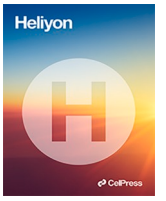
Heliyon 9 (2023) e20768

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Big data sentiment analysis of business environment public perception based on LTP text classification ——Take Heilongjiang province as an example

Kan Liu , Xueying Sun *, Hongrui Zhou

School of Management, Harbin University of Commerce, Heilongjiang, 150028, China

ARTICLE INFO

Keywords:
Public perception
Business environment
Big data text mining
Sentiment analysis

ABSTRACT

At present, the research of business environment is limited to conducting surveys on specific groups or measuring data from official databases. The assessment of the business environment largely depends on public perception. Aiming to explore the public perception of business environment, this paper organically combines the big data text mining and sentiment analysis (SA). The results show that the combination of big data text mining and SA can reflect the theme characteristics, reduce the bias of sentiment and text analysis, and clearly show the public perception of the business environment. The empirical study found that the public perception of business environment depends on not only the four dimensions of business environment, but also the influence of public opinion that cannot be ignored. The public's low recognition of the business environment in Heilongjiang Province mainly includes backward economic development, serious brain drain, low government efficiency, imperfect policies, administrative law enforcement, regional climate and urban construction. In order to solve these problems, it is necessary to improve the high-standard market system to promote economic development, enhance the efficiency of government services, improve government policies, effectively enhance law enforcement, strengthen infrastructure construction and promote cultural innovation.

1. Introduction

The business environment is crucial to the development of a country and a region [1], which is the premise of a country's high-quality development [2] and an important embodiment of economic soft power [3]. In theory, the business environment is a systematic environment, emphasizing the "soft environment" of marketization [4], legalization [5], facilitation and internationalization [6]. In practice, the business environment is the soil for the survival and development of start-ups and enterprises, representing the regional soft power and competitiveness [7]. A good business environment can attract the inflow of investment, technology and talents, and promote the high-quality development of regional economy [8]. For example, the State Council of China launched a supervision platform to collect clues about improving the business environment in 2023, stating that enterprises and the public can enter the column to reflect the clues, opinions and advice, and make suggestions for optimizing the business environment and promoting high-quality development, and the effects remain to be seen. Klapper et al. (2011) believe that positive business environment as a force behind business creation process is conducive to the economic development [9]. Liu et al. (2022) taking small and

* Corresponding author.
E-mail address: xueying_sun0518@163.com (X. Sun).

https://doi.org/10.1016/j.heliyon.2023.e20768
Received 10 June 2023; Received in revised form 5 October 2023; Accepted 5 October 2023
Available online 6 October 2023

medium-sized enterprises in China as examples suggest that funding results in a performance growth by forming an appropriate business environment [10]. Nayal et al. (2022) argue that the business environment contains the digitalization of supply chain [11]. The existing literature mainly focus on the perspective of firms or industries, failing to consider the public perception or public sentiment while the public perception is always thought related to sentiment [12].

According to the sentimental cognition theory in social cognitive theory, the participation of cognition and the evaluation process of cognition on environmental and physiological arousal are the mechanisms of sentiment generation [13]. Therefore, the evaluation of business environment is closely related to public perception [14]. However, in the existing researches, scholars focus on specific groups or enterprises, or study a specific dimension of the business environment, but fail to understand the public perception of the business environment in a broad and objective way. Additionally, various research has focused on the analysis and application of data [15–17]. The public perception of the business environment is not explored on the basis of data mining and SA.

Based on the above analysis, the public perception of business environment still faces great challenges such as vague dimensions and limited subjects. Also, the measurement of public sentiment is still unclear. Thus, this paper proposes the following research questions:

(i) From the public perspective, what are the problems of the business environment in the region?
(ii) What are the public concerned topics and sentiment feedback on business environment?
(iii) How does the government guide the improvement of good business environment and make economic policies in order to boost the economic development?

In order to solve these problems, this study measures public perception of business environment based on big data SA. The innovations of this research are as follows. (1) On the basis of big data mining and SA, this paper directly classifies big data texts into four dimensions, so as to deeply and accurately explore the state of public perception and sentimental value of the business environment in different dimensions. (2) This research expands the depth of text mining, increases the accuracy of perception state and sentimental value, and enriches the research paradigm of big data text mining. (3) This study also provides theoretical and methodological basis for network information mining, making the public perception of business environment more accurate, detailed and persuasive.

Through big data SA, this study aims to discover the public's concerned topics, topic strength and sentimental strength for the four dimensions of the business environment, and obtain the problems existing in the business environment from the public perspective, so as to provide a reference for improving the business environment.

The organization of the sections is arranged as follows: The first part is introduction. The second part is conceptual framework. The third part is the research design and method, determining the research idea and the specific method used. The fourth part is empirical analysis, using the identified big data text mining and SA methods, to explore the public perception topic, perception intensity and sentimental intensity of the business environment in Heilongjiang Province. The fifth part is the conclusion, explaining the research significance, suggestions and future research of the paper.

2. Conceptual framework

A good business environment is characterized by convenience, fairness, transparency, rule of law, and internationalization. At

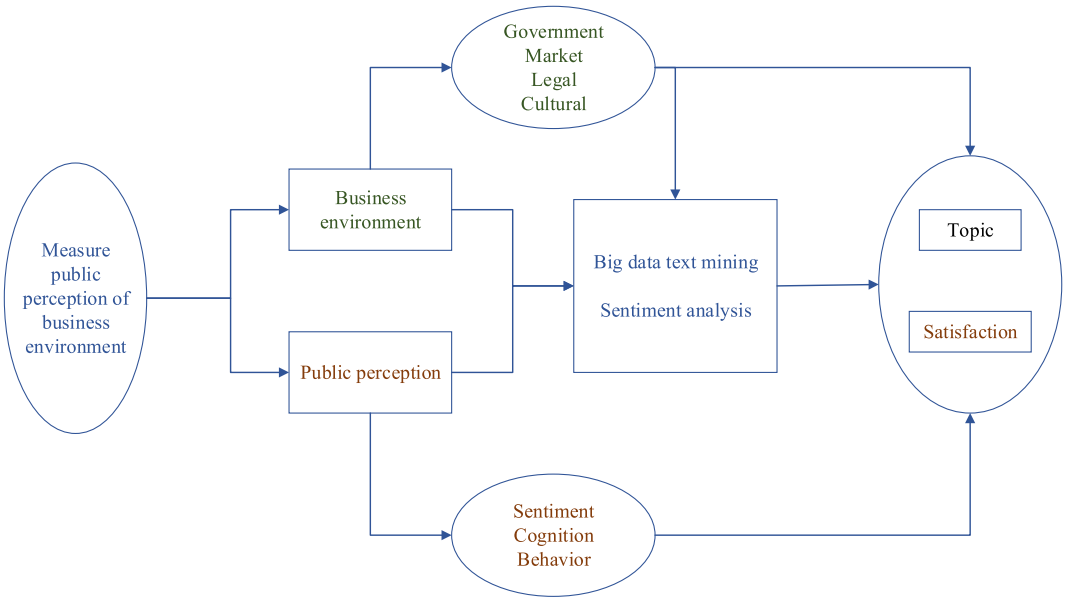


Fig. 1. Conceptual framework diagram.

present, China's business environment has improved significantly, but there are also weak areas such as unbalanced reform and institutional blocking points, which needs to further improve the business environment.

