Topic Mining of Economic and Trade Cooperation between Guangxi Province and ASEAN Countries based on LDA Model*

Chunlan Li, Chongjian Song*

School of Economic and Management Guangxi Normal University, Guilin, China

lcl_lee@126.com,scj199700@163.com*

Abstract. The LDA Model is a topic model commonly used in natural language processing. It mines the topic in three steps. First, calculates the topics of each document in the document set in the form of probability distribution. Second, extracts the actual topics after analyzing those documents in the same subject. Third, mines those factual topics by topic clustering or text classifying. This paper uses the LDA model to analyze the articles and news texts published on the topic of Economic and Trade exchanges between Guangxi Province and ASEAN Countries on the website of the People's Government of Guangxi Zhuang Autonomous Region. It summarizes 11 topics of economic and trade activities between the two sides from 2017 to 2021. It calculates topic intensity and discovers the cooperative industries with high frequency. It provides a reference direction for further promoting the cooperation and development of the two sides.

Keywords: LDA Model; Guangxi Province and ASEAN Countries; Economic and Trade Cooperation; Topic Mining

1 Introduction

Over the years, Guangxi has insisted on promoting the construction of China (Guangxi) Pilot Free Trade Zone, the construction of new land and sea corridors in the west, and the formation of an important gateway for the organic convergence of the Belt and Road, which is highly consistent with the proposal of General Secretary Xi Jinping in the report of the Twentieth National Congress to "promote the construction of the Belt and Road Initiative, raise the level of opening up of the central and western parts of the country, accelerate the construction of the new land and sea corridors in the west, and implement the strategy of upgrading the Pilot Free Trade Zone". Guangxi has demonstrated a high level of opening up to the outside world. Guangxi has the unique geographical position of being connected to Southeast Asian countries by sea and land and

* Chunlan Li is a professor of the School of Economics and Management at Guangxi Normal University, with research interests in enterprise information resource management.
Chongjian Song is a postgraduate student of the School of Economics and Management at Guangxi Normal University, with research interests in business management.

© The Author(s) 2023
B. K. Kandel et al. (eds.), *Proceedings of the 2023 3rd International Conference on Business Administration and Data Science (BADS 2023)*, Atlantis Highlights in Computer Sciences 19,
https://doi.org/10.2991/978-94-6463-326-9_13

thus has a unique advantage in the development of economic cooperation and trade relations with Southeast Asian countries. ASEAN has maintained its position as Guangxi's top trading partner for 21 consecutive years since 2000 [1], and ASEAN will also become China's largest trading partner in 2020. Win-win cooperation between Guangxi and Southeast Asian countries is a very important part of China's basic national policy of opening up to the outside world and is one of the important paths to building a new type of state relationship.

2 Literature review

2.1 Guangxi province and ASEAN countries

As early as 25 November 2000, Premier Zhu Rongji first proposed establishing a China-ASEAN free trade area at the fourth China-ASEAN "10+1" Leaders' Meeting, which was positively responded to by the leaders of various countries. At the same time, many scholars have considered and constructed the direction and trend of economic and trade cooperation between China and ASEAN under this concept [2]. The China-ASEAN Free Trade Area (CAFTA) was formally established on January 1, 2010, after 10 years of efforts by both sides [3]. Due to its unique geographic location and strategic positioning, the Guangxi region is a bridgehead and frontier for the construction of CAFTA [4]. In September-October 2013, during his visit to Central Asia and Southeast Asia, General Secretary Xi Jinping put forward the major initiatives of the "Silk Road Economic Belt" and the "21st Century Maritime Silk Road." [5] As Guangxi is located in the organic convergence zone of the "Belt and Road" initiative, economic and trade cooperation between Guangxi and ASEAN countries has also become an important part of Guangxi's active integration into the construction of the "Belt and Road". Many scholars have studied China and ASEAN countries [6-8] under this perspective and the Guangxi region with ASEAN countries economic and trade cooperation between China and ASEAN countries. On August 30, 2019, the official operation of China (Guangxi) Pilot Free Trade Zone, which aims to build a "high-standard and high-quality free trade park leading China-ASEAN opening up and cooperation", marks a higher level of economic and trade cooperation between the Guangxi region and ASEAN countries. In addition, the RCEP agreement involving 15 countries, which was officially signed on November 15, 2020, has led to the further development of cooperation between China and ASEAN countries, marking the official launch of the world's most populous free trade area with the largest economic and trade scale and the greatest potential for development. The signing of this agreement has also led scholars to re-examine the challenges and opportunities in the issue of economic and trade cooperation between Guangxi and ASEAN countries. By January 1, 2022, the formal entry into force of the RCEP agreement [9], the rest of China and non-ASEAN member countries will have unprecedented attitudes and measures to deal with economic and trade cooperation activities with ASEAN countries, Guangxi region, and ASEAN countries' economic and trade cooperation activities will be impacted and challenged.

