The Lexicon Construction and Quantitative Research of Digital Economy Policy Texts

Ye Li(✉), Ziqiang Shen, Cunyang Zhang, and Linfang Zhao

Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, China
liye@sdas.org

Abstract. At present, China's digital economy policy documents have a large number, rich topics and an increasingly large system. Traditional content analysis methodss have been difficult to achieve a large number of policy text mining and quantitative research. In this paper, we build a professional lexicon of digital economic policy to capture emerging concepts in this field through the new word discovery algorithm and artificial supplement method. The experimental results show that the professional lexicon can increase the F1 score of text segmentation from 57.40% to 77.09%. On the basis of the policy lexicon, using natural language processing technologies such as text classification, keyword extraction, subject modeling, etc., to conduct quantitative research on policies related to the digital economy since the "Fourteenth Five-Year Plan" period is conducive to the intelligent interpretation of the digital economy policy text. The research shows that deepening application, standardizing development and inclusive sharing are the keynote of China's digital economy development. The digital industrialization policy text focuses on five themes: key technologies, emerging industries, new models and new formats, intelligent equipment and support platforms.

Keywords: digital economy · Policy texts · Neologism discovery · Text mining

1 Introduction

The digital economy has become an important driving force for China's economic growth. According to the data of the Ministry of Industry and Information Technology of China, the scale of China's digital economy will reach 45.5 trillion yuan in 2021, accounting for 39.8% of GDP. At present, China has intensively issued a large number of policy documents related to the digital economy, and created a good policy environment for strengthening, optimizing and expanding the digital economy. The policy and regulation system for the development of China's digital economy is composed of these policy documents, which provide rich materials for researchers at all levels of government departments, enterprises and institutions to explore the layout and development trend of China's digital economy.

However, with the increasingly complex policy issues, the scale of data has gradually expanded from small samples to big data. Policy analysis has gradually shifted from traditional policy quantification to policy informatics. Therefore, in the context of the

K. Hemachandran et al. (Eds.): ICAID 2023, AHIS 9, pp. 23–33, 2023.
https://doi.org/10.2991/978-94-6463-222-4_3

increase in the number of policy texts, the shortening of the iteration cycle and the increasingly rich topics, it is of great significance to carry out information extraction, quantitative research and intelligent analysis of policy texts with the help of computer technologies such as natural language processing. The basic steps of quantitative research on policy text are to accurately segment the policy text and convert the unstructured content of the policy text into structured data information.

Therefore, this paper starts with the lack of digital economic policy thesaurus in the existing word segmentation tools. Through the new word discovery algorithm and manual supplement, the digital economy policy thesaurus is constructed, which improves the application effect of text mining technologies such as policy text segmentation, keyword extraction, automatic classification and topic modeling, and is used for quantitative research of digital economy policy text.

2 Research Status

The digital economy policy is an important strategic guide to ensure the high-quality development of the digital economy. At present, scholars' research on digital economy policy mainly focuses on summarizing foreign experience, policy evolution and policy research in specific fields. For example, Li et al. combed the digital economy development strategies of the United States, the European Union, the United Kingdom and other developed economies [1]. Wu et al. revealed the development hotspot of digital economy in the United States and the European Union from the perspective of think tanks [2]. Shi et al. combed the policy evolution and theoretical development of China's digital economy development [3]. Jiang et al. conducted a quantitative analysis of the policy tools used in 45 digital government policies issued by local governments in China [4]. At present, using NVivo software as a tool and adopting content analysis method to conduct quantitative research on digital economic policy text is still the main method adopted by scholars. For example, Lei et al. proposed a three-dimensional analysis framework of "subject tool cycle", which quantifies the text content of China's local government's digital economy policy [5]. Yang et al. made a quantitative analysis of China's provincial government's digital economy policy from two dimensions of policy objectives and policy tools [6]. However, the traditional content analysis method is limited by the huge workload of manual coding, and the amount of policy text it covers is usually small, limited to specific topics [7].