Business environment refers to the sum of external factors and conditions such as the government environment, market environment, legal environment and cultural environment involved in the process of market entities' entry, production and operation, and exit [18].

The economy in Heilongjiang Province continues slumping. There exists an obvious gap of business environment between Heilongjiang Province and coastal areas in southeast China, especially in investment and talent introduction. Understanding the public perception of business environment and optimizing the business environment in Heilongjiang Province are urgent, so taking Heilongjiang Province as an example can better improve the practical value on the basis of big data mining and SA.

The results of the evaluation of various services in the business environment largely depend on public satisfaction [19]. The essence of the source of public perception is the progressive relationship of "sentiment, behavior and cognition". Media such as social public affairs, economy, culture, environment, life and individuals will all affect the public's direct reaction to things, and the public's sentimental attitude and psychological reaction tendency to various influences are the public's subjective perception [19]. Therefore, public perception is the perception after processing, and the information received in the cognitive stage affects the public's emotional judgment, and then affects the public's reaction tendency and behavior [20].

Big data mining and SA methods are used to build a big data analysis system for public perception, comprehensively identify fragmented and high-dimensional public perception data, study the regional business environment from the perspective of public perception, identify opportunities and problems existing in the regional business environment according to public perception, and promote the optimization of the business environment. The conceptual framework of the paper is shown in Fig. 1.

3. Research design and methods

In the relevant research on the business environment, there are few studies using the text mining method of big data. For example, using the data of private enterprises as samples, big data research methods are used to explore how to optimize the business environment and promote development [21]. Based on the big data text mining method, the government public opinion related to the business environment is mined and mapped to understand public opinion, laying a foundation for the government to accurately govern and implement relevant policies. Big data web crawler technology is used to retrieve big data information of power business environment [22]. In the business environment research conducted by using big data, the research object is single, the research content is not comprehensive enough, and the subjective perception of the public cannot be obtained, the public opinion cannot be fully understood, and the public's sentimental perception of the business environment is not measured. This paper uses big data mining and SA to measure the public perception of the business environment, so as to achieve a wide range of research objects and comprehensiveness of research content, fully and accurately excavate the public perception content, perception state and sentimental value, understand public opinion, and provide an effective basis for further improving policies related to the business environment.

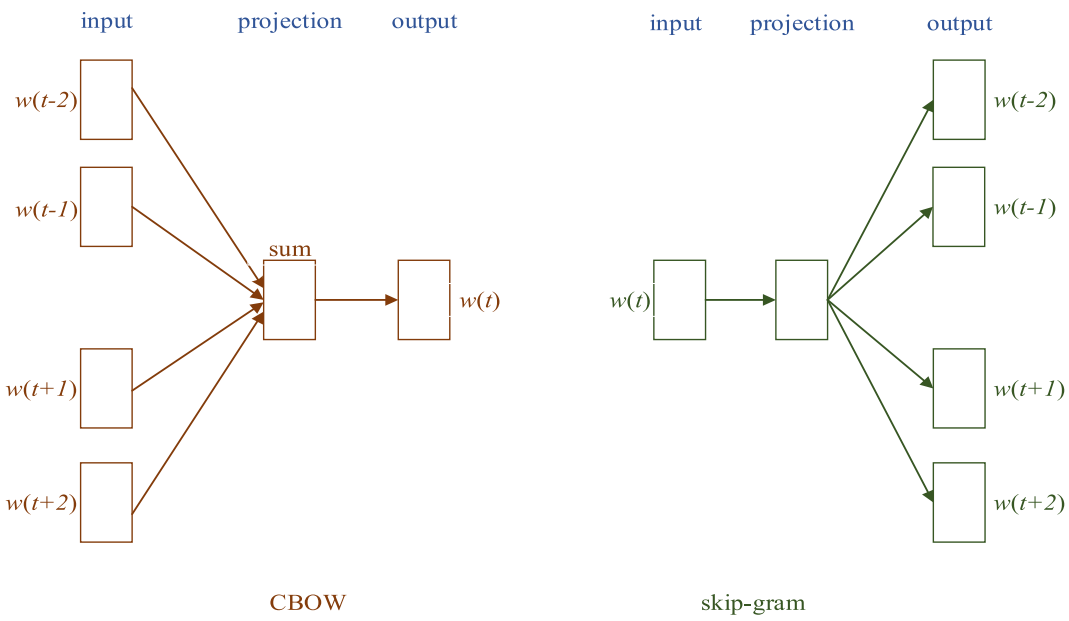


Fig. 2. Comparison of CBOW and Skip-gram.

3.1. Semantic network analysis

Semantic network analysis methods is a formal analysis method based on semantic network knowledge [23], which models elements such as people, things, relationships and attributes through nodes and relational chains. The formal expression of the relationship between conditions, objectives and elements, as well as the expression of logic, is mainly applied to the simulation relationship between concepts and things. Based on the transformation of the model, the relationship between concepts, things and elements is analyzed and reasoned, and problems are found and contradictions are analyzed [24], so as to provide a basis for finding and solving problems. Semantic network analysis methods can intuitively present the problem or the topic, making complex problems simpler and easier to understand.

3.2. Vectorization of comment words based on Word2vec

Word2vec is a model in python's Gensim toolkit for generating word vectors, using this model to calculate the word vectors in the comment text, convert natural language symbols into numerical information in the form of vectors, convert natural language problems into machine learning problems, and represent the relationship between words. Word2vec is implemented by two different ideas, CBOW (Continuous Bag of Words) and Skip-gram. CBOW, also known as word bag model, is a simplified model in information retrieval and natural language processing, similar to the text information such as files or sentences in a bag, and then expressed in the form of words. By default, CBOW does not consider the order of words, and all words in the context have the same influence weight on the current word. This model will predict the probability of the current word according to the context. Both methods use artificial neural network as classification algorithm, and finally get the optimal vector of each word. In order to obtain a more appropriate and accurate word vector, this study adopts the CBOW model framework to calculate word vector. As shown in Fig. 2.

