Postprint: News Topic Mining and Analysis of the Belt and Road Initiative on the Chinese Government Website

Authors: Qin Yue, Wu Yaping, Wang Jimin

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] To investigate the topic content and popularity evolution of "Belt and Road" related news on the Chinese government website, present the themes and dynamics of the "Belt and Road" Initiative, identify the focus of the initiative in different periods, and provide references for related research.

[Method/Process] Construct a basic framework for news topic content based on the LDA model, restrict the dataset to "Belt and Road" related news from 2015-2017, employ the LDA model for topic extraction, and analyze the popularity evolution of each theme across different time periods based on probability distribution calculations between documents and topics.

[Results/Conclusion] The extraction yielded 30 sub-topics, categorized into seven major classes: policy coordination, facilities connectivity, unimpeded trade, financial integration, people-to-people bonds, the impact of the "Belt and Road" on China's economy, and government work. Among these, the policy coordination category exhibited the highest popularity throughout the entire timeframe, followed closely by unimpeded trade and the impact of the "Belt and Road" on China's economy. The popularity of sub-topics such as "import and export" continued to rise, while that of sub-topics like "reform and transformation" declined, demonstrating that official media news content and its associated attention evolve over time.

Full Text

Preamble

Mining and Analysis of "Belt and Road" News Topics on the Chinese Government Website *ChinaXiv Collaborative Journal, Vol. 63, No. 15, August 2019*

Qin Yue[1], **Wu Yaping**[2], **Wang Jimin**[1] [1]Department of Information Management, Peking University, Beijing 100871 [2]Peking University Library, Beijing 100871

Abstract

[**Purpose/Significance**] This study investigates the topic content and popularity trends of "Belt and Road" related news on the Chinese government website, presenting the themes and dynamics of the Belt and Road Initiative, identifying its priorities across different periods, and providing references for related research. [**Method/Process**] The study constructs a basic framework for news topic analysis based on the LDA model, focusing on Belt and Road news data from 2015 to 2017. Using the LDA model for topic extraction, the research analyzes the evolution of topic popularity across different time periods based on probability distributions between documents and topics. [**Result/Conclusion**] Thirty sub-topics were extracted and categorized into seven major categories: policy coordination, facilities connectivity, unimpeded trade, financial integration, people-to-people bonds, the Belt and Road Initiative's impact on China's economy, and government work. Among these, policy coordination topics showed the highest popularity across all time periods, followed by unimpeded trade and the Initiative's impact on China's economy. The popularity of sub-topics such as "import and export" continued to rise, while topics like "reform and transformation" declined, reflecting how official media content and its attention evolve over time.

2 Related Research

2.1 Belt and Road News Research

Current research on Belt and Road news primarily focuses domestically and concentrates on two aspects: news reporting framework studies and quantitative analysis of news content. News frameworks refer to the specific principles that news media apply when selectively processing news facts. Reporting framework research typically involves analyzing news length, sources, and topic selection to summarize framework characteristics. For instance, Yao Yujiao selected Belt and Road reports from *People's Daily*, examining material selection, construction, content, and themes to explore the production process, information framing, and fact construction features, concluding that *People's Daily* formed a news framework emphasizing positive guidance, policy propaganda, and achievement demonstration. Zeng Runxi et al. analyzed 118 Belt and Road news reports from 18 mainstream media outlets including People's Daily Online and Xinhua Net, finding that current news content suffers from heavy propaganda tones, repetitive focus points, and singular emphasis on China's role, which hinders society's correct understanding of the Initiative.

Quantitative analysis research on news content focuses more on the content itself, attempting to explore thematic content. Wang Haizao selected 114 articles

from *China Daily* and 466 from *China Daily USA Edition*, extracting keywords through word frequency statistics and categorizing them, finding that *China Daily* focused on economic categories while the US edition emphasized comprehensive reporting. Tian Zuoyu selected 594 Indian English news articles on the Belt and Road, combining corpus discourse analysis with sentiment dictionaries and using word analysis, diachronic keyword analysis, and word cluster analysis to explore India's interpretation and evaluation of the Initiative, discovering that Indian news coverage focused on leader visits, China's cooperation with India's neighbors, and the AIIB establishment, with media attitudes containing complex emotions including skepticism and speculation.

Overall, current Belt and Road news research is dominated by news reporting framework studies, with fewer quantitative content analyses, and these quantitative analyses mostly use word frequency statistics and word cluster analysis, making deep topic mining difficult.

2.2 News Topic Extraction and Evolution Analysis Research

Topic Detection and Tracking (TDT) technology automatically identifies new topics and continuously tracks known topics in news media information streams, becoming a hotspot in information processing in the era of information explosion. Topic extraction and evolution analysis are applications of TDT. Common news data modeling methods include vector space model-based approaches, language model-based approaches, and probabilistic topic model-based approaches.