In recent years, scholars have paid more and more attention to the use of natural language processing technology to process policy text information in policy quantitative research to reduce subjective bias. Wang et al. analyzed the thematic features of the EU, UK and US policy texts related to disruptive technologies using the thematic model analysis technology combining word2vec and LDA [8]. Moritz et al. used supervised machine learning algorithm to conduct cross-domain topic classification research on political texts of political parties in Britain and the United States [9]. Zhang et al. compared the similarities and differences of policies by calculating the similarity value of China's big data policy text [10]. Hu et al. used LDA model and improved TextRank model to enhance the effect of strategy text representation, and built an integrated framework of strategy text representation and classification to improve the effect of automatic classification [11].

However, in the process of policy text mining, it is difficult to capture the emerging concepts in this field due to the poor adaptability of the general segmentation tools to the policy text [12]. This may lead to inaccurate subsequent analysis. Therefore, some scholars also apply the new word discovery algorithm to the quantitative research of policy text. Zheng et al. constructed a dictionary of the degree of science and technology policy and applied it to the quantification of policy paragraphs according to the functional orientation and language characteristics of science and technology policy [13]. When studying China's environmental policies and regulations, Wang obtained the professional thesaurus of environmental management through the new word discovery algorithm, which effectively improved the effect of text segmentation, keyword extraction and text classification [7]. Wan et al. used Word2vec, K-means and other natural language processing technologies to analyze the theme and theme changes of China's AI policy at various stages [14]. In order to improve the clustering effect, they also used a new word discovery algorithm to construct new words in the domain. The construction of these specialized lexicon has greatly improved the effect of quantitative analysis of policy text using natural language processing technology.

To sum up, the digital economy policy has become a hot research topic. Scholars spend a lot of time and resources manually marking policy text data for content analysis of digital economic policies. However, manual labeling is subjective and inefficient. This has greatly hampered the development of quantitative policy research. With the rise of natural language processing technology, more and more scholars begin to apply keyword extraction, topic mining, text classification and other methods to the quantitative study of policy text. However, there are relatively few such studies in the area of digital economic policy. In addition, in the field of digital economic policy, the accurate and efficient discovery of new words in the field can promote the deep application of natural language processing technology in policy text quantification and analysis, make up for the inefficiency of manual analysis, and thus achieve intelligent management of policy data. Therefore, this paper constructs a thesaurus of digital economic policy domain through new word discovery algorithm and a few manual supplementary methods, verifies its application effect in policy text segmentation, text classification, keyword extraction and theme modeling technology, and makes a quantitative study of digital economic policy text, which is of great research value.

3 Lexicon Construction of Digital Economic Policy

3.1 Data Sources

In order to comprehensively discover the professional vocabulary in the field of digital economy policy, this study searched and collected the development plans, action plans and other policy documents related to the digital economy in recent years in the official websites of various departments and governments of the country and 31 provinces (cities, autonomous regions), and constructed a corpus of new words discovery policy. The policy selection includes the following principles: First, the policy title must contain the keywords "digital economy", "data", "digital" and "intelligence"; Second, announcement, solicitation, application, interpretation and other types of policy documents are not included in the scope of collection; Third, the policy content is consistent with the theme of digital economy development and the text is complete. After screening, 34 China's policy documents and 450 provincial and municipal policy documents were finally obtained, covering the period from 2015 to 2022.

3.2 Neologism Discovery

New word discovery is the process of analyzing the text in a certain field to obtain words or phrases related to that field. In order to accurately identify the new words and technical terms in the field of digital economy policy text, this study constructs the domain lexicon of digital economy policy through four steps. 1. Carry out n-gram segmentation of policy text, generate all possible word fragments, and count the frequency of each segment; 2. Calculate the solidification degree of each segment according to the frequency [8], set different thresholds for segments of different lengths, and discard certain segments according to the thresholds; 3. Use the left word fragments to segment the text again, count the frequency, and calculate the boundary entropy. 4. Get the candidate word set by setting the threshold and sorting, and use rule filtering to assist manual completion to get the thesaurus of digital economic policy field.