3.3. Word clustering based on K-means clustering algorithm

K-means clustering algorithm is one of many clustering algorithms, the operation is simpler, faster, and easier to understand its principle. The *K*-means clustering algorithm should be manually specified in advance to divide the data into several categories, and the algorithm can only be applied to continuous data. Let the data sample be X,X = {X_1,X_2,X_3, …,X_n}, the initial *k* centers of mass are {C_1, C_2,C_3 …, C_n}, 1<*k* ≤ n. Euclidean distance calculation formula as shown in formula (1).

$$\text{Euclidean distance of cluster center : } \mathrm{dis}(\mathrm{Xi}\ ,\ \mathrm{Cj}) = \sqrt{\sum_{t=1}^{m} (X_{it} - Cit)^2} \tag{1}$$

X_i represents the i-th object 1 ≤ i ≤ n, C_j represents the j-th cluster center 1 ≤ *i* ≤ k, X_{it} represents the t-th attribute of the i-th object,

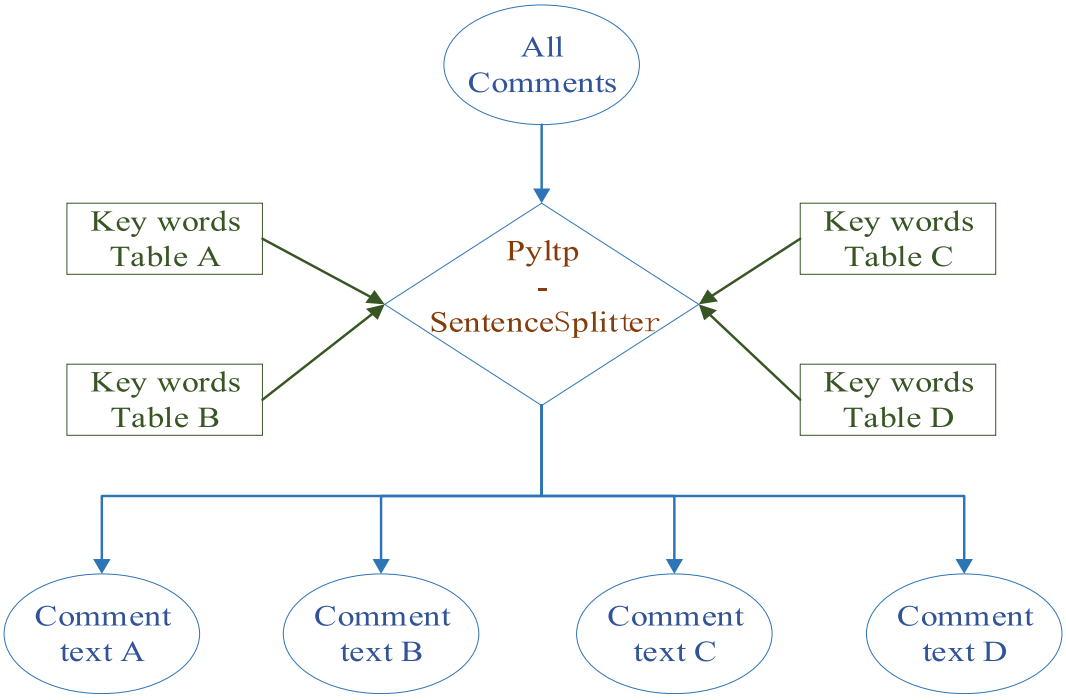


Fig. 3. Text classification diagram based on LTP.

$1 \leq t \leq$ m, C_{jt} represents the t-th attribute of the j-th cluster center.

The specific execution process of K-means is as follows:

(1) First of all, input the value of k, indicating that we need to divide the data into k classes according to the demand through the clustering algorithm;
(2) Randomly select k data from the data set as the initial centroid;
(3) For each word in the set, formula (1) is used to calculate the distance between the word and the center of mass. The nearest center of mass to the word will be classified into which category;
(4) At this time, all the words are clustered into class k, and then a new centroid is selected from each class through the algorithm;
(5) The distance between the recalculated centroid and the original centroid tends to be stable, which is less than a set threshold, indicating that the clustering has reached the expected expectation, and the algorithm terminates;
(6) If the distance between the centroid obtained by recalculation and the original centroid is still large, then the algorithm continues to iterated steps (3) to (5) until the result is stable, and the algorithm terminates.

3.4. Text classification based on LTP

LTP (Language Technology Platform) is the most influential Chinese processing platform at home and abroad, pyltp is the python encapsulation of LTP. Using the SentenceSplitter clause model in the pyltp toolkit, all comments are classified according to the topic keywords. Read the keyword documents of each topic in turn, and extract the comments of the corresponding topic from all the comments according to the keywords of each topic, and then write them into the corresponding documents respectively. The flow diagram of this model is shown in Fig. 3.

Read all comment documents - Read each topic keyword - Extract each topic comment using the LTP split-sentence model - write to the document separately.

3.5. Analysis of sentimental tendency

ROST Content Mining 6 software was used to analyze the sentiment tendency, and the four dimensions of documents after text topic classification were analyzed respectively. After analysis, six analysis result files are obtained from the comment text of each dimension, which are positive sentiment results, neutral sentiment results, negative sentiment results, sentiment distribution view, sentiment distribution statistical results, and SA detailed results. Among them, the sentiment distribution statistical results give the number and