To summarize, at the present stage, the economic and trade activities between Guangxi and ASEAN countries are carried out under the establishment of the China-

ASEAN Free Trade Area, the initiative of building the " the Belt and Road", the official operation of China (Guangxi) Pilot Free Trade Zone and the entry into force of the RCEP agreement. National strategic planning for the Guangxi region and the entry into force of international cooperation agreements have significantly impacted the current study of the subject of economic and trade activities between the two.

2.2 Policy text analysis

Scholars have begun to use data analysis methods to study government policy and regional development news texts. Qiu et al. [10] selected the policy texts related to cross-border e-commerce for LDA topic analysis based on the background of China's "double-cycle" policy. The relevant topics of the policy texts are extracted, and the importance of the topics is determined according to the topic strength, so as to establish a game model between the government and the cross-border electronic commerce. Li et al. [11] used the LDA topic model to extract the topic contents of the policy texts on new energy vehicles in China, and then argued the influence mechanism of these topic contents on the development of new energy vehicles through empirical analysis. Bing et al. [12] took the provincial intellectual property policy text as the research object, from which they analyzed the degree of importance attached to intellectual property rights and the degree of completion of supportive policies in the implementation of intellectual property strategy in the provinces and regions; Bing et al.[13] used LDA topic modeling to capture four topic contents of environmental policy texts issued by four urban agglomerations in East China as the four independent variables for the subsequent fsQCA analysis. These results were used to explore the mechanisms by which environmental policies influence regional eco-efficiency.

Concerning the cooperative relationship between Guangxi and ASEAN, researchers have analyzed certain aspects at the micro level through various reports and structured data. However, there are few macro-level analyses of policy texts or cooperation texts to summarize the key points of the cooperation between the two sides, and few analyses use big data methods in the research process.

3 Research design

3.1 Data sources

Between 2017 and 2021, 544 news and policy texts were published in the Economic and Trade Exchanges topic of the Guangxi and ASEAN section of the official website of the Guangxi Zhuang Autonomous Region People's Government. Excluding six of the video news, a total of 538 news and policy texts on economic and trade cooperation between Guangxi and ASEAN were collected in this paper.

3.2 Research process

The research process included three broad steps: data collection, data preprocessing, and LDA topic modeling. The data collection step completes the collection of the research content; the data preprocessing step removes the text that affects the results of the research; and the LDA topic modeling step presents the underlying topics of the research content. The specific process and introduction included in the three processes are shown in Table 1.

Table 1. Specific process and introduction

	Process	Introduction
Data collection	Crawling texts	Save research content locally in large batches
Data preprocessing	Text preprocessing	Delete fixed-format text in bulk using "find and replace".
	Add user-defined dictionaries	Ensure that proper nouns consisting of multiple words are not automatically broken up by the Jieba library.
	Segmentation	Split all coherent sentences in the text into separate words for further analysis by the software.
	Rejecting stop words	Remove words from the text that do not contribute to the purpose of the study.
Latent Dirichlet allocation model	Topic consistency calculation	Calculate the topic consistency of the LDA model to determine the number of topics in the LDA model.
	Optimal number of topics	Set the number of topics that maximize topic consistency
	Extracting topics	Run the LDA theme model to extract the theme information of the text.
	LDA topic visualization	Visualize the results of extracting themes from the LDA theme model.

Data collection.

Through Octopus V8 collector, relevant content of government websites was collected locally in large batches to get the research content for subsequent text analysis.

Data preprocessing.

