in Figure 3. The segmentation threshold was determined based on the significant "jumps" in the merge distances within the linkage matrix. The orange and green highlighted sections correspond to two semantically similar topic groups, while the remaining gray branches represent other topics with excessively large merge distances that were not assigned to the main clusters. A total of 5 clusters (major themes) were generated, where the small topics within the same cluster exhibited high semantic consistency, while distinct differences existed between clusters. The following section will name and analyze each major theme in conjunction with the cluster labels (1-5) in Figure 2.

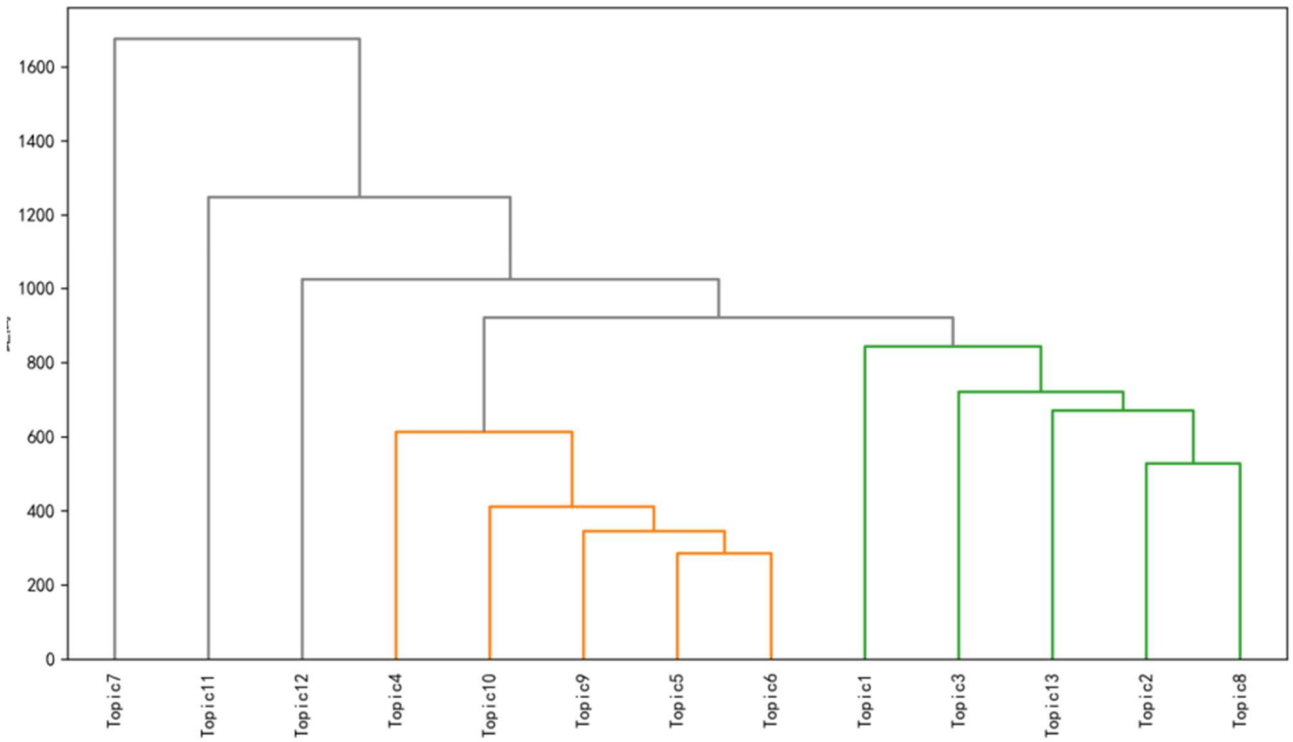

Figure 2. Hierarchical clustering tree with K=5

Through LDA model analysis, we obtained thirteen sub - themes, each with fifteen keywords, as shown in Table 1.

Table 1. Thirteen Sub-Themes and Their Keywords

Topic	Keywords
Topic 1	low-carbon, standard, electricity, power grid, mechanism, field, renewable energy, energy storage, project, clean, resource, new energy, new-type, base, planning
Topic 2	manufacturing, calculation, intelligence, field, industry, energy efficiency, equipment, product, peak regulation, energy conservation, resource, 5G, project, management
Topic 3	5G, industry, digitization, small and medium-sized enterprises, city, data security, integration, scenario, transformation, digital, build, product, collaboration, pilot, field
Topic 4	green certificate, electricity, transaction, consumption, market, mechanism, project, power generation, renewable energy, institution, power quantity, issuance, price, reform, management system
Topic 5	absorption, renewable energy, electricity, IPv6, weight, competent department, undertake, complete, power grid, grid connection, market entity, proportion, municipality directly under the Central Government, provinces, transaction
Topic 6	5G, competent department, park, production capacity, informatization, production, wisdom, safety, reform, department, according to law, backward, coal, supervision, standard
Topic 7	communication, safety, internet, information, 5G, infrastructure, platform, management, telecommunication, integration, network security, collaboration, optimization, business, sharing
Topic 8	industry, product, standard, material, field, research and development, management, electronics, informatization, emission, integration, measurement, future, design, information
Topic 9	new-type, management, dispatching, electricity, project, informatization, energy storage, grid connection, main body, equipment, safety, grid-related, communication, institution, environmental protection
Topic 10	computing power, resource, government, information, platform, safety, unit, society, research and development, management, sharing, field, product, collaboration, formation
Topic 11	data center, computing power, hub, node, collaboration, resource, region, new-type, safety, reform, cluster, informatization, integration, layout, national
Topic 12	energy conservation, emission reduction, standard, supervision, energy consumption, energy efficiency, implementation, equipment, product, management, system, information, environmental protection, project, guidance
Topic 13	demonstration, intelligence, system, research and development, high-efficiency, intelligentization, energy storage, key technology, breakthrough, battery, research, power generation, science and technology, development, test

