Quantitative Evaluation of Pharmaceutical Industry in Jilin Province Based on Text Mining

Liang Huo and Chengyou Cui(✉)

College of Economic and Management, Yanbian University, Yanji, China
cycui@ybu.edu.cn

Abstract. The pharmaceutical industry has evolved into one of the most strategically focused developing sectors as a result of the high prevalence of several epidemic viruses in recent years. Analysis of the effectiveness of Jilin Province's pertinent pharmaceutical policies is necessary since the province is a significant source of medicinal resources in China with a wealth of herbal medicine reserves as well as clear advantages in the growth of the pharmaceutical industry. The text mining methods employed in this paper include LDA topic modelling, word frequency analysis, and keyword extraction. To visually analyze the policy themes and priorities of China's pharmaceutical industry during the past ten years, 33 important pharmaceutical industry policies that were made public at the national level in that country between 2012 to 2022 are utilized as text mining objects. Along with pertinent references, the text mining results are used as a blueprint to create a more rigorous pharmaceutical industry assessment index system. Finally, we adopt the PMC model to quantitatively evaluate the key pharmaceutical policies issued by the Jilin government in recent years, analyzed the heterogeneity, strengths, and weaknesses of each key pharmaceutical policy in Jilin Province from a variety of dimensions and put forward reasonable suggestions for the improvement of relevant policies.

Keywords: Pharmaceutical Industry Policy · Text Mining · PMC

1 Introduction

The Internet has given the general population access to a large amount of electronic data information against the backdrop of information age. Data mining technology can replace manual work to extract valuable and potential knowledge from a vast amount of data, which has steadily taken over as one of the crucial research methods in both social science research and natural science research now [1]. Text mining is an extension and expansion of data mining, when semi-structured or unstructured textual data is the subject of data mining, the process is known as text mining [2]. Text mining technology is an enlargement and extension of data mining. Early adopters of text mining techniques employed statistical approaches to finish simple word segmentation work in 1959. Yet Feldman didn't first propose the idea until 1995.Currently, there are extensive applications and relatively sound theoretical results on the basis of text mining techniques in the

P. Qi and Z. Chen (Eds.): ICBDIE 2023, AISR 178, pp. 580–598, 2024.
https://doi.org/10.2991/978-94-6463-238-5_80

areas of text topic modelling, text sentiment analysis, text clustering, text classification, and semantic analysis [3]. Text data typically comes from online news, public policy, web articles, social media, etc. Jia and Wu adopted LDA to mine the textual themes of a large number of academic papers on new energy vehicles from the China National Knowledge Infrastructure, analyzing the evolution of high-frequency words and main topics in the papers to explore the development trend of new energy vehicle research in China [4]; Baris Ozyurt and M. Ali Akcayol introduced emotion layer into LDA algorithm and created SS-LDA algorithm to analyze the emotional polarity of Turkish comments in online social media [5]. Rothwell and Zegveld have proposed a traditional division of policy instruments into supply-based, demand-based and environment-based categories in the field of policy research [6]. On the basis of this, subsequent scholars frequently introduced two-dimensional or even multidimensional policy analysis framework. For example, Yue et al. conducted qualitative analysis of China's national comprehensive nursing policy documents from 2009 to 2019 and explored the application of policy tools in the development of Integrated Care in China through a three-dimensional analytical framework from the dimensions of policy tools, stakeholders, and health service supply chains [7]; Zhang et al. made a comprehensive overall analysis of medical service policies in Sichuan Province of China and pointed out the advantages and disadvantages of relevant policies through a three-dimensional medical service policy analysis framework from the dimensions of policy tools, policy intensity, and types of medical service activities [8].

In recent years, the field of policy research has made extensive use of the quantitative policy assessment model known as PMC. It includes a wide range of sectors, including agriculture, industry, resources, public safety and biomedicine. In the field of pharmaceuticals, Xu et al. used the PMC model to quantitatively evaluate ten representative pharmaceutical industry policies in China through text mining and provided pertinent suggestions for policy improvement [9]; Han et al. integrated the PMC model with word frequency analysis and semantic association analysis of China's rare disease policies to conduct a thorough evaluation of national rare disease policies [10].

This paper, which is based on the research of numerous scholars, uses text mining to analyze 33 important pharmaceutical policies that were published at the national level in China from 2012 to 2022. It relies on the high directivity of national-level policies and applies the text mining results to create a more scientific pharmaceutical policy evaluation index system. Finally, this paper statistically assesses the key. Pharmaceutical Industry policies adopted by the Jilin Provincial Government through the PMC model, offers quantitative policy recommendations and puts forward rationalized suggestions for policy optimization.