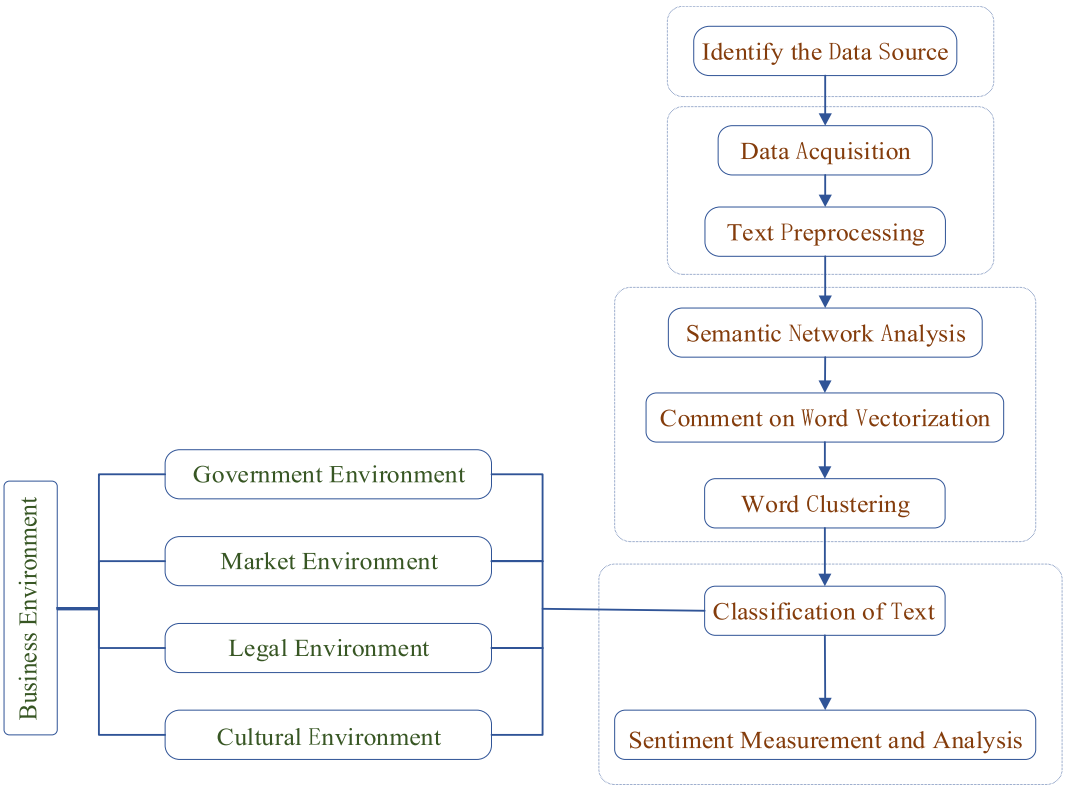


Fig. 4. Measurement model of public perception of business environment.

percentage of comments with positive sentiment, neutral sentiment and negative sentiment, and the number and percentage of comments with segmentation results (general, moderate and high) are given for positive sentiment and negative sentiment, respectively. According to the research demand of this paper, only three sentiment indexes of each dimension comment text need to be obtained, so this study only extracts two parts of the view of sentiment distribution and the statistical results of sentiment distribution for research and analysis, and arranges, summarizes and statistically analyzes these two parts of the files.

Suppose that the number of comment texts in each dimension is Q, and the number of comments whose affective tendency is positive/neutral/negative (affective index greater than/equal to/less than 0) is X, Then the public sentiment perception index Ie in this dimension is calculated by formula (2) :

$$Ie = \frac{X}{Q} \tag{2}$$

4. Empirical analysis

On the basis of clarifying the measurement idea and drawing up the measurement formula, the measurement model based on semantic network analysis, LTP text classification and sentiment analysis is constructed, as shown in Fig. 4.

4.1. Preliminary work preparation

Determine the source of data collection. 'Zhihu' is the most popular knowledge-based question-answering community on the Internet in China, and its questions, answers and comments are of high quality compared with other platforms. Considering the data quantity and value, the network platform 'Zhihu' was selected as the data collection source. Identify research collection topics. When mining the public perception of the business environment in Heilongjiang Province, only a few topics related to "business environment in Heilongjiang Province" are searched and the public participation is very low. Therefore, according to the four dimensions of business environment: government environment, market environment, legal environment and cultural environment, to search the related topics of business environment in Heilongjiang Province respectively. From these four dimensions as the entry point, input "Heilongjiang government", "Heilongjiang market", "Heilongjiang humanities", "Heilongjiang rule of law", "Heilongjiang logistics", "Heilongjiang production and operation", "Heilongjiang government services" and other keywords to search.

4.2. Text acquisition and preprocessing

According to the data collection platform and collection related topics determined by the previous preparation work, the topics that need to be collected are summarized. And then, using python crawler method, a total of 40,781 public comments on 58 valid issues related to the business environment in Heilongjiang Province were crawled, and 5937 invalid comments were removed, and 34,844 effective public comments were finally obtained. In this study, the data preprocessing is carried out, including word segmentation, part-of-speech tagging and stop words removal.

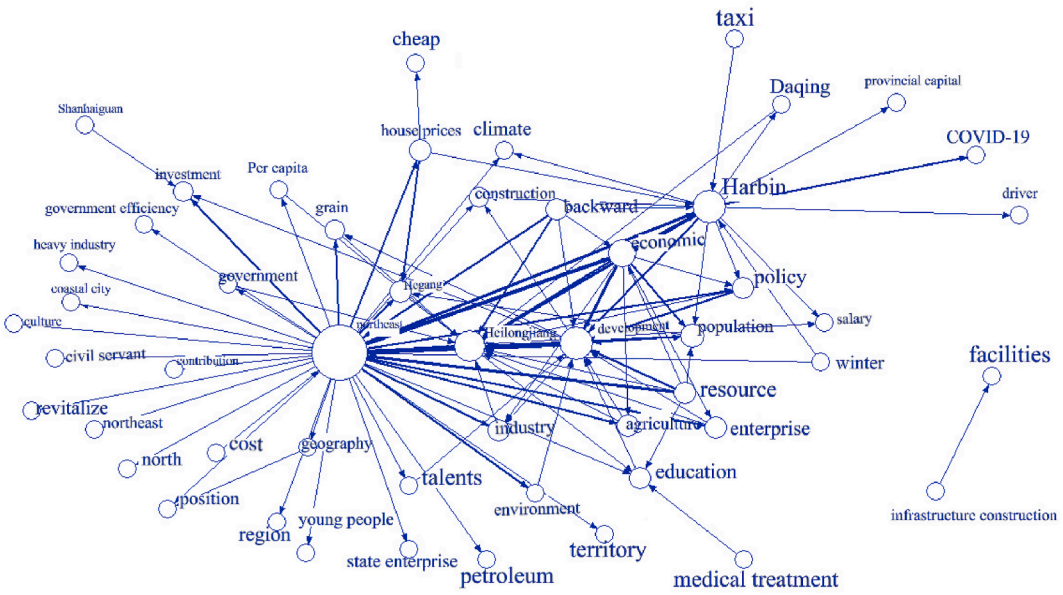


Fig. 5. Semantic network diagram of business environment public perception.

4.3. Semantic network analysis of business environment public perception in Heilongjiang province

On the basis of data preprocessing, semantic network analysis of public perception comments on the business environment can understand the main concerns of the public about the business environment and the main semantic structure of the discussion topics on the whole. The public perception comments of the business environment were imported into the ROST Content Mining 6 software. By extracting high-frequency words, filtering out meaningless words, extracting feature words, constructing the feature word matrix, and then using the semantic network diagram tool NetDraw to generate semantic network diagram for visualization (Fig. 5). It lays the foundation for semantic network analysis of public perception of business environment. In the semantic network graph, the size of the graph node is measured by the centrality of the node degree. The larger the node is, the stronger the centrality of the corresponding content in the whole semantic network, and the more nodes are connected with it. The thickness of the lines between nodes is also different, and the thicker the lines are, the higher the co-occurrence times between two nodes are, and the stronger the relationship strength is.

