Open Access Library Journal
2019, Volume 6, e5742
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Related Research on "The Belt and Road" Initiative Based on Big Data Text Mining: Taking the Domestic Area and the Korean Peninsula as an Example

Hongyi Li[1], Zhezhi Jin[2*]

[1]Department of Mathematics, Yanbian University, Yanji, China
[2]Department of Economics and Management, Yanbian University, Yanji, China
Email: *780831601@qq.com

How to cite this paper: Li, H.Y. and Jin, Z.Z. (2019) Related Research on "The Belt and Road" Initiative Based on Big Data Text Mining: Taking the Domestic Area and the Korean Peninsula as an Example. *Open Access Library Journal*, **6**: e5742.
https://doi.org/10.4236/oalib.1105742

Received: August 29, 2019
Accepted: September 13, 2019
Published: September 16, 2019

Copyright © 2019 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

As one of the core contents of China, "The Belt and Road" is very necessary to analyze the impact and significance of "The Belt and Road" on the domestic areas and Korean Peninsula areas in the context of big data. The text mining method was used to extract the core content and hot topics from the news reports of "The Belt and Road" in China and the Korean Peninsula in recent domestic and international news data. Using the unique attributes of public opinion media and news reports. Focusing on the key theme of "The Belt and Road", using clustering algorithm and TF-IDF method, combined with LDA topic model, the structure and content of the theme are analyzed in detail on the pre-built big data analysis platform. The establishment of a big data platform and the use of TF-IDF and LDA topic models have improved the efficiency of data analysis. On the domestic front, the eastward extension of "The Belt and Road" is imperative and the prospects are clearer. On the Korean Peninsula, its own security situation is the core issue, and the United States is the biggest drag factor. In the future, it is necessary to focus on the common focus and solve the problem.

Subject Areas

Statistics

Keywords

The Belt and Road, Text Clustering, TF-IDF, LDA Topic Model

1. Introduction

Since "The Belt and Road" was introduced in 2013, it has been regarded as the focus of attention and research in the whole society and even in the world. The analysis of the influencing factors is also one of the research hotspots. "The Belt and Road" is not just an initiative and policy, it is a trend. Since September and October 2013, Chinese President Xi Jinping officially proposed the cooperation initiative for the construction of "The New Silk Road Economic Belt" and "The 21st Century Maritime Silk Road", and after the official release of "The Vision and proposed actions outlined on jointly building Silk Road Economic Belt and 21st-Century Maritime Silk Road" on March 28, 2015. Leading export index of foreign trade, maritime silk road trade index, and "The Belt and Road" aviation trade index all showed steady growth. At present, a large number of scholars have studied the related content of "The Belt and Road". For example, Changxiu Hu, Jinglei Lin and others [1] have studied the situation of "The Belt and Road" in regional economic development. Li Wang [2] has studied the path of strengthening the "The Belt and Road" foreign cultural communication, while Yumiao Lv and Hongchun Wang and others [3]. The application of big data in logistics management under the background of "The Belt and Road" has been studied. However, there are relatively few research literatures on foreign countries, such as the Korean Peninsula, which is closer to China and more closely related to it. Because of the difficulty in obtaining data, language inconsistency, and inability to accurately analyze foreign language data, the relevant research content is very less, but it is essential. Accurately analyzing the development and changes of the influence factors and influencing factors of "The Belt and Road" on the Korean Peninsula and other regions not only provides a scientific basis for the targeted strengthening of the relationship and development between the two parties, but also gives the implementation of the "The Belt and Road" policy. This paper makes use of the text mining method [4] [5], and based on a large amount of news data suitable for time series analysis, clusters [6] [7] [8] [9] [10] the data objects on the big data analysis platform. By finding the TF-IDF value of each word, the correlation and importance of each topic in each text are analyzed and compared. Using the topic model approach, combined with the LDA topic model [11], to more accurately screen the specified topics in each text data, and to derive the changes in the factors affecting countries since "The Belt and Road" initiative, and provide the valuable reference for the subsequent development.

2. Research Steps and Related Theories

Text mining is the process of taking a large amount of unstructured text data through computer natural language processing technology, mastering the pattern of its text data and extracting useful information for analysis. The purpose of this work is to find a periodic, associative pattern between words and sentences after processing the data in the proposed file. Using computer algorithms to extract the meaning hidden in the text, as one of the research methods specif-

ically applied to text analysis, recently also used in the study of Korean language.

The first step is to collect data. So far, there is no automated method for data collection. A large part of the data needs to be searched and collected by human means. This analysis uses the data reported by People's Daily in China, and the data on the Korean Peninsula is the use of RodongSinmun and the news of ChosunIlbo, totaling nearly 100,000. News reports with "The Belt and Road" from 2015 to 2019 were adopted, and news reports containing "North Korea", "South Korea", "Korean peninsula" and other words were selected as specific analysis objects.