2 Methodology

2.1 Research Framework

Chinese state-level pharmaceutical industry policies are used as the corpus source in this paper. Due to the relatively small number and dispersed distribution of policies, it is difficult to complete the information acquisition at one time by crawling technology. Hence, this paper adopts a manual approach to screen official websites to collect the

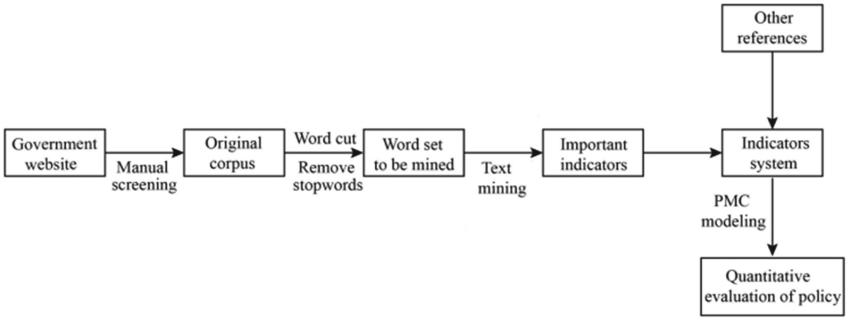

Fig. 1. Research framework. Source: Designed by the authors

important medical policies issued by the State Council of China, ministries, and commissions of the State Council or agencies directly under the State Council in the past ten years to obtain the original corpus of policy texts.

After acquiring the original corpus, it is of great importance to firstly conduct data processing, i.e. text pre-processing, which includes word cut, removing stopwords, information cleaning and merging, etc., before carrying out the mining of the intrinsic information of the text [11]. In this paper, we employ word frequency analysis, keyword extraction, and LDA topic modelling to mine the intrinsic information of the text, extract key indicator variables, and refer to relevant references to construct a rather rigorous review system for pharmaceutical policies. Finally, we adopt the PMC model to quantitatively evaluate key pharmaceutical policies in Jilin Province. The specific research framework is shown in Fig. 1.

2.2 Word Frequency and Keyword

Word frequency statistics, a fundamental and mechanical approach to text mining, appears to be an essential quantitative tool to extract key information from text data. Simple word frequency statistics, however, don't always offer a reliable analysis of the significance of words. For instance, some commonly used words may appear frequently throughout several publications despite having no bearing on the article's subject. Moreover, TF-IDF, as an algorithm for extracting keywords, is relatively more widely used in the field of text mining. In addition, the keyword extraction results may be represented graphically through data visualization technology to create a word cloud map to display the strength information of keywords [12].

Term frequency-inverse document frequency, often known as TF-IDF, is a composite statistic made up of two statistics to evaluate the importance of a word in a collection of documents [13]. A word's frequency in a document is indicated by TF, while a keyword's prevalence in a collection of documents is shown by IDF. The value rises as the term appears more frequently and falls as the number of occurrences in the document collection grows. The formula for the TF-IDF algorithm is shown in Formulas (1), (2), and (3).

$$TF_{\mathrm{Ii,m}} = \frac{n_{I,\mathrm{m}}}{\sum_k n_{k,m}} \tag{1}$$

$$IDF_i = log\frac{|D|}{1 + |m : t_i \in d_m|} \tag{2}$$

$$TF - IDF = TF \cdot IDF \tag{3}$$

In the above formula, $n_{i,m}$ denotes the number of occurrences of the word t_i in the document d_m; $TF_{i,m}$ represents the frequency of occurrences of the word t_i in the document d_m; |D| represents the total number of documents, and $|m{:}t_i \in d_m|$ signifies the total number of documents containing the word t_i. In order to keep the denominator of Formula (2) from 0, Constant 1 must be a part of the denominator.

2.3 LDA Modelling

Word frequency analysis and keyword extraction, which typically reflect only the quantitative status and strength of words, in most cases fail to delve into the semantic linkages between words and texts, nor can they uncover more valuable implicit information. To address this challenge, Blei, et al. proposed the Latent Dirichlet Allocation (LDA) in 2003, which can provide a probabilistic distribution of topics for each document in a document set to accomplish a powerful topic modelling exercise. LDA, one of the most widely used text mining methods today, is essentially an unsupervised learning model for the semantic clustering of text [14]. The LDA algorithm views a document as a collection of unordered words. Each word in the document is generated by one of the document's topics, which is a triple probabilistic structure comprising the document, topic, and word. The probability of each word occurring in the document is shown in Formula (4).

$$P(w|e) = n\sum P(w|t) \cdot P(t|e) \tag{4}$$

In Formula (4), the letters **w**, **e**, **t**, and **n** stand for the word, article, subject, and the number of topics respectively. Figure 2 depicts the specific topology diagram.

Extracting topic feature words is the primary goal in the LDA model refers to completing text semantic clustering. The generation method of characteristic words is as follows:

(1) Sample from the Dirichlet distribution α to produce the topic distribution $\boldsymbol{\theta_k}$ for the article **k**.
(2) Sample from $\boldsymbol{\theta_k}$ to create the topic $\mathbf{w_{k,m}}$ for the **m-th** word in the article **k**.
(3) Sample from the Dirichlet distribution **β** to generate word distribution **Ψ** for topic $\mathbf{w_{k,m.}}$
(4) Sample from the polynomial **Ψ** of words to create the feature word $\mathbf{z_{km}}$.

The joint distribution of all variables during the entire operation is shown in Formula (5).

$$P(w_k, z_k, \theta_k, \psi|\alpha, \beta)\prod_{m=1}^{N} = P(\theta_k|\alpha)P(w_{km}|\theta_k)P(\psi|\beta)P(z_{km}|\varphi_{w_{km}}) \tag{5}$$

The Bayesian network model of the feature word generation process is illustrated in Fig. 3.