Through the measure of node degree centrality, it can be obtained that in the whole public perception of the semantic network graph, Northeast China (NrmDegree = 15.130), Development (NrmDegree = 6.610), Heilongjiang (NrmDegree = 5.754), Harbin (NrmDegree = 5.394), Economy (NrmDegree = 4.540), population (NrmDegree = 2.308), resources (NrmDegree = 2.207), policy (NrmDegree = 1.489), backward (NrmDegree = 1.483) and other words have relatively high degree centrality, and closely related words are investment, grain, talent, agriculture, industry, enterprise, environment, education, government, COVID-19, etc. Through the semantic network diagram of public comments, it can be seen that the public pays high attention to topics such as economic development, government policies, population changes, resources and agriculture, investment status, COVID-19, backward development, per capita income and brain drain, government efficiency and house price.

4.4. Keyword cluster analysis based on Word2vec and K-means

Based on the semantic network analysis of public perception, Word2vec, a model for generating new word vectors in the Gensim toolkit, is used to calculate the word vectors in the comment text, and the natural language symbols are transformed into digital information represented by vectors. Then the *K*-means algorithm is used to cluster the topic words of the comment text according to the obtained word vector of the comment text. By constantly debugging the value of *K*, it is found that the word clustering effect is relatively good when the value of *K* is set to 16. Further remove words that are not relevant to the topic. The 16 clusters of words are divided into four dimensions of business environment, namely market environment, government environment, legal environment and cultural environment. Thus, the comment text keywords are divided into four dimensions, as shown in Table 1, which only shows 15 keywords in each dimension and their frequencies.

In order to better show the key words perceived by the public of the business environment, a key word cloud map is generated for visualization for each topic after the key word clustering (Fig. 6), and the key words with high frequency in the public comment text data are displayed in the form of visualization, so that people can more easily appreciate the problem expressed by the public comment data. In the word cloud map, the size of the word represents the word frequency. The larger the keyword is, the more times the public mentions the keyword in the comment text, that is, the issue is valued by the public. The word cloud map of each dimension of business environment is shown in Fig. 6.

4.5. Measurement of business environment public perception based on LTP text classification

Based on the keyword clustering of the comment text, the keyword table of four dimensions of business environment (government environment/market environment/legal environment/cultural environment) is obtained respectively. The split-sentence model in Python-LTP toolkit is used to read all comments and the keyword table of four dimensions, extract the comment text of corresponding

Table 1
Key keywords and their frequency.

Number	Government	Frequency	Market	Frequency	Legal	Frequency	Cultural	Frequency
1	Policy	567	Development	1845	Taxi	263	Education	495
2	Government	471	Economy	1333	Society	251	Environment	402
3	COVID-19	448	House price	1313	Ability	210	Area	356
4	Civil servant	396	Work	893	Police station	168	Climate	344
5	Efficiency of service	339	Resources	789	Objective	149	Quality	228
6	Leadership	268	Backward	717	Take a taxi	96	Culture	221
7	System	227	Salary	434	Phenomenon	93	Tourism	215
8	Nucleic acid	156	Investment	398	Public security	91	College students	203
9	Infrastructure	145	Grain	393	Department	88	Comfortable	191
10	Officials	124	Agriculture	361	News	87	Pension	185
11	Corruption	116	Enterprise	322	Publicity	85	Attitude	179
12	Finance	116	GDP	312	Fairness	70	Effort	178
13	Bureaucracy	106	Revitalization	260	Process	61	Thoughts	151
14	Grassroots level	106	Shanhaiguan	157	Law	57	Attract	107
15	Prevention and control	99	Geographical location	150	Slander	42	Building	59

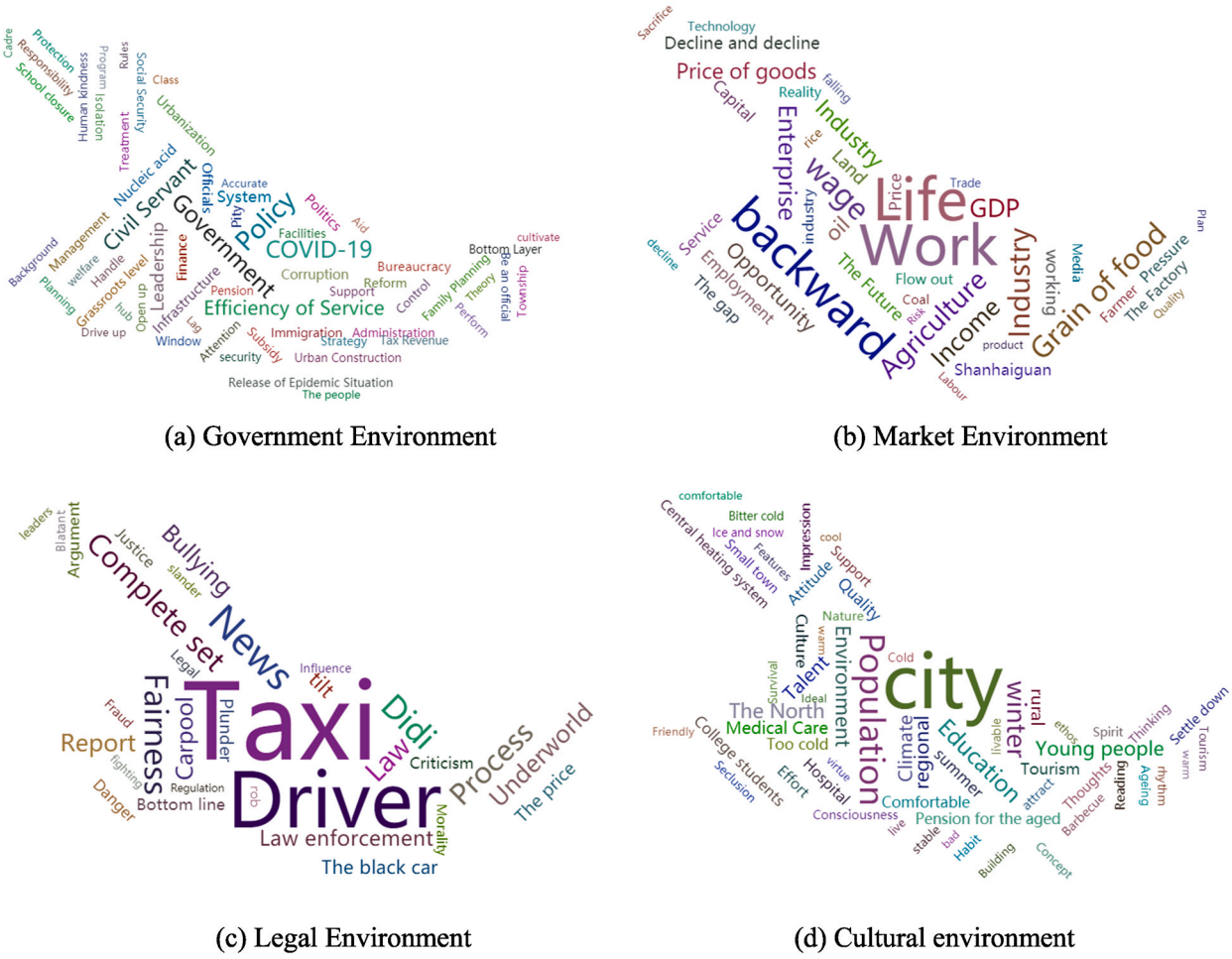


Fig. 6. Cloud map of business environment public perception.

dimensions from all comments according to the keyword table, and then write it into four corresponding folders respectively, thus achieving the text classification of comments.