The second step is the stage of data cleansing. The process of transforming human natural language into a computer-understandable language is called natural language processing, referred to as NLP. Data cleansing is the process of converting collected text data into natural language processing (NLP) and transforming it into a form that is easy to analyze, which needs to be done through data preprocessing and morphological analysis. In the process of data preprocessing, it is necessary to repeatedly remove unnecessary words and sentences from the collected text data, and unify the words that are similar or identical but have different expression meanings. Then you need to remove stopwords that are not meaningful for analysis, such as special symbols, punctuation, numbers, onomatopoeia. The process of data pre-processing may be repeated, and the scope of its work will produce some differences depending on the analyst. After the pre-processing work, the text of the essay text needs to use the KoNLP analysis package and classify according to the shape. By extracting the stem of the word, removing the affix and other stem extraction work, the unit form with the smallest meaning is separated, and the word class paste work is performed.

The third step is to analyze and visualize the data. First of all, in order to grasp the attention of countries on "The Belt and Road", analyze the number of news reports on "The Belt and Road" first report, and then understand the similarities, core words and topics between the documents. In order to grasp the change of news data and whether there is similarity in each level, the analysis of text clustering by using the similarity between texts is used, which is a method of clustering and distinguishing similar words in text by text clustering algorithm. Multivariate data for cluster analysis does not require additional reaction variables, but only needs to be clustered based on similarity between individuals.

Secondly, we need to find the TF-IDF value to analyze the core words in the text data. Through Term Frequency (TF), according to the frequency of occurrence of specific words in the text, you can grasp the importance of the word. However, the TF value alone cannot grasp the keywords in the text, so it is necessary to use TF-Inverse Document Frequency (TF-IDF) value to calculate the frequency based on the probability of occurrence.

As shown in the Table 1, the TF-IDF value is calculated by multiplying the TF value of specific words in the text by the IDF value, while the TF-IDF analysis,

Table 1. TF·IDF·TF-IDF output formula.

TF	$f_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$ n_{ij} : The number of times a word appears in the text. $\sum_k n_{kj}$: The number of times all words appear in the text.
IDF	$idf_i = \log \frac{\lvert D \rvert}{\lvert \{ d_j \mid t_j \in d_j \} \rvert}$ $\lvert D \rvert$: The amount of text contained in the text collection. $\lvert \{ d_j \mid t_j \in d_j \} \rvert$: The number of texts with the word t_j.
TF-IDF	$\mathrm{TFIDF}_{ij} = tf_{ij} \times idf_i$

which calculates the occurrence probability of words instead of occurrence frequency, is used to evaluate the importance of designated words. And the frequency analysis of the topic words is effective for confirming the appearance of the topic. But it is difficult to grasp the unexpressed subject content only by the presence or absence of words. In order to accurately grasp the theme of each cluster about "The Belt and Road" related content, you need to use the Topic Modeling method. The topic model approach refers to the method of automatically extracting specific topics and combinations based on the frequency and usage of the topic words used in the collection of text. This time, the LDA topic model approach is used. The LDA method [12] is to calculate the appearance probability of each word w constituting the text M, and to judge the composition of the theme, and to connect the individual words and the theme, and then recombine the words together.

The LDA method is composed of various unknown themes in individual texts, and is based on the premise that different words are represented and processed for most topics, and the potential causes of the constituent texts are analogized, and the complex data is reduced and simplified. LDA methods have some similarities with exploratory factor analysis in terms of effective understanding. Compared to other topic model methods, the results of the LDA topic model approach are easier to interpret and can solve overfitting problems. Therefore, it is beneficial to derive topics from a large amount of unstructured data [13] [14]. A large number of recently published extension models for topic models are also based on the LDA method, so there is considerable practicality in terms of continuation of future research.

Finally, it is necessary to visualize the results. Visualization by means of charts and images is a method to display a large amount of data and intelligence in an efficient way. The results of text cluster analysis that analyzes the similarity between texts will be visualized in the form of a dendrogram. In order to achieve the purpose of visualizing the theme model, it is necessary to use LDAvis to visualize and accurately derive the theme of each text and the core words that appear in each theme.

3. Construction of Big Data Platform

3.1. Component-Based System Architecture

The big data platform deployment design is as follows (Figure 1):

Transwarp Data Hub is a high-level platform that combines technical performance and is the most widely use version. It is a platform that supports Spark's Hadoop distribution, which is faster than the open source Hadoop 2 version. This platform incorporates the memory computing techniques, it can handle massive amounts of data, and contains efficient indexing technology, the degree of contain is any size of enterprises, the coverage of data volume is very high. At the same time, the platform can continuously expand its capacity. Without the downtime, the data can be grown without fear. The more significant advantage is that the performance is the highest so far.