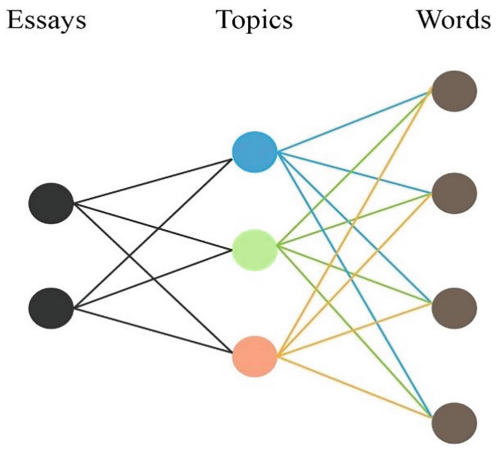

Fig. 2. Topology structure model of LDA. Source: Plotted by the authors according to the reference [14]

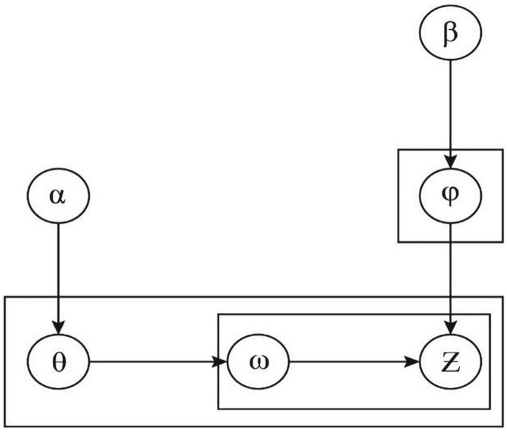

Fig. 3. LDA Bayesian network structure diagram. Source: Plotted by the authors according to the reference [4].

2.4 PMC Modelling

The Policy Consistency Model is the full name of PMC, which was originally proposed by Ruiz Estrada et al. in 2011. In the PMC index model, any factor variable is dynamic and connected, where no feasible variable should be disregarded during conducting an overall policy evaluation, and neither the quantity nor the weighting of variables should be restricted [15]. The steps in constructing a PMC index model involve the following: designing and constructing a multiple input-output table, identifying the parameters of the variables, and performing the calculation of the PMC index. The variables of the

policy index system in this paper are mainly determined collaboratively on the basis of the findings of text mining and other references, including the first-level variable and secondary variables. The value of each first-level variable is summed up to the PMC value of the policy with the average value of each secondary variable being equal to the value of its first-level variable. The policies can be ranked using the PMC value. In addition, Excel can be used to create a PMC visualization curved surface so that the advantages and disadvantages of policies in each dimension can be seen clearly. The PMC modeling flowchart is shown in Fig. 4.

Among the following formulas, Xi and Xij denote the first-level variables, and the secondary variables respectively. The specific calculation formulas of PMC model are shown in Formulas (6), (7), (8), and (9).

$$X \sim N(0, 1) \tag{6}$$

$$X = \{XR : (0, 1)\} \tag{7}$$

$$X_i = \sum_{j=1}^{t} \frac{X_{ij}}{T(X_{ij})} \tag{8}$$

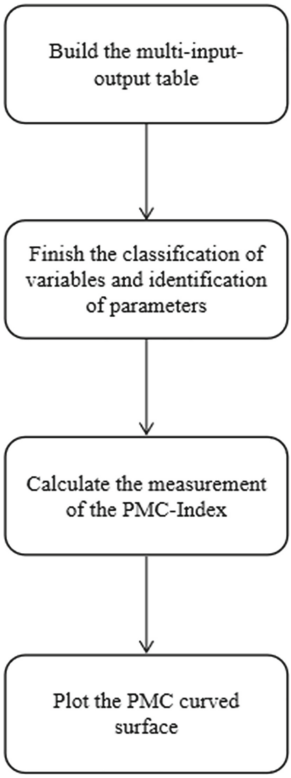

Fig. 4. PMC modeling flowchart Source: Plotted by the authors according to the reference [15]

Table 1. Score evaluation criteria of PMC

9.0–7.5	7.49–6.0	6.0–4.5	4.49–0
Excellent	Well	Acceptable	Unqualified

Source: Plotted by the authors according to references [9] and [10]

$$PMC = \sum_{i=1}^{m} X_i \tag{9}$$

According to Formula (6) and Formula (7), all secondary variables are shown to follow the [0,1] distribution, which means they are given a value of **1** if they appear and a value of **0** if they do not. **R** indicates that the values of secondary variables are all integers, and **T** is the number of secondary variables that are a part of its first-level variable. Assuming that **m** is the number of the first-level variables in the policy evaluation index system, the score of each first-level variable of a policy is the average of the secondary variables under the first-level variable, and the sum of the scores of all first-level variables is the PMC score. In addition, we are able to construct a variable matrix to draw surface curved plots of PMC, intuitively observing the advantages and disadvantages of each policy. The smoothness of the surface curve plots can also reflect the synergy of policies. The higher the smoothness of the plot is, the more comprehensive the policy is. After consulting other sources, the number of first-level variables of the indicator system in this paper is found to total 9, making the full score for each first-level 1 and the overall score of each policy 9. The precise score policy evaluation criteria of PMC are shown in Table 1.