The content of the collected textual information contains fixed formats (reporter, editor, source of the article, etc.), insubstantial words (of, all, with, between, etc.), and intonational auxiliaries (ah, oh, what about, etc.) that interfere with the process of textual analysis. Therefore, maximizing the removal of these words can help the text analysis process to be more accurate and scientific. The steps of manual preprocessing, adding user-defined dictionaries, word splitting, and removing deactivated words were used in this research to assist in removing these words that interfere with the analysis process. Among them, manual preprocessing refers to the use of Microsoft Word 2016 software's Find and Replace function to delete and change the collected text in large quantities; adding user-defined dictionaries to assist the software to more accurately divide the overall text process, to ensure that the proper nouns are not jieba library is cut into several words; participle is a prerequisite for the analysis of the text of the Chinese context, the content of the text into paragraphs and sentences, the participle will be used in the analysis process. Segmentation of the text into paragraphs and sentences into words, to continue text analysis. However, because the Chinese text, in addition to punctuation, there is no split sign between the words, so we have to use the Python jieba library for Chinese word separation.

Latent Dirichlet allocation model.

LDA (Latent Dirichlet Allocation), proposed by Blei D M et al. in 2003[14], is a topic model that gives the topic of each document in a collection of documents in the form of a probability distribution. LDA is the conjugate prior probability distribution of the polynomial distribution. Given a corpus D of M documents, where document d has N_d words d（$d \in 1, \dots, M$）. LDA models D according to the following generative process:

(1) Select the multinomial distribution φ_t of subject t（$t \in 1, \dots, T$） from the Dirichlet with hyperparameter β;

(2) Select the multinomial distribution θ_d of document d（$d \in 1, \dots, M$） from the Dirichlet with hyperparameter α;

(3) For a word W_n（$n \in 1, \dots, N_d$） in a document d,

a. Select topic Z_n from θ_d.

b. Select a word W_n from φ_{zn}.

In the above generation process, the words in the document knowledge observed variables, while the others are latent variables （φ and θ） and hyperparameters (β and α).

The probability of observing the document set D is derived from the corpus computation. The formula is as follows:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_{dn}|\theta_d) p(w_{dn}|Z_{dn},\beta)\right) d\theta_d \quad (1)$$

In the equation(1), α, β are the hyperparameters that need to be determined. D represents the test document set, M represents the number of texts. θ_d represents the topic distribution of document d, while w_{dn} refers to the nth word in the dth document. Z_{dn} indicates the topic of the nth word in the dth document [14].

The key aspect of the LDA topic model lies in the setting of the number of topics, denoted as k. Common methods for determining the value of k include heuristic empirical setting, perplexity [15], and topic coherence analysis [16]. In text analysis research utilizing the LDA method, a significant proportion of studies calculate perplexity to determine the value of k. However, some studies argue that the topic coherence score is the optimal method for determining the number of topics in LDA, and there has been an increase in the use of this method in recent years [17]. Therefore, this paper adopts the Topic Coherence Score to determine the value of k. After conducting repeated experiments, as shown in Fig. 1, the model achieves the best fit when the number of topics, k, is set to 11. Hence, in running the LDA algorithm, this paper sets the number of topics to 11. To display the top keywords for each topic more effectively, the desired number of displayed high-frequency keywords is set to 10. The setting of the two prior parameters, α and β, follows the research findings of Maier et al. [18], where α is set to 0.5, and β is set to 1/k = 0.09.