Discover data mining machine learning component is an important component of Transwarp Big Data integration platform. It needs to be emphasized here. It can become a representative technology in the field of Big Data Mining because of its high coverage. Especially in the field of visualization, it has shown excellent ease of use, and built-in distributed implementation of many classic machine learning algorithms. For example, clustering, classification, regression, neural network, association, etc. in statistical algorithms. It can also perform text analysis, risk analysis, transaction anti-fraud, etc., inheriting more than 6000 machine learning and data mining algorithms in R language, such as including R language MapReduce distributed computing framework and so on. The distributed machine learning engine provided by the massive data platform is highly easy to use. The R software can be used to execute R language programs to access the data in HDFS, Hyperbase and Inceptor, and achieve high-speed analysis to the massive data in the platform. Discover can support the graphical development interface Midas, which is used in combination with algorithms in the R language to provide the necessary technical support for advanced data mining services in many areas and industries such as user behavior analysis, financial fraud, and marketing strategies.

The big data platform also includes the Hadoop basic platform and the Inceptor distributed memory analysis engine. The independent distributed memory column is a hybrid storage architecture that stores Holodesk and supports

Rack		Rack1			
Node		Node1	Node2	Node3	Node4
TranswarpZookeeper	Zookeeper	√	√	√	
Transwarp HDFS	Namenode	√	√		
	Journal Node	√	√	√	
	Date Node	√	√	√	√
Transwarp Yarn	Resource Manager	√			
	Node Manager	√	√	√	√
Transwarp Inceptor	Metastore		√		
	Inceptor Server		√		
Transwarp Manager	Manager	√			

Figure 1. KPID big data platform deployment design.

memory/SSD cache. It provides higher data access function and it supports the processing of GB to hundreds of terabytes of data by combined linear expansion processing capability with the cluster size expands. In addition, Inceptor supports high-performance add-delete operations on ORC tables and Hyperbase tables through SQL. Compared with open source solutions, Inceptor is far ahead in support. Secondly, the platform includes Hyperbase Distributed Trial Online Data Processing Engine and Stream Processing Engine.

The functions provided by the integrated components can meet the requirements of the enterprise for all scenarios of the data platform. The description of the important components involved is as follows (Table 2).

3.2. Big Data Platform Capabilities

The metadata of the file system is stored on a cluster of JournalNodes. To make HDFS always reliable, maintain a Namenode with hot standby, and avoid a single point of failure, better choose the Namenode HA solution. NameService is used to process files on different HDFS. If you encounter a large amount of data, you could get into troubles during processing. In this case, you must start the HDFS Federation function and use the copy mechanism to deal with data storage security issues.

As shown in the Figure 2, it is configured as three replicas, with different data blocks distributed on a set of different rack servers, and servers closest to the network are provided for access.

When faced with a huge amount of data, the file will be split and distributed among the servers to increase the access width of the huge file. This is because the system can read in parallel, and read in parallel from multiple servers at the

Table 2. Name and description of the component.

Number	Component Name	Description
1	Inceptor (Spark)	Distributed memory computing engine
2	Hyperbase (HBase)	Distributed real-time online NoSQL data service engine
3	Stream (Streaming)	Real-time data processing engine
4	Discover (R)	Encapsulates the R language
5	Manager	Independently developed graphical cluster management tool
6	HDFS	Hadoop distributed file system
7	MapReduce	Distributed data computing model and execution environment
8	Yarn	Unified resource management system
9	Zookeeper	Distributed, highly available distributed coordination services
10	Sqoop	Hadoop relational database synchronization tool
11	Flume	Distributed massive log collection system
12	Oozie	Workflow engine
13	Elastic Search	Full staff search service

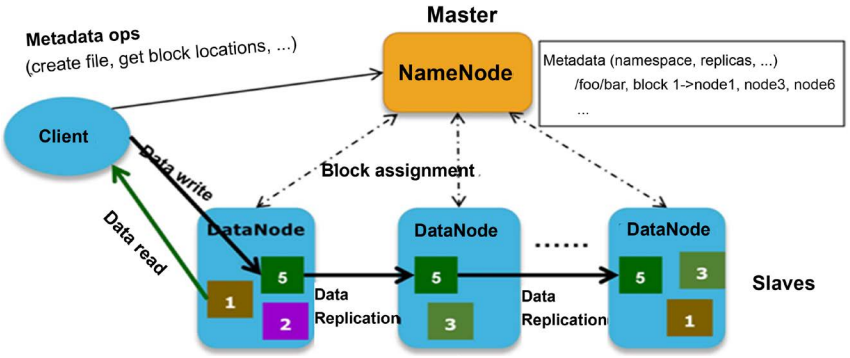

Figure 2. HDFS replication mechanism configuration.

same time. The DataNode of Name Node server is a data block used to store HDFS files. The capacity of the system can be expanded without manual maintenance and intervention. Then the system can match the newly added DataNode in the overall array in real time, and the data block will be moved to the new DataNode. Based on the above, it can be seen that not only can increase the throughput of data, but also make a new breakthrough. When the data is evenly spread across different servers, the limitations of a single server can reach hundreds of times to complete distributed computing.