In order to ensure the speed of the experiment, the calculation code of the freezing degree and boundary entropy of the new word discovery algorithm refers to the Github open source algorithm [15]. By comparing the screening results of different threshold combinations in the second step, the useful word fragments are retained as much as possible, and the occurrence of meaningless word fragments is reduced. Finally, 12061 words fragments with threshold greater than 0 are screened in the fourth step. According to the rules, the word fragments containing stop words are eliminated, leaving 9623 words. In addition, in order to ensure the professionalism and comprehensiveness of the obtained thesaurus, the study also added 406 words manually according to the National "Fourteenth Five-Year Plan" for Digital Economy Development (hereinafter referred to as the National Plan), and finally the thesaurus in the field of digital economy policy was obtained, totaling 10026 words, including the words in the field of digital economy, such as intelligent manufacturing, digital trade, data resources, digital technology, digital government construction, etc.; And policy terms, such as pilot demonstration, typical application, efficient allocation, innovation guidance, etc.

3.3 Comparison of Segmentation Effect

Since the subsequent research processes of policy text keyword extraction, text classification and topic modeling are based on the results of text segmentation, the accuracy of word segmentation will have a direct impact on the results of the entire study. In order to test the application effect of thesaurus in word segmentation, this study selects the precise mode of the most commonly used jieba word segmentation tool to segment the digital economic policy text. In order to comprehensively evaluate the effect of domain thesaurus in digital economic policy text segmentation, this study takes recall rate (R), precision rate (P) and F1 score as evaluation indicators, where R is the percentage of the correct number of words in algorithm segmentation and the number of words in manual segmentation; P is the percentage between the correct number of words in the algorithm

Table 1. Comparison of Text Segmentation Effects (%)

	R	P	F1
No lexicon added	69.90	48.91	57.40
Add lexicon	**81.58**	**73.40**	**77.09**

and the number of words in the algorithm; F1 score is the harmonic average of recall rate and accuracy rate, and the formula for calculating F1 score is shown in Formula (1).

$$F1 = \frac{2PR}{(P + R)} \tag{1}$$

Taking 50 randomly selected policy paragraphs as the test object, after comparing the segmentation results of artificial experience knowledge, the segmentation results without adding new words and the segmentation results with adding new words are shown in Table 1. According to the results in Table 1, it can be seen that the effect of text segmentation has been improved significantly after adding the domain lexicon, and its recall rate, accuracy rate and F1 score have increased by 11.68%, 24.49% and 19.69% respectively.

4 Classification of Policy Text

4.1 Experimental Data

Automatic classification of policy text is the important links to realize intelligent analysis of policy data. In order to automatically identify the development task of the digital economy policy text, this study selected 42 digital economy planning policy documents, developed 8 classification indicators (digital infrastructure, data elements, industrial digital transformation, digital industrialization, digital public services, digital economic governance system, digital economic security system, international cooperation and regional cooperation), and marked a total of 2519 policy paragraphs.

4.2 Model and Parameter

In the experiment, three common neural network models are selected for text classification effect comparison, including TextCNN and LSTM_ Attention, GRU_ Attention. Randomly select 80% of the policy text as the training set, and the other 20% as the test set. Using word2vec to train the text word vector, select the jieba word segmentation tool and set whether to add the domain word library when the policy text word segmentation. The loss function uses cross entropy loss function, and the activation function uses softmax. Other experimental parameter settings are shown in Table 2.

Table 2. Parameter Setting

Parameter	Value
embedding size	300
train_epochs	20
batch_size	32
learning_rate	5×10^{-5}

Table 3. Experimental Result (%)

Model		Accuracy	F1
TextCNN	No thesaurus added	78.96	78.88
	Add thesaurus	**81.74**	**80.75**
LSTM_ Attention	No thesaurus added	85.11	84.37
	Add thesaurus	**85.91**	**84.75**
GRU_ Attention	No thesaurus added	85.25	85.51
	Add thesaurus	**85.75**	**86.11**

4.3 Experimental Result

Compare the automatic classification results with the manual labeling results, and calculate the accuracy and F1 value. The results are shown in Table 3. The results show that different neural network models have different effects on text classification. In general, GRU_Attention model has the best classification effect, while TextCNN model has the worst classification effect. In addition, after comparison, adding new words in the field will help improve the classification accuracy. This improvement is most obvious in the TextCNN model, and the accuracy rate of automatic classification has increased from 78.96% to 81.74%. The improvement effect on the other two models is small. The reason may be that many of the same policy expressions appear in similar or different words in different policy texts, but these words have differences after quantitative expression, such as "pension" and "smart pension"; "Internet Hospital" and "Smart Hospital". In the case of a small amount of data, such differences make the effect of text classification not significantly improved. However, on the whole, it can be seen that adding thesaurus is still helpful to improve the effect of text classification.