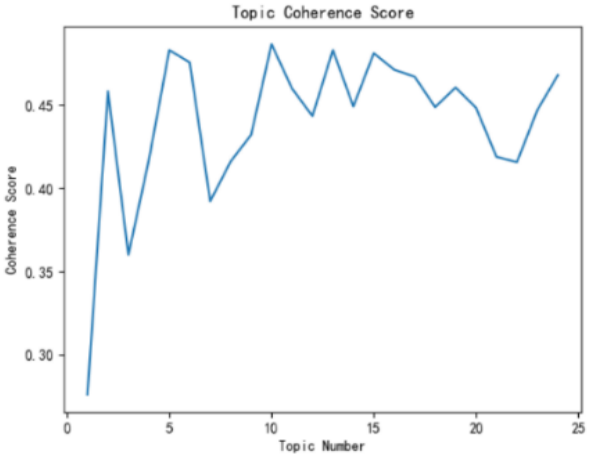

Fig. 1. Experimental plot of topic coherence scores

4 Result

4.1 The result of the Latent Dirichlet Allocation mode

The 11 topics obtained in this study, along with the keywords contained within each topic, are presented in Table 2. The high-frequency keywords generated by the LDA topic model are listed without specific topic names but are arranged in order of their relevance within the topic. Therefore, the "Topic" column in the table is assigned based on the logical relationships among the keywords to provide a comprehensive topic name. Furthermore, there is considerable disparity in the relevance scores of keywords within the same topic. For instance, in the "Information Port and Financial Development" topic, the keyword "construction" has a relevance score of 0.044, while "experimental zone" has a relevance score of 0.010. Considering the relevance of keywords, the naming of each topic is primarily determined by the top five keywords with the highest relevance within the topic.

Table 2. Table of LDA themes for cooperation between Guangxi and ASEAN countries

Number	Topic	High Frequency Keywords
Topic1	Information Harbor and Financial Development	construction, development, finance, China-ASEAN Information Harbor, opening-up, cross-border, cooperation, economy, innovation, pilot zone
Topic2	Bilateral Investment and Belt and Road Construction	cooperation, development, enterprise, investment, expo, exchange, belt and road, international, construction, economy
Topic3	Import and Export of Goods	port, Vietnam, Pingxiang, Dongxing, fruit, customs clearance, import, customs, goods, Youyi Pass
Topic4	Agricultural Cooperation	Agriculture, Vietnam, business, investment, Cambodia, cooperation, Myanmar, Indonesia, Brunei, Laos
Topic5	Foreign Trade	E-commerce, trade, growth, cross-border, import and export, enterprises, RMB, export, settlement, market
Topic6	Economic and Trade Cooperation with Malaysia	Malaysia, logistics, two countries, enterprises, Singapore, liner, China-Malaysia Qinzhou Industrial Park, dual park, international, seafood
Topic7	International Maritime Transportation	Construction, international, Qinzhou, Beibu Gulf, channel, bonded, Fangchenggang, port, route, Belt and Road
Topic8	Cooperation with Mekong Basin Countries	Cooperation, Laos, Lan Mekong, Creativity, Nanning, Thailand, Cambodia, Vietnam, International, Two Countries
Topic9	Economic and Trade Cooperation with Thailand	Thailand, market, commodity, Malaysia, exhibition, durian, product, group, exhibitor, enterprise
Topic10	Culture and Tourism	Culture, tourism, education, bridge, fund, students, animation, villagers, poverty alleviation, experts
Topic11	Bilateral Trade in Specialty Products	Tea, Henan, mahogany, Laos, specialty, ancient tree, special train, travelers, liaison office, tea

4.2 The result of the pyLDAvis visualization

Based on the obtained 11 topics, this paper utilizes pyLDAvis [19] to visualize the relationships between keywords and topics, as well as the interrelationships among the topics, as shown in Fig. 2.