4. Results of the Analysis

Using the web crawler and manual retrieval methods, the data of the analysis object was collected, and the news data with "The Belt and Road" keywords from 2015 to 2019 was extracted. And the news data including "North Korea", "Korea" and "Korean Peninsula" are selected as specific analysis objects, and the results are shown in the following Table 3.

The greater the concern about "The Belt and Road", there will be more relevant reports in the news of various countries. Based on this premise, the number of publications was sorted out. The number of publications on "The Belt and Road" related content in the news reports of China and the Korean Peninsula is shown in the following Figure 3.

As can be seen from the left picture, in the 2015 Korean News, the number of "The Belt and Road" related reports showed the highest value in the past five years, accounting for 30% of the entire report. After that, there was a slight decline, but the trend tends to be stable. In the picture on the right, before 2017, the number of publications of relevant Chinese reports showed a trend of increasing year by year, reaching a peak in 2017, accounting for 40% of the entire report. And then reduce year by year. Although the number of publications through news reports can measure the public's interest in "The Belt and Road", it cannot be based solely on the frequency of "The Belt and Road" related reports, and it is concluded that the scale of concern has increased. Because there are certain limitations on how to report related content in the "The Belt and Road" related report. Therefore, it is necessary to use text mining method to analyze and process news data in more detail.

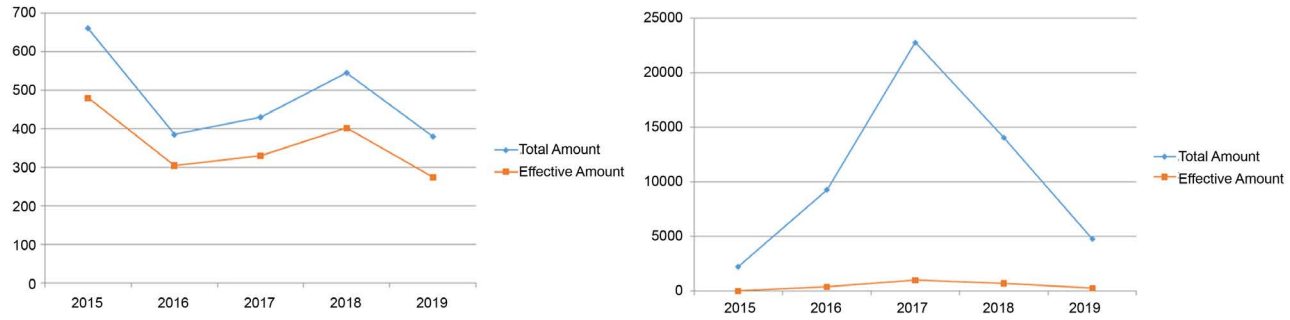

Figure 3. Number of related reports on "The Belt and Road".

Table 3. Data collection methods and results.

Data Sources	Total Amount	Effective Amount
People's Daily	53,911	2454
RodongSinmun & ChosunIlbo	1715	619

4.1. Cluster Analysis Results

In order to carry out a more in-depth analysis of the news data, and to obtain the similarity of the topic words in the text data, and to analyze its relevance, it is necessary to use the cluster analysis methods. Among them, the hierarchical clustering method is used. When performing simple text clustering, the effect of hierarchical clustering is very good, and because clustering uses the principle of distance metric, it can visually see the correlation between words in text data. The maximum number of words in each hierarchical cluster image can be set by itself. Because the clustering results of data of all years cannot be displayed, the use of the korean news data in 2015 and 2018, the relative number of two related reports is relatively high. High-year data is clustered. Set the maximum number of words that can be displayed to 25, and analyze the relationship changes between the core words appearing in each year.

As shown in the **Figure 4**, the news data of two representative years is selected to cluster the objects. From the news of 2015, the main core words include "North Korea", "Russia", "United States", "Seoul", "Transportation", "Development", "Economy", "Logistics", etc. In the clustering diagram, the closer the words are between each branch, the shorter the distance between them, which means the stronger the correlation between them. In the picture on the left, "Culture" and "Future", "Economy" and "World", "United States" and "Russia", "Mainland" and "Transportation" are all strongly related words. In the 2018 cluster map, the words with higher frequency are "United States", "Strengthen", "Construction", "Development", "City", "Demand" and other words. From the branch of clustering, the correlation between "Strengthen" and "Construction", "Standard" and "Effort", "Supply" and "Demand", and "City" and "Development" are relatively strong. It is also possible to visually see the correlation between each core words of the news reports.