The Vector Space Model (VSM), proposed by G. Salton et al. in the 1970s, represents documents as vectors and simplifies text content processing to vector operations in vector space. J. Allan et al. represented broadcast news reports as feature vectors, using VSM to identify feature vectors corresponding to news topics and determine whether new articles belong to known topics or represent new topics, achieving topic detection and tracking. Lin Nan proposed a TD-VSM model suitable for topic detection based on news reporting structure and temporal characteristics, using information entropy and structural features to improve TF-IDF weight calculation and combining temporal features to improve cosine similarity calculation for news topic identification.

The Language Model (LM), first proposed by M. Spitters in 2002, performs mathematical abstraction of language based on objective linguistic facts, including N-gram models and decision tree models. V. Lavrenko et al. used a special unigram language model—the relevance model—to dynamically expand information on existing topic-related news documents, improving the comprehensiveness of topic models. C. Zhai et al. studied language model smoothing issues and their impact on retrieval performance, finding that retrieval performance is not only sensitive to smoothing parameters but also that sensitivity patterns are affected by query types, demonstrating performance advantages.

Probabilistic Topic Model (PTM) theory originated from T. Hofmann's probabilistic latent semantic analysis (pLSA) model proposed on the basis of latent

semantic analysis (LSA). This model assumes that each document is randomly generated from a multinomial distribution of topics, and different topics generate different words. In 2003, D. M. Blei et al. proposed the LDA (latent Dirichlet allocation) model, a three-layer Bayesian generative probability model that simulates document collections as finite mixtures of latent topics, which are in turn composed of several feature words. Many improved and extended probabilistic topic models have since emerged, such as dynamic topic models considering temporal factors. L. Alsumait et al. proposed an improved online topic model OLDA that automatically identifies new topics in incoming news documents and incrementally updates the topic model based on information inferred from new data streams, timely grasping changes in each topic over time. Chu Keming et al., using Two Sessions news as an example, proposed a method to mine how news topics change over time by first using LDA to extract topics from document sets in different time periods, then calculating distribution distances between any two topics in adjacent periods to discover content associations and derive topic evolution.

Besides these three models, lexical chain models and graph models have also enriched news topic extraction and evolution analysis research. Overall, VSM, despite its wide application, has defects as it doesn't consider semantic associations between words. Language models lack accuracy for breaking news topics and haven't become mainstream. Probabilistic topic models have good generalization capabilities and can be extended to achieve good results in processing short texts, making them widely applied. Current Belt and Road news research mainly focuses on news framework studies, lacking in-depth research on news content itself. As a type of probabilistic topic model, LDA has strong topic identification capabilities and has been applied in topic discovery, text classification, text clustering, sentiment analysis, and other fields with good results. Therefore, this paper selects the LDA model to analyze Belt and Road news and explore topic composition and popularity evolution.

3 Topic Extraction and Evolution Analysis Framework

3.1 Topic Extraction Framework

The LDA model is a three-layer Bayesian generative probability model containing word, topic, and document structures, simulating documents as finite mixtures of latent topics. In LDA's three-layer structure, words are first assumed to be mixed from topic probability distributions, and documents are then assumed to be mixed from latent topic probability distributions. For each document, the topic proportion contained in the document is first sampled from a Dirichlet distribution, and then each word in the document is generated based on the topic and word probability distributions. The following steps describe the document generation process in the LDA model in detail, with symbols and meanings shown in Table 1 .

(1) For document d in the document collection, generate the topic distribution

on the document according to d~Dirichlet(α);

(2) Generation of the ith word wdi in document d:

- Generate a topic zk~Multinomial(d);
- Generate a word that maximizes p(wdi|ϕk) according to ϕk~Dirichlet(β).

Table 1: Symbol Meanings in the LDA Model

Symbol	Meaning
d	Document
i	The ith word in document d
wdi	The ith word in document d
α	Prior parameter for topic-word distribution
β	Prior parameter for document-topic distribution
d	Multinomial distribution of topics in document d
ϕk	Multinomial distribution of topic k on the vocabulary

The LDA model introduces α and β to complete the document generation process and uses Gibbs Sampling, expectation propagation, and other methods to approximate inference for parameters and ϕ, obtaining document topics. The key lies in solving the current word sampling probability, with the posterior estimates for and ϕ expressed as:

$$\hat{\theta}_{dj} = \frac{C_{dj}^{DK} + \alpha}{\sum_{j=1}^{K} C_{dj}^{DK} + K\alpha}$$

$$\hat{\phi}_{ij} = \frac{C_{ij}^{VK} + \beta}{\sum_{i=1}^{V} C_{ij}^{VK} + K\beta}$$

where K represents the number of topics, C_{dj}^{DK} represents the number of words in document d assigned to topic j, $\sum_{j=1}^{K} C_{dj}^{DK}$ represents the total number of words assigned in document d, C_{ij}^{VK} represents the number of times word i is assigned to topic j, and $\sum_{i=1}^{V} C_{ij}^{VK}$ represents the total number of words assigned to topic j.