The results of text classification are shown in Table 2 below.

Based on the comment text classification, the ROST Content Mining 6 software was used for sentiment tendency analysis. This study focuses on measuring the four dimensions of public comment texts separately. Considering that sentimental inclination (positive and negative) is mainly used to measure public sentiment perception, this study takes positive and negative sentiments as main factors and neutral sentiment as secondary factors in the sentiment calculation results of ROST Content Mining 6. Through sentiment analysis of the four dimensions of public comments on the business environment, the public's satisfaction with the government environment, market environment, legal environment and cultural environment of Heilongjiang Province was respectively excavated, so as to put forward countermeasures and suggestions for optimizing and improving the business environment.

The ROST Content Mining 6 software was used to conduct sentiment analysis and assign values to each public comment text. A value of positive indicates positive sentiment, a value of 0 indicates neutral sentiment, and a negative value indicates negative sentiment. At the same time, the advantage of this software is that in order to more accurately express the sentimental tendency, positive sentiments and negative sentiments are divided into three levels: general, moderate and high. The positive mood segments are general (0–10), moderate (10–20), high (above 20), and the negative mood segments are general (−10-0), moderate (−20–10), high (below −20).

Without classification, the general sentiment analysis of business environment public perception is shown in Table 3.

Table 2
Text classification quantity table.

Business environment Dimension	Government Environment	Market Environment	Legal Environment	Cultural Environment	Total
Number of comments (per piece)	5453	16750	4124	8517	34844

Table 3
Overall sentiment analysis table.

	Segmentation	Quantity (pieces)	Proportion (%)	Proportion (%)
Positive Sentiment	High (above 20)	3154	9.05	52.35
	Moderate (10–20)	4568	13.11	
	General (0–10)	10521	30.19	
Neutral Sentiment	(0)	7027	20.17	20.17
Negative Sentiment	High (below −20)	905	2.60	27.48
	Moderate (-20—-10)	1938	5.56	
	General (-10—0)	6731	19.32	
–	–	34844	100	100

Through text classification, the detailed results of public perception sentiment analysis under the four dimensions of business environment are shown in Table 4.

Without text classification, sentiment analysis was conducted on comments related to public perception of business environment (Table 3), and it was found that positive sentiments accounted for 52.35 % and negative sentiments accounted for 27.48 %, which could reflect the sentimental state of the public for the business environment of Heilongjiang Province to a certain extent, but it was not possible to understand and identify the differences in sentimental perception of specific dimensions, nor is it possible to obtain the public's perception of specific concerns about the business environment and their sentimental value.

On the basis of text classification, sentiment analysis is made for public perception comments on four dimensions of business environment (Table 4). Obviously, after classifying the text data through the LTP split-sentence model, it is easier and more intuitive to see the public's perceived state and sentimental value of each dimension of the business environment, which not only expands the depth of text mining but also increases the accuracy of the public's perceived state and sentimental value of the business environment in Heilongjiang Province.

4.6. Results and discussion

The research shows that the public's overall satisfaction with the business environment in Heilongjiang Province is not high. First of all, the public's satisfaction with the market environment of the business environment in Heilongjiang Province is the lowest, with positive sentiments accounting for only 47.30 %. Based on semantic network analysis and clustering analysis shows that in terms of market environment, the public think the economic development in Heilongjiang province is backward, excessive resource consumption, the geographical position compared to other provinces also have no advantage, serious brain drain, low pay, prices are higher. Secondly, the public's satisfaction with the government environment of Heilongjiang Province's business environment is only

Table 4
Sentiment analysis table based on text classification.

Government Environment Number of comments (/piece) Sentiment Category&segmentation Ratio (/%) Type (/%)		Market Environment Number of comments (/piece) Sentiment Category&segmentation Ratio (/%) Type (/%)		Legal Environment Number of comments (/piece) Sentiment Category&segmentation Ratio (/%) Type (/%)		Cultural Environment Number of comments (/piece) Sentiment Category&segmentation Ratio (/%) Type (/%)	
(2747) Positive Sentiment (50.38)	(372) High (6.83) (756) Moderate (13.86) (1619) General (29.69)	(7923) Positive Sentiment (47.30)	(1028) High (6.14) (1831) Moderate (10.93) (5064) General (30.23)	(2411) Positive Sentiment (58.46)	(644) High (15.62) (612) Moderate (14.84) (1155) General (28.01)	(5152) Positive Sentiment (60.49)	(1106) High (12.99) (1359) Moderate (15.96) (2687) General (31.55)
(852) Neutral Sentiment (15.62)		(4737) Neutral Sentiment (28.28)		(432) Neutral Sentiment (10.47)		(1017) Neutral Sentiment (11.93)	
(1854) Negative Sentiment (34.00)	(124) High (2.28) (335) Moderate (6.14) (1395) General (25.58)	(4090) Negative Sentiment (24.42)	(375) High (2.24) (822) Moderate (4.91) (2893) General (17.27)	(1281) Negative Sentiment (31.06)	(174) High (4.22) (287) Moderate (6.96) (820) General (19.88)	(2348) Negative Sentiment (27.57)	(230) High (2.70) (495) Moderate (5.81) (1623) General (19.06)
5453 Total (100)	5453 Total (100)	(16750) Total (100)	(16750) Total (100)	(4124) Total (100)	(4124) Total (100)	(8517) Total (100)	(8517) Total (100)

3% points higher than that of the market environment, with a satisfaction rate of 50.38 %. The public believed that the government's working efficiency was low, the government's policies were not perfect, and the COVID-19 prevention and control had loopholes [25, 26]. As a result, Heilongjiang Province had been affected by the COVID-19, which was not conducive to the urban development. The bureaucracy and corruption also had a negative impact on the government environment [27]. Then, compared with the market environment and the government environment, the public's satisfaction with the legal environment and cultural environment is relatively high, but the satisfaction is only about 60 %. In terms of the legal environment, the public generally has a strong reaction to "Take a taxi", "Didi Taxi" and "Unlicensed Taxi". The phenomenon of drivers taking detour, carpooling without passengers' permission, social governance ability, social security, legal strength and formalism of administrative law enforcement personnel are particularly concerned. Finally, the public's satisfaction with the cultural environment of the business environment in Heilongjiang Province is relatively high, accounting for 60.49 %, but there are also many problems. The public believes that Heilongjiang Province is affected by the geographical location, the winter climate is cold, the population is aging seriously, the environment needs to be reformed, and the urban construction needs to be improved [28].