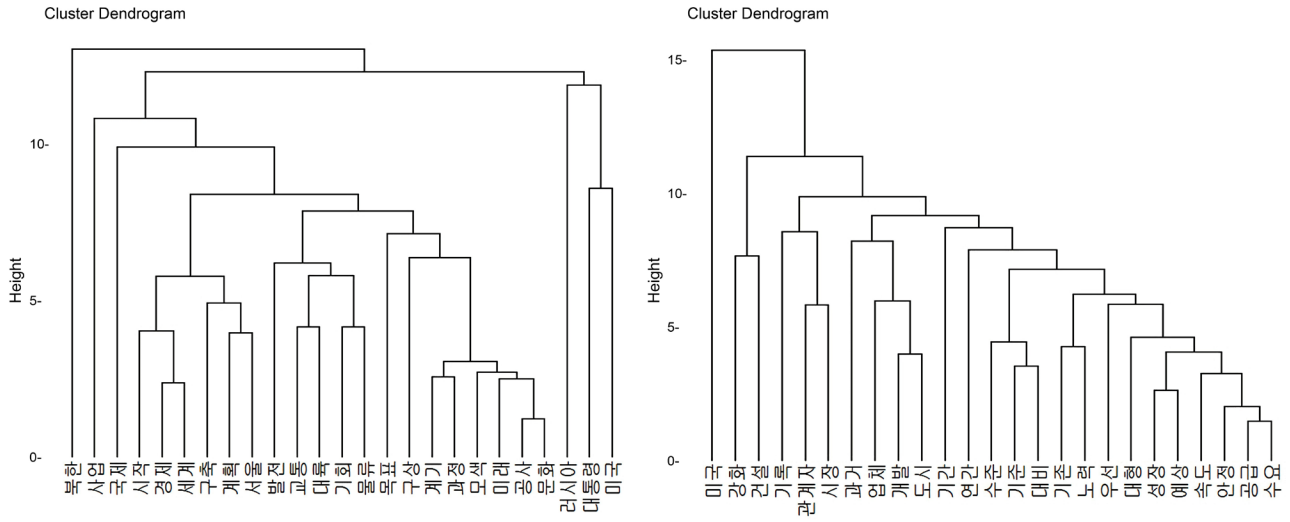

Figure 4. Hierarchical cluster map of "The Belt and Road" data.

4.2. TF-IDF Analysis Results

Cluster analysis can be used to see the correlation between core words and core words in text data, but it is still impossible to more accurately screen the core words in the text through cluster analysis. And it is impossible to know what is the most important core word among the text data by text clustering. The following **Table 4** shows the results of TF-IDF analysis of Chinese news and Korean news data for the past five years based on TF values and TF-IDF values.

In the process of analysis, words such as "China" and "The Belt and Road" exist in the data of each year, so the analysis of the results has no practical significance, so it is eliminated in the text data.

Firstly, among the data of the Korean News, the words with the highest TF and TF-IDF values are "the United States" and "North Korea", followed by "Economy", "South Korea", "Government", "Enterprises", and "Russia", "South Korea"," the United States", "Trump". But it can also be understood that words such as "the United States" and "South Korea" are the most important words because they not only appear the most frequently, but their values are also quite high in terms of TF-IDF values. Among the Chinese news data, the words with the highest TF and TF-IDF values are "Development" and "Tourism", followed by "Cooperation", "Nation", "International", "Economy" and "Culture", "Enterprise", "Trade", "Investment", "Sports".

Secondly, in the Chinese news, the TF-IDF value of the top ranked words such as "Tourism", "Culture", and "Enterprise" is much higher than other words. It can be intuitively seen that the importance of these key words in each text is quite high, and also in the Korean news data. Compared with the data of Korean news, it can be seen that the TF-IDF value of the core word in the Chinese data is generally higher than that of the core word in the Korean data. The difference in the data amount between the Korean news and the Chinese news may also be one of the reasons for this situation. Then it can be seen that the content of core

Table 4. TF·TF-IDF analysis results.