Five major themes were obtained through clustering. At the same time, we also obtained the importance scores of each theme. The higher the score, the more important the theme. It can be seen from this that among the five major themes, Digital Computing Power and Intelligent Manufacturing is the most important, far higher than other themes, with a score of 43.02. Green Energy and Market Absorption comes next, with a score of 21.36. The other three major themes, Computing Power Hub and Network Security Collaboration, Energy Efficiency, Emission Reduction Standards and Implementation Supervision, Demonstration Projects and Key Technology Research, each have only one sub - theme, that is, the major theme is its only sub - theme. This reflects that the content of policy texts related to these three major themes is highly relevant and centrally reflects one theme. The scores of the themes of Computing Power Hub and Network Security, Energy Efficiency, Emission Reduction Standards and Supervision are much higher than those of the Demonstration Projects and Key Technology theme, which well reflects the focus of policy texts in recent years in Table 2.

Table 2. Major Themes and Their Sub-Themes with Scores

Major Theme Content	ID	Sub-theme Content	Sub-theme Score	Major Theme Score
Green Energy and Market Absorption	4	Low-carbon Power Grid and Standardization of Renewable Energy	7.52	21.36
	5	Intelligent Manufacturing and Energy Efficiency Peak Regulation	2.85	
	6	5G-Driven Digital Transformation	1.92	
	9	Renewable Energy Certificates and Market Mechanisms	5.55	
	10	Grid Connection of Renewable Energy and Market Absorption	3.52	
Digital Computing Power and Intelligent Manufacturing	1	Smart Park Production Capacity and Compliance Supervision	11.49	43.02
	2	5G Communication Network and Security Collaboration	9.11	
	3	Industry Standards and Emission Measurement	9.48	
	8	New-type Power Dispatching and Energy Storage Integration	10.12	
	13	Government Platform-based Computing Power Resource Sharing	2.82	
Energy Efficiency, Emission Reduction Standards and Supervision	7	Energy Efficiency, Emission Reduction Standards and Implementation Supervision	16.95	16.95
Computing Power Hub and Network Security	11	Data Center Computing Power Hub and Integrated Layout	16.85	16.85
Demonstration Projects and Key Technology Research	12	Demonstration Projects and Key Technology Research	5.83	5.83

The themes of Digital Computing Power and Intelligent Manufacturing mainly focus on the digital and intelligent development of the computing power industry. Among them, except for the policies related to government platform - based computing power resource sharing, the other sub - themes such as Smart Park Production Capacity and Compliance Supervision, 5G Communication Network and Security Collaboration, Industry Standards and Emission Measurement, and New - type Power Dispatching and Energy Storage Integration are all relatively important.

4. Conclusion

This study applied LDA-based topic modeling to 104 Chinese government documents (2007–2025), extracting 13 sub-themes aggregated into five major themes. The themes, scored by importance, are: Digital Computing and Intelligent Manufacturing (43.02), Green Energy and Market Absorption (21.36), Computing Hub and Network Security (16.85), Energy Efficiency and Emission Reduction (16.95), and Demonstration Projects and Key Technology Research (5.83). Results demonstrate China's evolving policy focus in green computing power, revealing shifts in priorities over time. These data-driven insights into policy emphasis-digital integration, energy adoption, network collaboration, efficiency standards, and technology demonstration-offer a basis for targeted policy refinement.

References

[1] Wang G, Zhang Q, Su B, et al. Coordination of tradable carbon emission permits market and renewable electricity certificates market in China[J]. Energy Economics, 2021, 93: 105038.

[2] Springer U. The market for tradable GHG permits under the Kyoto Protocol: a survey of model studies[J]. Energy economics, 2003, 25(5): 527-551.

[3] Sun L, Xiang M, Shen Q. A comparative study on the volatility of EU and China's carbon emission permits trading markets[J]. Physica A: Statistical Mechanics and Its Applications, 2020, 560: 125037.

[4] Keohane N O. Cap and trade, rehabilitated: Using tradable permits to control US greenhouse gases[J]. Review of Environmental Economics and policy, 2009.

[5] Larsen B, Shah A. Global tradeable carbon permits, participation incentives, and transfers[J]. Oxford Economic Papers, 1994, 46(Supplement_1): 841-856.