3 Empirical Analysis

3.1 Source of Corpus

Due to the scattered nature of China's national-level pharmaceutical industry policy documents and their low number, this paper adopts a manual screening method to select from the Official Website of the Chinese Government, the Official Website of the National Healthy Commission of the People's Republic of China, the Official Website of the Chinese Medical Product Administration and the Official Website of the National Development and Reform Commission of China. A total of 33 policy documents were screened, but due to space limitations, Table 2 only shows partially acquired policy names because of space limitations.

3.2 Word Frequency Analysis and Keyword Extraction

After acquiring a corpus in text format, the unstructured text data into structured data, including Chinese word cut and removing stopwords. English text can be automatically separated by spaces or punctuation marks. In contrast, Chinese words lack separators, so Chinese word cut should be performed initially [16]. In this paper, the exact pattern of the jieba toolkit in Python is used to complete the task, which significantly lessens the ambiguity in the Chinese language. Additionally, the article, after the word cut, contains

Table 2. Some sources of policy corpus

The Institution issuing policy	The Policy name
The State Council of China	The Outline of "Healthy China 2030" Plan
The State Council of China	China's National Health Plan for the 14th Five-Year Plan
The State Council of China	The Several Policies and Measures on Accelerating the Development of TCM Characteristics
The National Healthy Commission of the China	The Chinese Health Standardization Work Plan for the 14th Five-Year Plan
The State Council of China	The Development of Health Services of China for the 12th Five-Year Plan
The Chinese Medical Product Administration	The Chinese Network security and information construction of drug supervision for the 14th Five-Year Plan
The Chinese Medical Product Administration	The Guiding Opinions on Strengthening and Promoting Scientific and Technological Innovation of Food and Medicine
The National Development and Reform Commission of China	The Bioeconomy Development Plan for the 14th Five-Year Plan

Source: Various official websites in China

a large number of useless words like dummy words, high-frequency common words, etc. Despite being used more frequently in the document, these term have no research significance. Consequently, it is necessary to utilize the process of removing stopwords to delete them. This paper summarizes the stopwords by referring to other references and the stopword list of the Harbin Institute of Technology. Some stopwords are shown in Table 3.

The data must first be pre-processed before starting the text mining process. In order to perform simple mechanical word frequency statistics, the first step is to sort the frequency values of the words by establishing a dictionary. The output of the top 15 words in the word frequency ranking is shown in Table 4.

Mechanical word frequency statistics sometimes fall short of accurately capturing the specificity and importance of words. Yet, it is possible to extract commonly used words but less relevant to the topic. Using this information as a foundation, this paper also presents a TF-IDF model for word vector weighting to enhance the importance of keywords and complete keyword extraction. The top 15 keywords with TF-IDF values are illustrated in Table 5.

From the word frequency list and keyword list, it can be seen that medicine, health, service, and regulation are at the top of both word frequency and keyword importance, whereas word development has a high word frequency but relatively low importance as a keyword. The importance of traditional Chinese medicine as a keyword is strong

Table 3. Stopword list

and
being
but
can
did
does
et
if
once
soon
because
consequently

Source: The stopword list of Harbin Institute of Technology and relevant references

Table 4. Word frequency list

Word	Frequency	Word	Frequency	Word	Frequency
Drug	1928	Traditional Chinese Medicine	861	Ability	608
Healthy	1403	Administration	713	Standard	597
Service	1150	Medical care	681	Innovation	583
Development	1019	Establish	648	Hygiene	574
Supervise	902	System	640	Institution	566

Source: Python

with a low frequency, which also provides a certain basis for the construction of the next evaluation index system of pharmaceutical industry policies.

3.3 LDA Topic Modeling

The importance of words was determined in a considerably comprehensive way of word frequency statistics and TF-IDF keyword extraction, but no semantic connection was established [17]. Due to its clear utility for creating correlations between the first-level and secondary variables in the index system scientifically, this paper also introduces the LDA algorithm to model the unknown corpus and complete topic-word clustering. We utilize the Gensim module as well as the pyLDAvis module within Python for LDA topic modelling and complete modeling visualization. The number of topics in this paper is chosen to be 5, and the topic-word association parameter λ is chosen to be 0.8

Table 5. Keywords list

Word	TF-IDF	Word	TF-IDF	Word	TF-IDF
Drug	0.15271	Medical care	0.04825	Assess	0.03953
Traditional Chinese Medicine	0.09247	Medical equipment	0.04823	Development	0.03925
Healthy	0.09018	Hygiene	0.04564	Medical hygiene	0.03718
Service	0.06618	Medical institution	0.04052	Medical insurance	0.03672
Supervise	0.05831	System	0.04020	Innovation	0.03606