Korean news data				Chinese news data			
Words	TF	Words	TF-IDF	Words	TF	Words	TF-IDF
미국 (US)	2199	조선 (North Korea)	6.09	发展 (Development)	22,327	旅游 (Tourism)	11.94
경제 (Economy)	1545	러시아 (Russia)	4.53	合作 (Cooperation)	19,552	文化 (Culture)	10.81
한국 (South Korea)	1217	한국 (South Korea)	4.43	国家 (Nation)	15,332	企业 (Enterprise)	10.25
정부 (Government)	1104	미국 (US)	4.27	国际 (International)	14,677	习近平 (Xi Jinping)	9.90
대통령 (President)	1022	트럼프 (Trump)	4.16	经济 (Economy)	14,087	贸易 (Trade)	9.22
기업 (Enterprise)	994	유라시아 (Eurasia)	4.12	世界 (World)	11,475	投资 (Investment)	9.01
시장 (Market)	865	협력 (Unite)	3.75	建设 (Construction)	7583	体育 (Sports)	8.87
세계 (World)	862	시장 (Market)	3.56	全球 (Global)	7241	中日 (Sino-Japanese)	8.77
투자 (Investment)	801	달러 (Dollar)	3.48	企业 (Enterprise)	7035	韩国 (South Korea)	8.72
협력 (Unite)	790	인도 (India)	3.45	文化 (Culture)	6699	亚洲 (Asia)	8.46
조선 (North Korea)	771	투자 (Investment)	3.44	贸易 (Trade)	6051	东盟 (ASEAN)	8.33
달러 (Dollar)	680	포럼 (Forum)	3.35	地区 (Region)	6037	外资 (Foreign Capital)	8.22
트럼프 (Trump)	654	사업 (Career)	3.30	推动 (Promote)	6012	北极 (Arctic)	8.05
시진핑 (Xi Jinping)	609	방문 (Visit)	3.14	美国 (US)	5848	中美 (Sino-US)	7.88
유럽 (Europe)	571	푸틴 (Putin)	3.13	问题 (Problem)	5803	中俄 (Sino-Russia)	7.78

words extracted based on TF value and TF-IDF value is almost different. Some core words may appear frequently in the text, but their TF-IDF value may not be very high. Therefore, it is not appropriate to judge the importance of words in the text based on the frequency of occurrence of words in the text.

Finally, according to the results of the Korean news data, it can be seen that the relationship between the Korean Peninsula and various countries, especially the relationship with the United States, is its core content. From the Chinese data, it can be seen that the core of "The Belt and Road" is not on the Korean peninsula. Strengthening relations with countries around the Korean Peninsula and countries such as the United States and Russia is also one of its core contents.

4.3. LDA Topic Analysis Results

By analyzing the TF-IDF values in each text, we can accurately grasp the core words appearing in "The Belt and Road" related reports. However, there are some limitations in understanding the vocabulary and ways in which core themes are used in various texts. Therefore, in order to understand in detail how the theme of the "The Belt and Road" related reports of each cluster is handled, it is necessary to use the topic model analysis. Use the LDA algorithm to extract the topics of "The Belt and Road" related reports and visualize the results by using LDAvis. The following Figure 5 is the result of visualizing the LDAvis data of the Korean news for nearly five years.

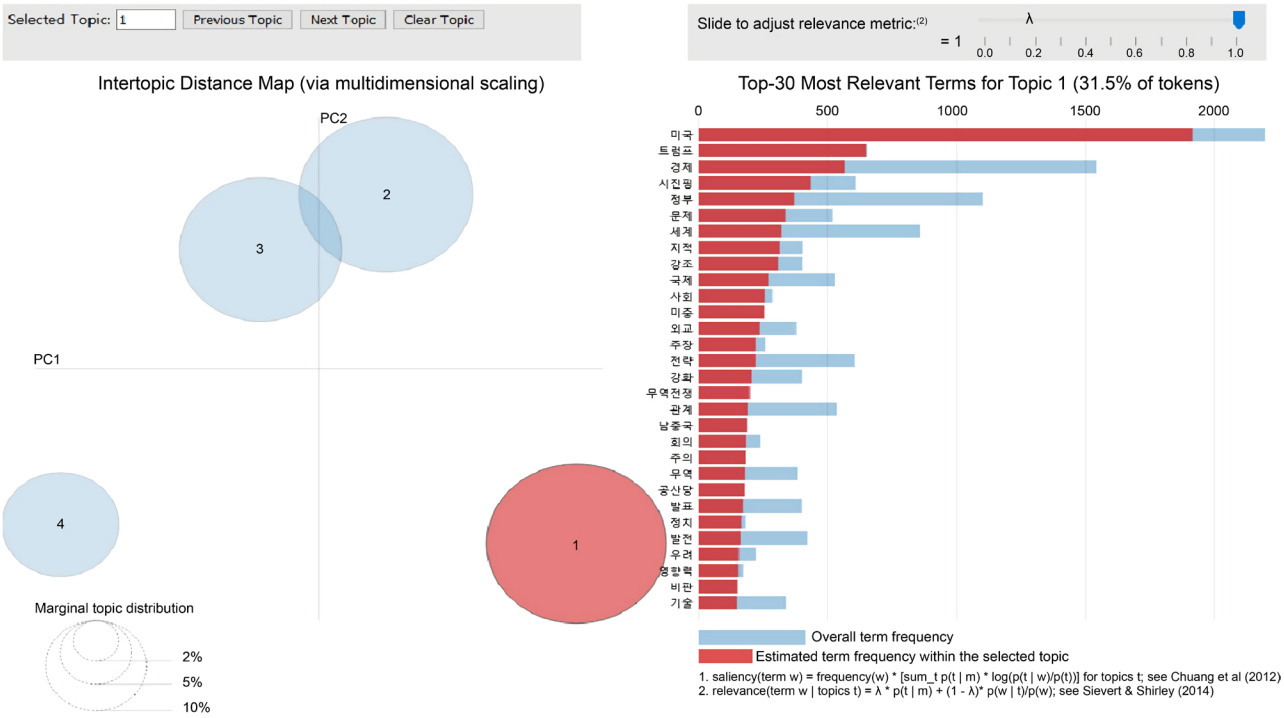

Figure 5. LDAvis visualization of "The Belt and Road" related data.