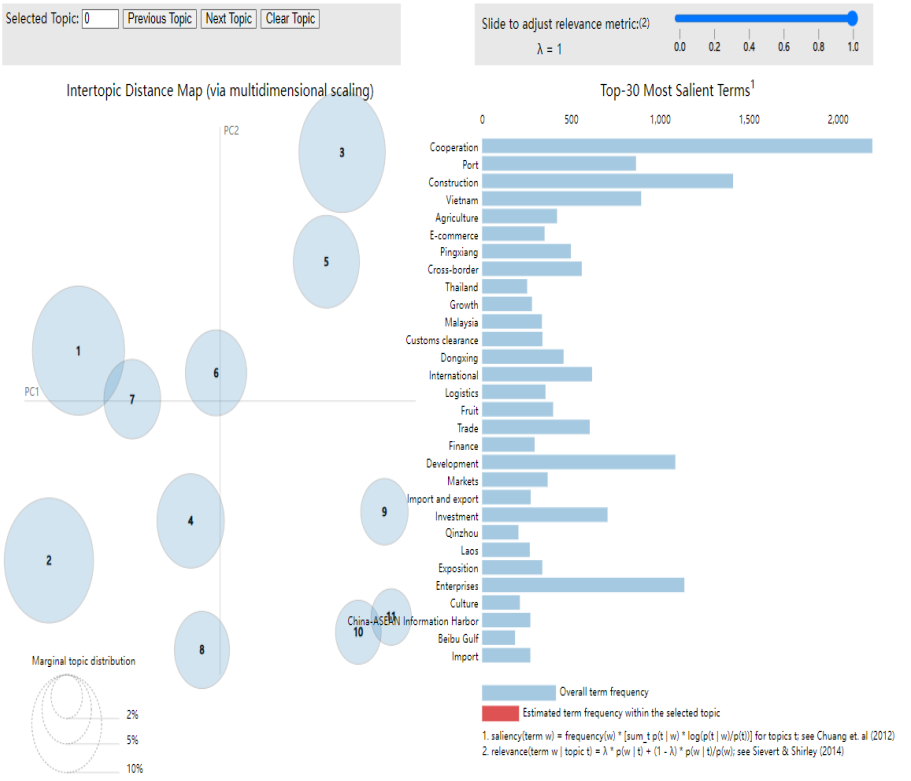

Fig. 2. LDA topic visualization of cooperation between Guangxi and ASEAN countries

In Fig. 2, the 11 shaded circles in the left area represent the 11 topics, and the numbers inside the circles correspond to the topic numbers in Table 2. The larger the area of the circle, the more texts are associated with that topic, indicating a greater influence. The distance between circles represents the dissimilarity between topics, with greater distance indicating lower similarity and vice versa. When one shaded circle is selected, the bar chart on the right displays the top 30 most relevant words within that topic and their respective influence scores, as illustrated in Fig. 3. In the bar chart on the right, the blue bars represent the proportion of each keyword in the overall text, while the red bars represent the relevance of the keyword to the corresponding topic.

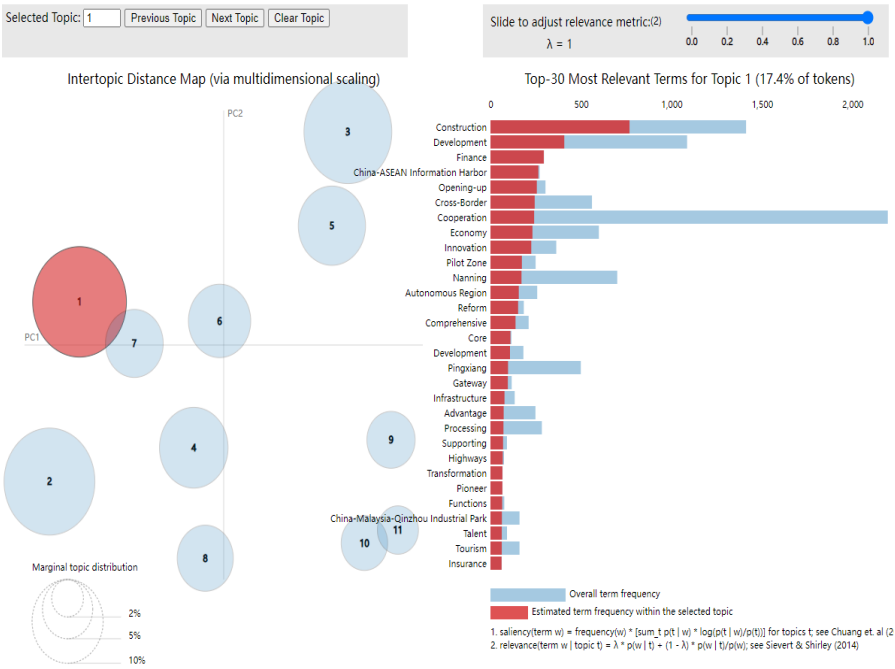

Fig. 3. Visualization of the interactive interface for selected Topic1

4.3 Topic intensity

The topic intensity is an indicator that describes the attention given to a particular topic in the overall document. The higher the number of documents related to a topic, the greater its topic intensity, indicating that the topic is more likely to be a hot topic. The formula for calculating it is as follows:

$$P_k = \frac{\sum_i^N \theta_{ki}}{N} \tag{2}$$

In the equation(2), P_k represents the topic intensity of the kth topic; θ_{ki} represents the probability of the kth topic word in the ith document; N represents the number of topic words under the topic [21]. The LDA topic intensity can be obtained using the equation(2) and is shown in Table 3.