Source: Python

Table 6. Theme-word model

Topic1	Topic2	Topic3	Topic4	Topic5
Traditional Chinese medicine	Healthy	Innovation	Drug	Service
Service	Service	Biology	Supervise	Medical care
Traditional Chinese medical science	Development	Technology	Medical equipment	Medical insurance
Healthy	Standard	Development	Review	Hygiene
Development	Hygiene	Production	Ability	Medical hygiene
Clinical	Medical care	Medicine	Administration	Development
Research	System	Quality	Approval	Healthy
Establish	Mechanism	Drug	Standard	Medical institution
Administration	Establish	R&D	Production	Drug

Source: Python

in combination with other scholars' research experiences. The specific extracted topic-word model is shown in Table 6, and the visual results are shown in Fig. 5. There are no overlapping areas for the five topics, which illustrates that the LDA modelling results are roughly reasonable, as shown by Table 6 and Fig. 5, which also provide the clustering findings and keyword rankings of each topic.

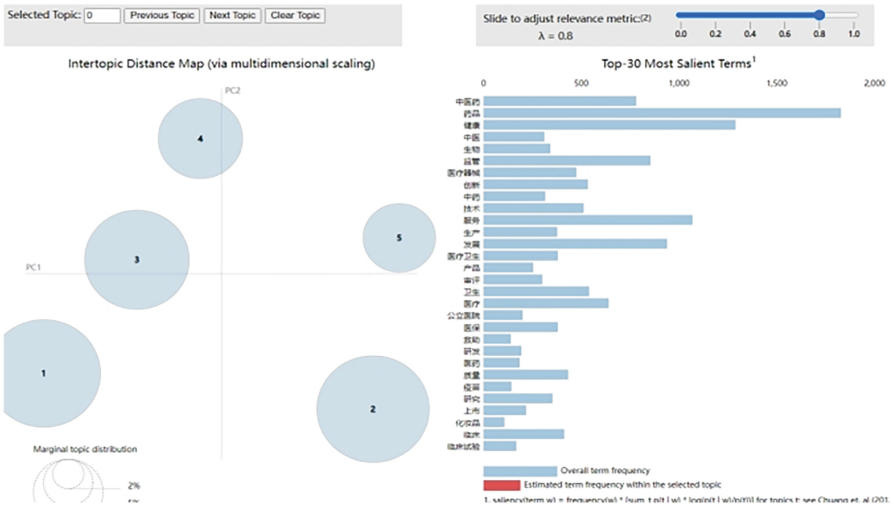

Fig. 5. Visualization of LDA modeling results. Source: Python

3.4 PMC Modelling

This paper builds a quantitative indicator variable system table for pharmaceutical industry policy as shown in Table 7, and adopts the PMC model to complete the quantitative evaluation of important pharmaceutical industry policies in Jilin Province on the basis of the index system constructed by other scholars for pharmaceutical industry policies as well as the information on China's national pharmaceutical industry policies mined by word frequency statistics, keyword extraction, and LDA topic modelling based on Python.

When referring to other references, it is found that most researchers set the issuing agency as a type of first-level variable and the administrative level of the issuing agency as the corresponding secondary variable. Since the research objects of this paper are all provincial-level pharmaceutical industry policies in Jilin Province, the administrative level of the issuing agency appears to have less influence on policy evaluation, so this variable is removed from the research. In addition, most academics would include the variable of government agencies for the secondary variables contained in the policy receptors. According to the analysis of this paper, all the policy documents involve government agencies, which have little impact on how well the policies are valued. As a result, this variable is also deleted in this paper. The specific pharmaceutical industry policy evaluation index system is shown in Table 7.

The aforementioned X1, X3, and X8 were constructed using the text mining results and reference [9] modified and enhanced, X2, X5, and X9 were built based on the text mining results and reference [10], and X4, X6, and X7 were constructed based on reference [18] and text mining results. The key pharmaceutical industries in Jilin Province are screened using this index system in order to calculate PMC values and complete the evaluation of some pharmaceutical policies in Jilin Province. The specific objects of analysis involve the six policy documents shown in Table 8, which are all important

Table 7. The quantitative rating system of medical policy

The first-level variable	Number	The secondary variable
Policy nature (X_1)	$X_{1:1}$	Suggestion
	$X_{1:2}$	Support
	$X_{1:3}$	Guide
	$X_{1:4}$	Supervision
	$X_{1:5}$	Regulation
	$X_{1:6}$	Forecast
Policy measures (X_2)	$X_{2:1}$	Financial allocation
	$X_{2:2}$	Legal protection
	$X_{2:3}$	Talent development
	$X_{2:4}$	Technical support
	$X_{2:5}$	System optimization
	$X_{2:6}$	Convenient approval
Policy focus (X_3)	$X_{3:1}$	Public healthy
	$X_{3:2}$	Drug innovation
	$X_{3:3}$	Chinese medicine industry
	$X_{3:4}$	Drug safety
	$X_{3:5}$	Biotechnology R&D
	$X_{3:6}$	Medical device update
	$X_{3:7}$	Service quality
	$X_{3:8}$	Regulations and standards
	$X_{3:9}$	Medical insurance
	$X_{3:10}$	Review system
	$X_{3:11}$	Health management system
	$X_{3:12}$	Community health care
	$X_{3:13}$	Pharmaceutical resources
	$X_{3:14}$	Clinical application
Policy function (X_4)	$X_{4:1}$	Promoting scientific innovation
	$X_{4:2}$	Drug quality upgrading
	$X_{4:3}$	Improving health industry
	$X_{4:4}$	Optimizing the industrial structure
	$X_{4:5}$	Boosting informatization level