5 Analysis of Policy Hotspots

5.1 Policy Keywords

In order to further explore the more fine-grained hotspots in the eight key tasks of the National Plan, the TF-IDF keyword extraction technology was adopted in this study to extract the top 20 keywords of the eight parts respectively, as shown in Table 4.

Table 4. TOP 20 KEYWORS OF NATIONAL PLAN

1	Infrastructure\Gigabit optical network\6G\Data center\Application\Upgrade\Network\Network infrastructure\Spatial data infrastructure\Cloud network collaboration\Arithmetical power\Optical network\Intelligent scheduling\5G network\IPv6 transformation\Coordination\Integrated\Artificial intelligence\Intelligence\Transmission network
2	Data\Data resources\Data trading platform\Data asset evaluation\Data service\Market subject\Data element\Data assets\Data transaction\Data value\Data security\Government data\Data element market\Pricing mechanism\Deal matching\Data development and utilization\City data\Digital technique\System\Data management
3	Digital transformation\Service\Intelligent manufacturing\Industry\Industrial cluster\Digital Transformation Promotion Center\Transformation\Innovate\Application\Digital level\Industrial Park\Digitization\All links\Pilot demonstration\Digital technique\Dig data\Small and medium-sized enterprises\Digital solution provider\Logistics\public service
4	Innovate\Platform\Service\share\Digital technique\Industrialization\Industrial innovation\New type\Advantage\Resource sharing\Artificial intelligence\Supply chain\Integrate\Application\On-line\Quantum information\Big data\Blockchain\Convergence application\Platform enterprise
5	Service\New smart city\Digitization\Application\Public service\Smart city\Data sharing\Urban and rural\Big data\Contingency management\Medical health\Service capability\Intelligence\Community service\Resources\Service level\Education\Accurate\Operate\Intelligence
6	Supervise\Platform\Mechanism\Statistical monitoring\Big data\Digital economic governance\Artificial intelligence\Market subject\Cross-departmental\government\Operator\standard\Digital economy\Innovative vitality\Digital governance\Blockchain\Problem study and judgment\Risk warning\Industry self-control\Core industries of digital economy
7	Network security\Security\Data security\System\Risk\standard\Data security protection\Network security protection capability\Industrial system\According to law and regulations\Prevent various risks\Flexible employees\Personal information\Safe and reliable\Facilities\Administration\Transmission\Information\Technology\Originality
8	International co-operation\Cross-border e-commerce\Cyberspace\Trade digitization\Restricted Lane\Partnership\Openness\ASEAN\Bilateral\Trade\Cross-border\Service industry\Introduction\Space sovereignty\Govern\International\Rule\Ecosphere\Digital trade

Based on the statistical analysis of key words, the following hot spots are found in the Plan: Firstly, on the basis of basic support, it is required to focus on building the infrastructure base of the digital economy in the aspects of information network infrastructure, cloud network collaboration and computer network integration, and take data elements as important production elements of the digital economy, and attach importance to the systematic management of various data element types in the links of transaction, pricing, development and utilization. Secondly, on the main task, the Plan emphasizes the "two-wheel drive" of industrial digital transformation and digital industrialization, give full play to the role of third-party service-oriented institutions, promote the innovative application and pilot demonstration of digital technology in all aspects of digital transformation, and promote the construction of industrial clusters and industrial parks. In terms of digital industrialization, it requires the rapid development of digital industries such as artificial intelligence, quantum information, big data and blockchain. Thirdly, in terms of inclusive sharing, promote the construction of a new smart city, the integration of digital urban and rural development, and promote the development of the digital economy towards inclusive and convenient development in emergency management, medical health, community services, smart education and other aspects; We will work hard to promote international cooperation in the digital economy, digital trade and multilateral governance. Fourthly, the policy requires that the two systems of digital economic governance and digital economic security be coordinated, the regulatory mechanism be strengthened, the government' digital governance capacity be improved, and the pluralistic governance be promoted. We should strengthen network security and data security, and prevent and resolve various major risks. Lastly, according to the observation of key words, there are many related words such as "application", "sharing", "service", "standardization" and "governance", highlighting the keynote of deepening application, standardized development and inclusive sharing in the development of China's digital economy.