Table 3. Topic Intensity

Topic clustering	Number	Topic	Topic intensity
Focus Areas	Topic1	Information Port and Financial Development	0.174
	Topic2	Bilateral Investment and Belt and Road Construction	0.163
	Topic3	Import and Export of Goods	0.153
Hot Areas	Topic4	Agricultural Cooperation	0.093
	Topic5	Foreign Trade	0.090
	Topic6	Economic and Trade Cooperation with Malaysia	0.077
	Topic7	International Maritime Transportation	0.066
	Topic8	Cooperation with Mekong Basin Countries	0.062
General	Topic9	Economic and Trade Cooperation with Thailand	0.046
	Topic10	Culture and Tourism	0.043
	Topic11	Bilateral Trade in Specialty Products	0.033

4.4 Research result

The LDA topic visualization results for the cooperation between Guangxi and ASEAN countries, as constructed in this study (see Fig. 2), reveal that, apart from minor overlaps between Topic 1 and Topic 7, as well as between Topic 10 and Topic 11, there is a significant separation between the distributions of other topics. Overall, the level of overlap between the topics is minimal, indicating a reasonable choice of the number of topics [20].

Through the analysis of textual data on economic and trade activities between Guangxi and ASEAN countries from 2017 to 2021 using the LDA topic model, it is evident that there is close cooperation in various aspects, such as Information Port and Financial Development, Bilateral Investment and "Belt and Road" Construction, Import-Export of Goods, Agricultural Cooperation, Foreign Trade, Economic and Trade Cooperation with Malaysia, International Shipping, Cooperation with "Lancang-Mekong" River Basin Countries, Economic and Trade Cooperation with Thailand, Culture and Tourism, and Bilateral Specialty Trade. However, the degree of cooperation varies across different areas. Based on the topic strength information, it can be observed that the cooperation in Information Port and Financial Development, Bilateral Investment and "Belt and Road" Construction, and Import-Export of Goods are the most significant, representing the focal areas of economic and trade cooperation between the two sides. Additionally, Agricultural Cooperation, Foreign Trade, Economic and Trade Cooperation with Malaysia, International Shipping, and Cooperation with "Lancang-Mekong" River Basin Countries show relatively less cooperation compared to the aforementioned three areas, indicating popular areas of economic and trade cooperation. Lastly, Economic and Trade Cooperation with Thailand, Culture and Tourism, and Bilateral Specialty Trade represent general areas of economic and trade cooperation between the two sides (as shown in Table 3).

5 Conclusion

This study collected textual information from the Website of the People's Government of Guangxi Zhuang Autonomous Region, specifically from the Guangxi-ASEAN column, covering the period from 2017 to 2021, regarding the economic and trade exchanges. The collected data underwent manual preprocessing, including the addition of user-defined dictionaries, tokenization, and removal of stopwords, to prepare it for analysis and processing with the LDA topic model. Subsequently, the LDA topic modeling technique was applied to these preprocessed texts, and topic coherence scores were computed to determine the optimal number of topics. After setting the number of topics to the optimal value, 11 topics related to Guangxi's economic and trade cooperation with ASEAN countries were obtained. Based on the formula for calculating topic strength, the topic strength for each of the 11 topics was determined. By comparing and clustering the topic strengths, three key areas were identified in the economic and trade activities between Guangxi and ASEAN countries: Information Port and Financial Development, Bilateral Investment and "Belt and Road" Construction, and Import-Export of Goods. Additionally, five popular areas were identified: Agricultural Cooperation,

Foreign Trade, Economic and Trade Cooperation with Malaysia, International Shipping, and Cooperation with "Lancang-Mekong" River Basin Countries. Lastly, three general areas were identified: Economic and Trade Cooperation with Thailand, Culture and Tourism, and Bilateral Specialty Trade.

ACKNOWLEDGMENTS

This paper is supported by the funding of Natural Science Foundation of China (No:71663009) and Graduate Student Innovation Program, School of Economics and Management, Guangxi Normal University (No: JG2023024).