As shown above, the histogram on the right is the 30 most important words in each topic. The four circles on the left represent the four topics derived from the LDA algorithm, and the area of the circle indicates the relative dominance of each topic. The distance between the circles indicates how well the judgment is done. In other words, the degree of discrimination of the topic is proportional to the distance of the circle. If the circle and the circle overlap, it means that the degree of discrimination is low, so the area of overlap between the circle and the circle is large enough or a circle contains a small circle, which can be regarded as similar content. As can be seen from the above figure, the degree of independence between the four themes is very high. Although the circle 2 and the circle 3 overlap slightly, the overlapping area is not very large, so the content of the topic is not greatly affected. Table 5 is the result of sorting the topic words in the LDA visualization result of the above figure in order of appearance probability from high to low. The probability mentioned here refers to the probability of a word appearing to form a topic.

Under the premise of not arbitrarily excluding high frequency words, among the 30 words appearing in each topic, they are selected and selected if necessary, and the topics are extracted from the 10 words extracted. The results are shown in the Table 6.

The name of the topic represented by the word extracted according to the LDA algorithm should be given the name by the researcher. For example, in Topic 1, the main core words are "the United States", "Trump", "Xi Jinping", "Sino-US", "Diplomacy" and other words. According to the meaning of each

Table 5. LDA topic model analysis results.

Topic 1	Topic 2	Topic 3	Topic 4
미국 (US) (0.35)	기업 (Enterprise) (0.27)	한국 (South Korea) (0.23)	지역 (Region) (0.20)
트럼프 (Trump) (0.33)	시장 (Market) (0.26)	조선 (North Korea) (0.22)	인도 (India) (0.19)
경제 (Economy) (0.28)	투자 (Investment) (0.23)	협력 (Unite) (0.20)	중앙아시아 (Central Asia) (0.17)
시진핑 (Xi Jinping) (0.20)	달러 (Dollar) (0.21)	경제 (Economy) (0.18)	사업 (Career) (0.15)
정부 (Government) (0.16)	경제 (Economy) (0.18)	러시아 (Russia) (0.17)	유럽 (Europe) (0.14)
문제 (Problem) (0.11)	위안 (Yuan) (0.18)	양국 (Two Countries) (0.15)	고속철 (High-speed Rail) (0.12)
세계 (World) (0.06)	수출 (Output) (0.17)	방문 (Visit) (0.15)	세계 (World) (0.12)
지적 (Indication) (0.05)	해외 (Aboard) (0.15)	아시아 (Asia) (0.13)	연구 (Research) (0.10)
강조 (Emphasize) (0.04)	정부 (Government) (0.11)	관계 (Relationship) (0.11)	몽골 (Mongolia) (0.08)
국제 (International) (0.04)	규모 (Scale) (0.09)	정상회담 (Summit) (0.10)	도시 (City) (0.06)

Table 6. The content of the exported topic.

Topic	Main Words	Main Content
1	미국 (US) 트럼프 (Trump) 시진핑 (Xi Jinping) 정부 (Government) 세계 (World) 국제 (International)사회 (Society) 중미 (Sino-Us)외교 (Diplomacy) 전략 (Strategy)	Sino-US Diplomacy
2	기업 (Enterprise) 시장 (Market) 투자 (Investment) 달러 (Dollar) 경제 (Economy) 수출 (Output) 해외 (Aboard) 규모 (Scale) 수입 (Income) 은행 (Bank)	Foreign Capital Situation
3	한국 (South Korea) 조선 (North Korea) 협력 (Unite) 러시아 (Russia) 방문 (Visit) 아시아 (Asia) 관계 (Relationship) 유럽 (Europe) 정상회담 (Summit) 유라시아 (Eurasia)	Inter-Korean Relations
4	지역 (Region) 인도 (India) 중앙아시아 (Central Asia) 사업 (Career) 유럽 (Europe) 고속철 (High-speed Rail) 몽골 (Mongolia) 연결 (Connect)력사 (History) 카자흐스탄 (Kazakhstan)	Central Asia

word and the actual situation, it can be judged that this is related to the diplomacy between China and the United States, so that the main content name can be set as Sino-US diplomacy. And the other three themes are the same. Using this method, you can accurately understand what vocabulary and how the core theme is used in each text.