It can be seen that, compared with previous surveys using questionnaires, specific groups and a single dimension of business environment [29,30], the combination of big data mining and SA can fully and comprehensively understand the subjective perception and real sentimental response of the public, and the results show clear comparability and obvious importance. This approach which does not make investigation and research is limited to a specific problem, respondents can give full play to the individual aspiration, the research results are abundant, comprehensive, real and concrete [31]. In the existing researches related to big data SA, scholars either directly collect different topics and then carry out sentiment measurement, or collect a large amount of data to directly identify and analyze the theme, and finally carry out overall SA. However, using the above methods to analyze sentiment and perceptions cannot fully understand the public's concerns, nor accurately grasp the public's sentimental feedback on various topics. The combination of big data mining and SA can supplement the research loopholes and make the research results more objective, accurate and comprehensive [32].

In addition to the four dimensions of the business environment, the perceived state and sentimental value of the public on the business environment are also affected by public opinion and have a negative value. According to the results of thematic word clustering, in terms of legal environment, in addition to keywords such as public security, law enforcement, bottom line, agency and red line, there are also keywords such as rumor, doubt, smear phenomenon, propaganda, helpless and society. These additional findings fully validate the effectiveness and advantages of big data mining and SA methods. The effective and fast extraction of new high-quality information from a large number of unordered and messy information shows the importance of the topic. These new findings are not limited to the questionnaire, and the respondents will not be subjectively guided [33]. They are the most real and effective public focus topics and sentimental feedback.

It can be seen that the business environment of Heilongjiang Province is facing the doubts of the public, promoting the negative information of the business environment of Heilongjiang Province and the frustration that the business environment of Heilongjiang province is deliberately smeared. These keywords reflect the public opinion on the business environment in Heilongjiang Province. It is closely related to the saying "investment except Shanhaiguan (meaning investment outside northeast China)" and so on. It alludes to the basic reality of low social recognition of Heilongjiang's business environment. A negative perception of the business environment will not only affect the inflow of capital and technology, but also lead to the loss of talent and serious brain drain [34]. Therefore, this study concludes that the public perception state and sentimental value of the business environment in Heilongjiang Province are not only based on the four dimensions of government affairs, market, legal and cultural in Heilongjiang Province, but also influenced by public opinion. It can be seen that the problems existing in the business environment are not only improving the business environment, but also the negative perception of the public on the business environment. These negative public opinions will reduce the enthusiasm of capital and labor inflow, and then continue to affect economic development and investment inclination, forming a vicious circle.

5. Conclusions

5.1. Significance

Although the existing studies have paid attention to the evaluation of business environment and its optimization path as well as the public's perception of policies, the current research still has the following two shortcomings: First, the research object is single, and the research only focuses on a specific group or a specific dimension of an event (such as students, specific occupational groups, enterprises, official databases, specific dimensions of the business environment, etc.) Second, the survey methods and contents are not comprehensive. For example, surveys take the form of questionnaires only. Fixed contents and topics in questionnaire cannot accurately reflect the subjective perceived content, focus and attention of the public, as well as the level of public sentimentality. In the research related to public perception, some scholars directly collect different topics and then measure sentiment separately. Some scholars directly identify and analyze the theme of the collected data, and finally conduct the overall SA. However, directly collecting topics for sentimental and perceptual analysis does not fully understand the public concerns. Big data's topic identification and direct SA cannot accurately grasp the public sentimental feedback on each topic.

This paper uses the research method of data science to explore the public perception of web information mining. Based on topic recognition, we propose to use LTP split-sentence model for text classification, and group a large number of reviews according to different topics, and then finally perform SA separately. It is possible to fully understand the public concerns and also accurately measure the sentiment value consistent with the topic, making big data SA more accurate, detailed and convincing. Utilizing big data mining methods to measure public perceptions after collecting the data from Zhihu (a famous public platform of social network in

China), this empirical study found that public perception of the business environment is not only a cognitive state resulting from the general awakening of simple emotions and environmental suggestions, but is also influenced by third-party social opinion factors that have great appeal and reflect the public's position and sentiment. The public will not only be infected by it, but also unconsciously accept the suggestions of social opinion. It expands the depth of text mining, and improves the accuracy of perceptual state and sentiment value mining, which is conducive to spread out the research paradigm of big data text mining, and to provide the theoretical and methodological basis for web information mining.

This research also provides an effective basis for further improvement of relevant policies and compensates for the shortcomings of existing studies with a single research object, incomplete research content, and inability to fully understand public concerns and sentiment feedback.

5.2. Suggestion

Through the analysis of the public perception of the business environment in Heilongjiang Province and the sentimental measurement of the dimensions, the existing problems are found, which provides a reference for improving the business environment. According to the problems existing in the business environment of Heilongjiang Province, this paper puts forward the following targeted suggestions. First, we will develop and improve a high-standard market system. Relax market access requirements, promote investment, and create a fair market competition order. Second, create an efficient and transparent government environment. Improve the efficiency of government service departments and improve policies related to the business environment. Third, focus on the business environment, rectify the chaos, and provide a fair and just legal environment. Fourth, foster a healthy and vibrant cultural environment. Continue to develop the characteristic tourism industry of Heilongjiang Province, develop the characteristic culture of Heilongjiang Province and constantly promote cultural innovation, strengthen the urban construction of infrastructure, and build an open and inclusive charming province. Fifth, in response to the public's doubts about the business environment, the external media and individuals have not demonstrated, no public reference value of sentimental deliberately smears, the relevant departments should give warnings, make reasonable corrections and explanations. Separately, relevant media and individuals should rectify the atmosphere, and do not continue to undermine regional unity, incite and stir up regional black topics, and intensify contradictions.

5.3. Future research

This study uses the proposed method to measure the public perception and sentiment analysis of the business environment in Heilongjiang Province. However, the evolutionary trend of public perceptions of the business environment in Heilongjiang Province has not been explored in depth; the social opinion factors affecting public perceptions of the business environment show negative effects on public perceptions of the business environment in Heilongjiang Province, which has yet to be verified for other provinces and regions, and the methods used for these issues can be further explored in future studies as well as multi-dimensional empirical studies.

Data availability statement

Data included in article/supplementary material/referenced in article.