REFERENCES

1. YI Y, GENG Y, YANG M. Has China-ASEAN Trade opening increased China's carbon emissions? [J]. Chinese Journal of Population, Resources and Environment, 2023, 21(2): 52-9.
2. CHIRATHIVAT S. ASEAN–China Free Trade Area: background, implications and future development [J]. Journal of Asian Economics, 2002, 13(5): 671-86.
3. PARK D, PARK I, ESTRADA G E B. Prospects for ASEAN-China Free Trade Area: A Qualitative and Quantitative Analysis [J]. China & World Economy, 2009, 17(4): 104-20.
4. ARIYASAJJAKORN D, GANDER J P, RATANAKOMUT S, et al. ASEAN FTA, distribution of income, and globalization [J]. Journal of Asian Economics, 2009, 20(3): 327-35.
5. TIAN X, SARKIS J, CHEN W, et al. Greening the Belt and Road Initiative: Evidence from emergy evaluation of China's provincial trade with ASEAN countries [J]. Fundamental Research, 2022.
6. BHOWMIK R, ZHU Y, GAO K. An analysis of trade cooperation: Central region in China and ASEAN [J]. PLOS ONE, 2021, 16(12): e0261270.
7. FOO N, LEAN H H, SALIM R. The impact of China's one belt one road initiative on international trade in the ASEAN region [J]. North American Journal of Economics and Finance, 2020, 54.
8. SONG A Y, FABINYI M. China's 21st century maritime silk road: Challenges and opportunities to coastal livelihoods in ASEAN countries [J]. Marine Policy, 2022, 136.
9. FENG T-T, GONG X-L, GUO Y-H, et al. Electricity cooperation strategy between China and ASEAN countries under 'The Belt and road' [J]. Energy Strategy Reviews, 2020, 30: 100512.
10. QIU Y, CHEN T, CAI J, et al. The Impact of Government Behavior on the Development of Cross-Border E-Commerce B2B Export Trading Enterprises Based on Evolutionary Game in the Context of "Dual-Cycle" Policy [J]. Journal of Theoretical and Applied Electronic Commerce Research, 2022, 17(4): 1741-68.
11. LI J J, JIAO J L, XU Y W, et al. Impact of the latent topics of policy documents on the promotion of new energy vehicles: Empirical evidence from Chinese cities [J]. Sustainable Production and Consumption, 2021, 28: 637-47.
12. BING S, MINGXING Y, ZAOLI Y, et al. An Algorithm Combining Latent Dirichlet Allocation and Bimodal Network for Evaluating Goal Deviation of Intellectual Property Strategy Execution in China [J]. Mathematical Problems in Engineering, 2020, 2020.

13. QIN M, SUN M, LI J. Impact of environmental regulation policy on ecological efficiency in four major urban agglomerations in eastern China [J]. Ecological Indicators, 2021, 130: 108002.
14. BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of machine learning research, 2003, 3(4/5).
15. ZHAO N, FAN G, QI Z, et al. Exploring the current situation of cultural tourism scenic spots based on LDA model——Take Nanjing, Jiangsu Province, China as an example [J]. Procedia Computer Science, 2023, 221: 826-32.
16. SHARMA A, RANA N P, NUNKOO R. Fifty years of information management research: A conceptual structure analysis using structural topic modeling [J]. International Journal of Information Management, 2021, 58.
17. O'CALLAGHAN D, GREENE D, CARTHY J, et al. An analysis of the coherence of descriptors in topic modeling [J]. Expert Systems with Applications, 2015, 42(13): 5645-57.
18. MAIER, WALDHERR, MILTNER, et al. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology [J]. Communication Methods and Measures, 2018, 12(2-3).
19. SIEVERT C S K. LDAvis: A method for visualizing and interpreting topics; proceedings of the the workshop on interactive language learning, visualization, and interfaces, F, 2014 [C].
20. GENCOGLU B, HELMS-LORENZ M, MAULANA R, et al. Machine and expert judgments of student perceptions of teaching behavior in secondary education: Added value of topic modeling with big data [J]. Computers & Education, 2023, 193: 104682.
21. CHEN J, WEI W, GUO C, et al. Textual analysis and visualization of research trends in data mining for electronic health records [J]. Health Policy and Technology, 2017, 6(4): 389-400.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.