5. Conclusions

By using text mining technology to analyze Chinese and Korean news data,

combined with the actual background and situation, we can draw relatively objective and accurate conclusions. In China, it can be divided into the following three points: first, the core of "The Belt and Road" is not in the Korean peninsula. Second, in terms of diplomacy, "The Belt and Road" extends eastward, and it is very necessary to strengthen the cooperative relations with Russia, North Korea and South Korea. Third, the United States will be the biggest drag factor in the implementation of the "The Belt and Road" initiative. Regarding the Korean Peninsula, it can be divided into two major aspects: First, the combination of the "The Belt and Road" and the Korean Peninsula is still based on the security situation of the peninsula itself. It can be seen from the analysis that the security risks of the Korean Peninsula still exist, and it is very important to solve its security problems. Second, the United States has a considerable influence on the Korean Peninsula, and its threat is not small.

In view of the influence and changes brought to various regions and countries after the "One Belt And One Road" policy was put forward, the following Suggestions and prospects are made in connection with the development and trend of the following. First of all, the trend of the "The Belt and Road" extension is imperative. The opening of the Arctic Route, the promotion of the China-Japan-Korea Free Trade Agreement, and the new era relationship between China and Russia have all had a positive impact and significance. Secondly, regarding the Korean Peninsula, the situation on the Korean Peninsula is crucial, and the future prospects are more optimistic. In the future, we need to focus on the common focus, solve problems, and need to strengthen communication and communication on the differences in understanding problems.

It can be seen that the actual effect of using text mining technology to analyze text data is still very obvious. The core topics and content in the text can be accurately extracted, so from the perspective of demonstrating whether the topic-based text analysis method is feasible, the answer is yes. In addition, considering that the topic model algorithm is simply used to extract topics from a large amount of text data, by analyzing the structure and content of individual topics. It is possible to give importance significance from the aspects of improving individual topic analysis functions. However, from the perspective of whether the analysis results have 100% objective accuracy, it is necessary to expand and study this aspect.

Funding

This work is supported by Jilin Province Education Department "13th Five-Year" Science and Technology Research Project (Project Number: JJKH20191119KJ).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hu, C., Lin, J. and Cui, Z. (2019) Research on the Regional Economic Development

Situation Based on the "One Belt and Road" of Big Data. *China Market*, **11**, 4-16.

[2] Wang, L. (2018) The Path of Strengthening "One Belt and Road" to Foreign Cultural Communication. *Youth Journalist*, **10**, 31-32.

[3] Lv, Y. and Wang, H. (2018) Application of Big Data in Logistics Management in Belt and Road Initiative. *Logistics Technology*, **37**, 25-28.

[4] Noh, Y., Kim, T., Jeong, D.-K. and Lee, K.H. (2019) Trend Analysis of Convergence Research Based on Social Big Data. *Journal of the Korea Contents Association*, **19**, 135-146.

[5] Kim, M., Koo, C. and Sohn, B. (2019) A Study on the Effectiveness of Education Welfare Priority Support Program through Text Mining. *Korean Journal of Youth Studies*, **26**, 313-332. https://doi.org/10.21509/KJYS.2019.02.26.2.313

[6] Zhou, J., Xiong, Z. and Zhang, Y. (2006) Multi-Center Clustering Algorithm Based on Max-Min Distance Method. *Journal of Computer Applications*, **26**, 1425-1427.

[7] Wang, Y. and Tang, J. (2014) High-Efficiency K-Means Optimal Clustering Number Determination Algorithm. *Journal of Computer Applications*, **34**, 1331-1335.

[8] An, J., An, G. and Shi, Z. (2015) An Improved K-Means Text Clustering Algorithm. *Transducer and Microsystem Technologies*, **34**, 130-133.

[9] Dhillon, I.S. and Modha, D.S. (2001) Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning*, **42**, 143-175. https://doi.org/10.1023/A:1007612920971

[10] Sun, J.G., Liu, J. and Zhao, L.Y. (2008) Clustering Algorithms Research. *Journal of Software*, **19**, 48-61. https://doi.org/10.3724/SP.J.1001.2008.00048

[11] Wang, C. and Zhang, J. (2014) Application of Improved K-Means Algorithm Based on LDA in Text Clustering. *Journal of Computer Applications*, **34**, 249-254.

[12] Blei, D.M., Ng, A.Y. and Jodan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[13] Steyvers, M. and Griffiths, T. (2007) Probabilistic Topic Models. In: Landauer, T.K., Mcnamara, D.S., Dennis, S. and Kinitsch, W., Eds., *Latent Semantic Analysis: A Road to Meaning*, Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

[14] Timothy, R.S. (2012) Deciphering North Korea's Nuclear Rhetoric: An Automated Content Analysis of KCNA News. *Asian Affairs: An American Review*, **39**, 73-89. https://doi.org/10.1080/00927678.2012.678128