5.2 Policy Topics

The policy topics are the core content of policy text and the focus of policy research. LDA theme model is one of the common theme modeling algorithms. It regards each policy as a mixed Dirichlet distribution of several potential topics, which can effectively aggregate the theme characteristics of the text, and present the "topic-theme words" of each policy in the form of probability distribution, without strict restrictions on the length of the text. Therefore, this study takes the policy text of the digital industrialization part with the largest number of data as an example, and selects 482 provincial policy text classification data. After many experiments, when the number of topics is set to 6, the modeling effect is the best. The first 10 theme words of each theme are shown in Table 5.

Summarize and analyze the subject words of different subjects as follows: Topic 1 can be summarized as key technologies. It covers the research and development of key technologies such as "big data", "cloud computing", "quantum communication" and "artificial intelligence". The policy requires accelerating the development of innovative technology towards high performance, intelligence and industrialization. Topic 2 and Topic 4 can be summarized as emerging digital industries. It mainly covers key industries such as "integrated circuit", "cloud computing", "big data", "5G industry", "industrial software", "artificial intelligence", and "new display". The policy requires to promote the forward-looking layout of digital industry and accelerate the expansion of industrial clusters through the construction of industrial parks and science parks. Topic 3 can be summed up as a new business model. The policy requires deepening the integrated application of "sharing economy" and "platform economy" in the field of life

Table 5. Topics of Digital Industrialization

Topic	Topic Words
1	Big data\Industrialization\Cloud computing\Key technology\ Quantum communication\Smart City\High performance\ Intelligent\Artificial intelligence\Technology R&D
2	Integrated circuit\Cloud computing\Big data\Focus areas\5G Industry\Digital technique\Industrial software\Artificial Intelligence\Information technology\Intelligent manufacturing
3	Blockchain\Platform economy\Convergence application\ Application scenario\New formats and models\Digital industry\ New model\Sharing economy\Big data\Supply chain
4	New display\Industrial cluster\Industrial park\Key Enterprises\ Science and Technology Park\Industrialization\Integrated circuit industry\OLED\High quality development\Industrial chain
5	Internet of Things\Intelligent terminal\Artificial intelligence\ Sensor\robot\intelligent sensor\Intelligent connected vehicle\ Industrialization\Edge computing\manufacturing
6	Artificial intelligence\Technological innovation\Industrial Innovation\laboratory\Platform construction\Development pilot area\technological innovation\New generation of artificial intelligence\Research institute\Application demonstration

services, and expanding the application scenarios of new formats and new models. Topic 5 can be summarized as key products. It mainly covers key products such as “Internet of Things”, “intelligent terminals”, “artificial intelligence”, “sensors”, “robots”, and “intelligent networked vehicles”. The policy requires to promote the industrialization and application of key software and hardware products. Topic 6 can be summarized as innovation platform system. The policy requires that based on the needs of “technological innovation” and “industrial innovation”, the innovation platform system mainly consisting of the new generation of information technology “laboratory”, “research institute” and “development pilot zone” should be gradually improved.