(*continued*)

Table 7. (*continued*)

The first-level variable	Number	The secondary variable
	$X_{4:6}$	Perfecting the medical service system
	$X_{4:7}$	Regional characteristic development
Policy effectiveness (X_5)	$X_{5:1}$	Short-term (0 to 3 years)
	$X_{5:2}$	Medium-term (3 to 5 years)
	$X_{5:3}$	Long-term (more than 5 years)
Policy integrity (X_6)	$X_{6:1}$	Clear goals
	$X_{6:2}$	Reasonable basis
	$X_{6:3}$	Explicit planning
Policy perspective (X_7)	$X_{7:1}$	Macrography
	$X_{7:2}$	Microcosmic
Policy areas (X_8)	$X_{8:1}$	Economy
	$X_{8:2}$	Politics
	$X_{8:3}$	Science and technology
	$X_{8:4}$	Society
	$X_{8:5}$	Hygiene
policy receptors (X_9)	$X_{9:1}$	Healthcare institutions
	$X_{9:2}$	Enterprises
	$X_{9:3}$	Patients
	$X_{9:4}$	Technical talents
	$X_{9:5}$	Medical equipment

Source: Some relevant references and text mining results

pharmaceutical industry policy documents issued by the Jilin Provincial Government in recent year.

After experts review and consultation, this paper assigns a value to the multiple input-output table of the PMC model, giving a value of 1 to the secondary variables involved in the policy and 0 to the secondary variables not involved. Each first-level variable receives a full score of 1, and since there are 9 primary variables, the full score of the PMC value for each policy is 9. The PMC score of the six Jilin pharmaceutical policies is shown in Table 9 with reference to the score evaluation criteria of PMC in Table 1.

Table 9 shows that the six important pharmaceutical industry policies in Jilin Province are all at a level above Acceptable, with the excellent policies scoring very closely together. The policies with lower grades are primarily a result of the gap between other

Table 8. Some important medical policies of Jilin Province

Number	The policy name
1	The Development of Traditional Chinese Medicine in Jilin Province for the 14th five-year plan
2	Several measures to support and promote the construction and development of "Changchun Pharmaceutical Valley"
3	Implementation Opinions on Accelerating the Construction of a Strong Pharmaceutical Province and Promoting the High-quality Development of the Pharmaceutical Health Industry
4	The Development of Pharmaceutical and Health Industry in Jilin Province for the 14th five-year plan
5	The medical device industry development plan for the 14th five-year plan
6	Implementation Opinions on Promoting the Development of Pharmaceutical and Health Industry in Jilin Province

Source: Official websites of Jilin Province

Table 9. PMC score chart of Pharmaceutical Policy in Jilin Province

Number	1	2	3	4	5	6
X_1	0,667	0.500	1.000	0.833	0.667	1.000
X_2	1.000	0.833	1.000	0.667	0.833	0.833
X_3	0.714	0.500	0.786	0.857	0.500	0.929
X_4	0.857	0.714	0.857	1.000	0.714	0.857
X_5	0.667	0.333	0.333	0.333	0.333	0.333
X_6	1.000	0.667	1.000	1.000	1.000	1.000
X_7	1.000	0.500	1.000	1.000	0.500	1.000
X_8	1.000	0.800	0.800	1.000	0.800	1.000
X_9	0.833	0.667	0.833	1.000	0.667	0.667
Score	7.738	5.514	7.609	7.690	6.014	7.619
Grade	Excellent	Acceptable	Excellent	Excellent	Well	Excellent

Source: Calculated by the authors and verified by experts

policies in the scores of policy focus, policy function, policy integrity and policy perspective. A 3×3 PMC surface evaluation matrix is constructed and is shown in Formula (10) to make it easier to observe each policy's advantages and disadvantages.

$$PMC_{SEM} = \begin{pmatrix} X_1 & X_2 & X_3 \\ X_4 & X_5 & X_6 \\ X_7 & X_8 & X_9 \end{pmatrix} \tag{10}$$

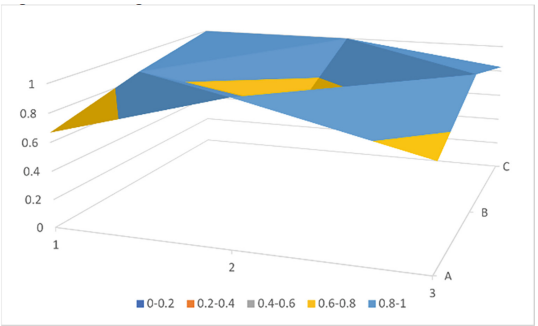

Fig. 6. Policy 1 curved surface

To further examine the synergy and heterogeneity of the individual policies visually, this paper uses Excel to complete visual PMC surface plots of Policy 1 to Policy 6 as illustrated from Fig. 6 to Fig. 11.