CRediT authorship contribution statement

Kan Liu: Conceptualization, Data curation, Supervision, Writing – review & editing, Formal analysis, Validation, Visualization, Writing – original draft. **Xueying Sun:** Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Hongrui Zhou:** Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N.O. Ndubisi, X.A. Zhai, K.H. Lai, Small and medium manufacturing enterprises and Asia's sustainable economic development, Int. J. Prod. Econ. 233 (2021), 107971.
[2] W. Gu, J. Wang, X. Hua, Z. Liu, Entrepreneurship and high-quality economic development: based on the triple bottom line of sustainable development, Int. Enterpren. Manag. J. 17 (2021) 1–27.
[3] N. Wang, D. Cui, C. Geng, Z. Xia, The role of business environment optimization on entrepreneurship enhancement, J Glob. Econ. Anal. 1 (2022) 66–81.
[4] J. Zhang, X. Chen, X. Zhao, A perspective of government investment and enterprise innovation: marketization of business environment, J. Bus. Res. 164 (2023), 113925.
[5] S. Li, An analysis on Regulation on optimizing the business environment: from the perspective of law, Business Administration and Management 2 (2020) 8–13.
[6] C.L. Mann, Supply chain logistics, trade facilitation and international trade: a macroeconomic policy view, J. Supply Chain Manag. 48 (2012) 7–14.
[7] M. Brychko, Y. Bilan, S. Lyeonov, D. Streimikiene, Do changes in the business environment and sustainable development really matter for enhancing enterprise development? Sustain. Dev. 31 (2023) 587–599.

[8] D.F.L. Santos, L.F.C. Basso, H. Kimura, The trajectory of the ability to innovate and the financial performance of the Brazilian industry, Technol. Forecast. Soc. 127 (2018) 258–270.
[9] L. Klapper, A. Lewin, J.M.Q. Delgado, The impact of the business environment on the business creation process, in: Entrepreneurship and Economic Development, Palgrave Macmillan UK, London, 2011, pp. 108–123.
[10] Y. Liu, A. Dilanchiev, K. Xu, A.M. Hajiyeva, Financing SMEs and business development as new post Covid-19 economic recovery determinants, Econ, Anal. Policy. 76 (2022) 554–567.
[11] K. Nayal, S. Kumar, M.M. Raut, Queiroz, P. Priyadarshinee, B.E. Narkhede, Supply chain firm performance in circular economy and digital era to achieve sustainable development goals, Bus. Strateg. Environ. 31 (2022) 1058–1073.
[12] A. Vohra, R. Garg, Deep learning based sentiment analysis of public perception of working from home through tweets, J. Intell. Inf. Syst. 60 (2023) 255–274.
[13] N.K. Singh, D.S. Tomar, A.K. Sangaiah, Sentiment analysis: a review and comparative analysis over social media, J. Amb. Intel. Hum. Comp. 11 (2020) 97–117.
[14] E. Martin, S. Shaheen, T. Lipman, M. Camel, Evaluating the public perception of a feebate policy in California through the estimation and cross-validation of an ordinal regression model, Transport Pol. 33 (2014) 144–153.
[15] G. Azzone, Big data and public policies: opportunities and challenges, Stat. Probab. Lett. 136 (2018) 116–120.
[16] E. Paffumi, M. De Gennaro, G. Martini, European-wide study on big data for supporting road transport policy, Case. Stud. Transp. Pol. 6 (4) (2018) 785–802.
[17] O. Enghoff, J. Aldridge, The value of unsolicited online data in drug policy research, Int. J. Drug Policy 73 (2019) 210–218.
[18] O. Taouab, Z. Issor, Firm performance: definition and measurement models, Eur. Sci. J. 15 (2019) 93–106.
[19] M. Fernandez, B. Allen, T. Wandhoefer, E. Cano, H. Alani, Using social media to inform policy making: to whom are we listening, in: Proceedings of the 1st European Conference on Social Media, 2014, pp. 174–182.
[20] H. Huang, W. Chen, T. Xie, Y. Wei, Z. Feng, W. Wu, The impact of individual behaviors and governmental guidance measures on pandemic-triggered public sentiment based on system dynamics and cross-validation, Int. J. Env. Re.s Pub. He. 18 (2021) 4245.
[21] R. Iqbal, F. Doctor, B. More, S. Mahmud, U. Yousuf, Big data analytics: computational intelligence techniques and application areas, Technol. Forecast. Soc. 153 (2020), 119253.
[22] R. Lin, Z. Xie, Y. Hao, J. Wang, Improving high-tech enterprise innovation in big data environment: a combinative view of internal and external governance, Int. J. Inf. Manag. 50 (2020) 575–585.
[23] M.L. Doerfel, What constitutes semantic network analysis? A comparison of research and methodologies, Connections 21 (1998) 16–26.
[24] J. Dong, G. Zou, L. Qi, Urban design logical scheming based on semantic network information, Inf. Technol. J. 12 (2013) 5725.
[25] K.I.M. Yuliya, S. Daribekov, L. Kundakova, D. Sikhimbayeva, G. Srailova, Role of state institutions in protecting the environment. Improving management system of the public services, Journal of Environmental Management and Tourism 14 (2023) 2379–2389.
[26] L. Paremoer, S. Nandi, H. Serag, F. Baum, Covid-19 pandemic and the social determinants of health, Bmj (2021) 372.
[27] J. Shen, H. Duan, B. Zhang, J. Wang, J.S. Ji, J. Wang, X. Shi, Prevention and control of COVID-19 in public transportation: experience from China, Environ. Pollut. 266 (2020), 115291.
[28] H.R. Güner, İ. Hasanoğlu, F. Aktaş, COVID-19: prevention and control measures in community, Turk. J. Med. Sci. 50 (2020) 571–577.
[29] C.R.G. Popescu, G.N. Popescu, An exploratory study based on a questionnaire concerning green and sustainable finance, corporate social responsibility, and performance: evidence from the Romanian business environment, J. Risk Financ. Manag. 12 (2019) 162.
[30] X. Zhao, C. Yi, Y. Zhan, M. Guo, Business environment distance and innovation performance of EMNEs: the mediating effect of R&D internationalization, J. Innov. Knowl. 7 (2022), 100241.
[31] W.T. Wu, Y.J. Li, A.Z. Feng, L. Li, T. Huang, A.D. Xu, J. Lyu, Data mining in clinical big data: the frequently used databases, steps, and methodological models, Military Med. Res. 8 (2021) 1–12.
[32] R.K. Behera, M. Jena, S.K. Rath, S. Misra, Co-LSTM: convolutional LSTM model for sentiment analysis in social big data, Inform. Process. Manag. 58 (2021), 102435.
[33] P.J. Castagna, D.E. Babinski, A.M. Pearl, J.G. Waxmonsky, D.A. Waschbusch, Initial investigation of the psychometric properties of the limited prosocial emotions questionnaire (LPEQ), Assessment 28 (2021) 1882–1896.
[34] M. Latukha, M. Shagalkina, E. Mitskevich, E. Strogetskaya, From brain drain to brain gain: the agenda for talent management in overcoming talent migration from emerging markets, Int. J. Hum. Resour. Manag. 33 (2022) 2226–2255.