6 Conclusion

This This paper constructs a professional Thesaurus of digital economic policy through the new word discovery algorithm, and illustrates the feasibility and necessity of domain thesaurus in the natural language processing of policy text by comparing the effects of text segmentation. In addition, this paper makes use of the thesaurus to conduct quantitative research on the text classification, hot spot analysis and theme modeling of China’s digital economic policies during the 14th Five-Year Plan, so as to expand the depth and breadth of the research. We summarize and analyze the hot spots and key construction areas of China’s digital economic development, as well as the five major thematic features of the text of the digital industrialization policy. The professional thesaurus built in this paper makes the results of quantitative research on policy texts intuitive and prominent, and is conducive to fine-grained analysis. It can meet the depth and speed requirements of large-scale policy text content analysis using natural language processing technology in the era of big data. However, there are still some deficiencies in this study. Due to the small size of the policy corpus, the effectiveness of the professional lexicon in policy text segmentation and text classification still needs to be improved. In the future, more models can be used to discover new words and to improve the feature input of word granularity in pre-training language models such as BERT, so as to make text classification more effective and promote the intelligent management of digital economic policy text.

Acknowledgment. This research was supported by the Shandong University Youth Innovation Technology Support Program "Industrial Reformation in the Intelligent Era: Technology, System and Entrepreneurship Orientation" (Project No.: 2020RWG009), Shandong Social Science Plan Digital Shandong Research Special Project "Research on Enterprise Digital Transformation Strategy and Implementation Mechanism" (Project No.: 21CSDJ01).

References

1. C. C. Li, G. Liu, "On Strategies for Developing Digital Economy in Developed Economies and Their Enlightenment for China," Contemporary Economic Management, vol. 44, No. 4, pp. 9–15, 2022.
2. J. Wu, F. Zhang, "An analysis of the trend of digital economy in foreign countries and China's countermeasures from the perspective of think tank," Scientific Research Management, "vol. 43, No. 8, pp. 32–39, 2022.
3. B. Shi, Q. Chang and L. Y. Zhang, "Policy Evolution and Theoretical Research on the Development of China's Digital Economy," Technical Economy, vol. 41, No. 8, pp. 1–10, 2022.
4. J. Jiang, Y. X. Jiang and H. Du, "Study on digital government from the perspective of policy tools—Based on quantitative analysis of 45 policy texts in China," Think Tank Theory and Practice, vol. 7, No. 2, pp. 14–23, 2022.
5. H. Z. Lei, Q. Wang. "Quantitative research on the text of China's local government's digital economy policy," Technical Economy and Management Research, No. 5, pp. 91–94, 2022.
6. Q. Y. Yang, Y. Y. Qiao and S. L. Liang, "Research on digital economic policy of provincial government based on compatibility of policy goals and policy tools," Economic System Reform, No. 03, pp. 193–200, 2021.
7. Z. J. Wang, S. Chang, L. Zhou, P. K. Guo and M. F. Gu, "Development of environmental management lexicon based on new word discovery and its empirical application," Journal of Environmental Engineering Technology, vol. 11, No. 2, pp. 385–392, 2021.
8. Y. Z. Wang, B. L. Hua, " Research on topic modeling of policies relate to disruptive technologies in European and American countries based on official website text data," Information Theory and Practice, vol. 45, No. 6, pp. 39–47, 2022.
9. Osnabrügge, Moritz, E. Ash, and M. Morelli, "Cross-Domain Topic Classification for Political Texts," Political Analysis, pp. 1–22, 2021.
10. T. Zhang, H. Q. Ma and Y. Yi, "Comparative Analysis of China's Big Data Policies from the Perspective of Text Similarity " Library and Information Work, vol. 64, No. 12, pp. 26–37, 2020.

11. J. M. Hu, W. L. Fu, W. Qian and P. L. Tian, “Research on policy text classification model based on topic model and attention mechanism,” Information Theory and Practice, vol. 44, No. 7, pp. 159–165, 2021.
12. W. Shao, B. L. Hua, “Unsupervised construction of thesaurus in the science and technology policy based on dependency syntax analysis,” Information Engineering, vol. 6, No. 6, pp. 33–44, 2020.
13. X. M. Zheng, Y. Dong, “Constructing Degree Lexicon for STI Policy Texts,” Data Analysis and Knowledge Discovery, vol.5 ,No. 10, pp. 81–83, 2021.
14. Y. Wan, M. H. Zhang and J. P. Gao, “Research on policies related to Chinese artificial intelligence based on text analysis,” vol. 6, No. 12, pp. 54–63, 2021.
15. https://github.com/Chuanyunux/Chinese-NewWordRecognition#chinese-newwordrecognition.