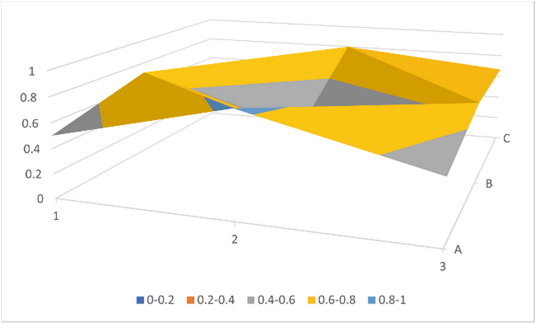

Fig. 7. Policy 2 curved surface

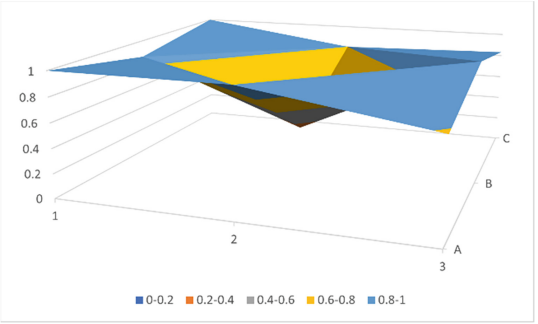

Fig. 8. Policy 3 curved surface

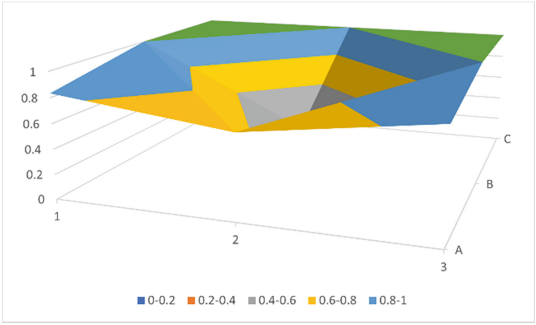

Fig. 9. Policy 4 curved surface

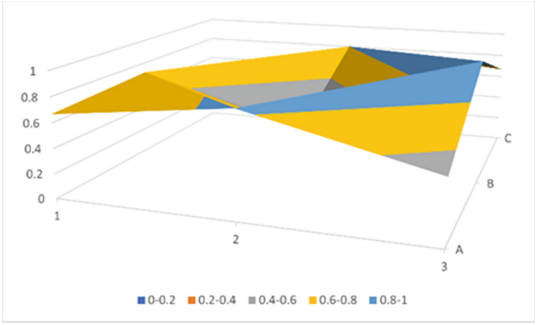

Fig. 10. Policy 5 curved surface

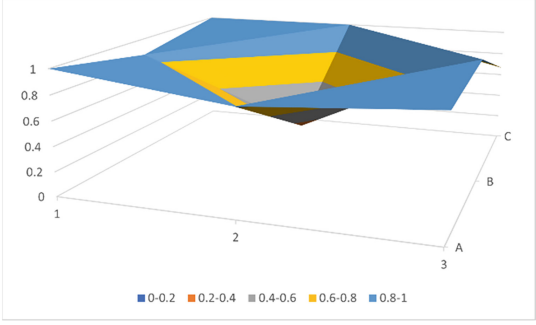

Fig. 11. Policy 6 curved surface. Source: Excel

4 Conclusions and Deficiencies

The following conclusions can be made from the aforementioned PMC surface plots. All six of Jilin Province's key pharmaceutical industry policies are within an acceptable range, although there are still significant gaps in them. Only P1 has a less depressed and smoother surface, indicating that P1 outperforms the other policies in terms of

policy coherence and comprehensiveness, while P1, P3, P4, and P6 have higher PCM scores with a grade of Excellent. On X5 Policy effectiveness, all policies generally have poor scores, revealing that most pharmaceutical industry policies in Jilin Province have a shortage of timeliness with the majority of policies concentrating on medium-term and long-term outcomes and neglecting short-term urgent goal. Due to P2's narrower policy perspective, P2 performs noticeably worse than other policies across all first-level indicators, resulting in low scores for policy focus, policy functions, and policy receptors involved in this policy. P5 has similar issues to P2, but P5 outperforms P2 in terms of policy nature and Policy integrity, raising P5's policy level above P2. Both P1 and P4 score higher and similarly mainly because they are more comprehensive and have a larger variety of policy receptors than other policies. P3 and P6 have similar scores and are also in the Excellent category. Nevertheless, P6 has fewer policy receptors and less policy functionality than P1 and P4, while P3 covers a narrower range of policy areas, ultimately making their PMC scores lower than P1 and P4.

By and large, the important pharmaceutical industry policies issued in Jilin Province are highly systematic as a whole, but there are still issues like insufficient efficacy, constrained policy perspectives, and a lack of policy receptors. In terms of specific implementation measures and policy priorities, Jilin's pharmaceutical industry policies are relatively detailed and comprehensive. In order to establish high-quality pharmaceutical industry policies, Jilin Province can continue to exert this advantage and make further additions in terms of policy effectiveness and policy perspectives as far as possible in the upcoming development of pharmaceutical industry policies.

In this paper, text mining techniques using word frequency statistics, keyword extraction, and LDA topic modelling are used to examine and analyze China's national-level pharmaceutical industry policies. The pharmaceutical industry policy evaluation index system constructed based on other references is further refined using the national-level policies as the guide and utilizing their directionality and scientific nature. Additionally, under the guidance of specialists, the index system is used to assign scores to evaluate key pharmaceutical industry policies in Jilin Province, bringing the empirical analysis process to a conclusion with certain reference significance. However, there are certain shortcomings in this paper. For instance, the premise of PMC modelling is that the number and weight of individual indicators are not constrained, and when a primary indicator contains multiple secondary indicators, each secondary indicator under that primary indicator has little influence on the evaluation of policy value, which may call for additional justification and verification in real-world scenarios. Moreover, there may be more near-sense terms in the LDA topic modelling, which has an impact on the construction of the indicator system to a certain extent. To further examine the topic terms and minimize the interference of synonyms. Subsequent research can consider applying Word2vec and other word vector similarity algorithms. This would help construct a more scientific quantitative evaluation system for pharmaceutical industry policies.

References

1. Yang J, Li Y, Liu Q, et al. Brief introduction of medical database and data mining technology in big data era [J]. J Evid Based Med, 2020, 13 (1): 57-69. DOI: https://doi.org/10.1111/jebm.12373.

2. Bezdan T, Stoean C, Naamany A A, et al. Hybrid Fruit-Fly Optimization Algorithm with K-Means for Text Document Clustering [J]. Mathematics, 2021, 9 (16): DOI: https://doi.org/10.3390/math9161929.
3. Tandel S S, Jamadar A, Dudugu S. A Survey on Text Mining Techniques [J]. Int Conf Advan Compu, 2019: 1022-1026.
4. Jia S, Wu B. Incorporating LDA Based Text Mining Method to Explore New Energy Vehicles in China [J]. IEEE Access, 2018, 6: 64596-64602. DOI: https://doi.org/10.1109/access.2018.2877716.
5. Ozyurt B, Akcayol M A. A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA [J]. Expert Systems with Applications, 2021, 168: DOI: https://doi.org/10.1016/j.eswa.2020.114231.
6. Su Y S, Lin C J, Li C Y. An assessment of innovation policy in Taiwan's electric vehicle industry [J]. International Journal of Technology Management, 2016, 72(1-3): 210-229. DOI: https://doi.org/10.1504/ijtm.2016.10001568.
7. Yue X, Mu K, Liu L. Selection of Policy Instruments on Integrated Care in China: Based on Documents Content Analysis [J]. Int J Environ Res Public Health, 2020, 17 (7): DOI: https://doi.org/10.3390/ijerph17072327.
8. Zhang H, Zhu L, Zeng C, et al. Text Mining and Quantitative Research of Medical Service Policy: Sichuan Province as an Example [J]. Front Public Health, 2020, 8: 509842. DOI: https://doi.org/10.3389/fpubh.2020.509842.
9. Li-ying X, Han Q, Xu L, et al. Policy evaluation of biomedical industry based on PMC index [J]. Chinese Journal of New Drugs, 2020, 29(13): 1501-1507.
10. Meng H, Zi-wen G, Ya-fei L, et al. Text mining quantitative analysis of policies related to rare diseases and their drugs [J]. Chinese Journal of New Drugs, 2022, 31(22): 2193-2201.
11. Wen P, Feng L, Zhang T. A hybrid Chinese word segmentation model for quality management-related texts based on transfer learning [J]. PLoS One, 2022, 17(10): e0270154. DOI: https://doi.org/10.1371/journal.pone.0270154.
12. Heung J D. English Bible Text Visualization Using Word Clouds and Dynamic Graphics Technology [J]. Korean Journal of Applied Statistics, 2014, 27 (3).
13. Havrlant L, Kreinovich V. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation) [J]. International Journal of General Systems, 2017, 46(1).
14. Jelodar H, Wang Y, Yuan C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey [J]. Multimedia Tools and Applications, 2018, 78 (11): 15169-15211. DOI: https://doi.org/10.1007/s11042-018-6894-4.
15. Ruiz Estrada M A. Policy modeling: Definition, classification and evaluation [J]. Journal of Policy Modeling, 2011, 33 (4): 523-536. DOI: https://doi.org/10.1016/j.jpolmod.2011.02.003.
16. Goh C L, Asahara M, Matsumoto Y. Pruning false unknown words to improve Chinese word segmentation[J]. PACLIC 18: Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation, 2004: 139-149.
17. Dai H, Dai X, Yi X, et al. Semantic-aware multi-keyword ranked search scheme over encrypted cloud data [J]. Journal of Network and Computer Applications, 2019, 147: DOI: https://doi.org/10.1016/j.jnca.2019.102442.
18. Yong'an Z, Haituo Q. Quantitative Evaluation of the Impact of Financial Policy Combination to Enterprise Technology Innovation—Based on the PMC-Index Model[J]. Science & Technology Progress and Policy, 2017, 34 (02): 113-121.