Topic coherence measures the semantic similarity between high-probability words under a topic and can be used to evaluate models like LSA and LDA. For topics in a model, if high-probability words under a topic have high semantic similarity, the topic is considered to have high coherence and the model performs well. By setting the number of topics to multiple equidistant values (N_1, N_2, … N) and calculating the topic coherence for each, the highest coherence value corresponds to the optimal number of topics. This paper uses this metric as the basis for selecting the number of topics.

The topic extraction process is shown in Figure 1 [Figure 1: see original paper]. After collecting the Belt and Road news document set from the Chinese government website, the dataset undergoes preprocessing including data cleaning and word segmentation. The optimal number of topics is selected based on topic coherence metrics, topics are extracted using the LDA model, and topic content is categorized to achieve content mining of news documents.

3.2 Topic Popularity Evolution Analysis Framework

Topic popularity is generally represented by the association between topics and documents. The same topic may appear in various documents with different importance levels; a topic mentioned by more articles has higher popularity. By calculating a topic's popularity in different time periods, we can reflect trends in topic popularity over time and achieve evolution analysis. Topic popularity is calculated based on document-topic distribution, specifically by computing the average probability of a topic appearing across all documents. For example, the popularity of topic zk in a certain time period can be expressed as:

$$\text{Popularity}(z_k) = \frac{\sum_{d \in D} \theta_{dk}}{|D|}$$

where D represents the document set in a certain time period, |D| represents the number of documents in set D, d represents a document in D, and dk represents the probability of topic zk appearing in document d.

Following the LDA model's definition of a topic as a set of semantically related words and their probability distribution values under the topic, which can be represented as:

$$Z = \{(w_1, p(w_1|z)), (w_2, p(w_2|z)), ..., (w_V, p(w_V|z))\}$$

where Z represents the topic, wi represents the ith word, p(wi|z) represents the probability of the ith word appearing under topic Z, and V represents the vocabulary size.

After extracting topics from all news documents, we first calculate the overall popularity ranking of each topic across the entire time period. The document set is then discretized into time windows according to publication dates. Using the document-topic distribution matrix obtained from the LDA model, we calculate the popularity of each topic in each time window to obtain topic popularity changes over time. Based on each topic's popularity trend, topics are classified into rising, declining, and fluctuating categories to obtain popularity evolution patterns. The specific topic popularity evolution analysis framework is shown in Figure 2 [Figure 2: see original paper].

4 Experimental Process and Results Analysis

4.1 Data Collection

When using the LDA model for topic extraction from news documents, setting the number of extracted topics is critical. Topic coherence serves as the metric for selecting the optimal topic number. Using "Belt and Road," "Silk Road Economic Belt," and "21st Century Maritime Silk Road" as search keywords, with time limited to 2015-2017, we obtained 8,069 news documents after deduplication. The annual distribution is shown in Table 2 .

To improve experimental accuracy, the collected raw data was preprocessed by: removing meaningless characters from news text; adding Belt and Road related terms to a user-defined dictionary to prevent incorrect segmentation; using Jieba for word segmentation and filtering stop words, personal names, and other words with low topic discriminative power.

4.2 Topic Extraction and Evolution Analysis Results

4.2.1 Topic Extraction Results Using topic coherence as the evaluation metric, we experimentally selected an appropriate number of topics. Results are shown in Figure 3 [Figure 3: see original paper]. When the number of topics is set to 30, the semantic similarity among words in topics is highest and topic coherence is strongest. Therefore, 30 was determined as the optimal number of topics, yielding 30 Belt and Road related news topics. Combining topic keywords and corresponding news document content, we categorized these 30 topics into seven major categories: policy coordination, facilities connectivity, unimpeded trade, financial integration, people-to-people bonds, the Belt and Road Initiative's impact on China's economy, and government work. The specific sub-topics, topic words, popularity values, and rankings are shown in Table 3 .

The analysis reveals: (1) Policy coordination covers intergovernmental cooperation and communication among Belt and Road countries, demonstrating how governments communicate and build political trust and consensus, including 7 sub-topics such as leader meetings; (2) Facilities connectivity covers infrastructure construction in Belt and Road countries, forming infrastructure networks connecting countries, including 3 sub-topics such as transportation construction; (3) Unimpeded trade covers investment and trade cooperation, addressing investment and trade facilitation issues, including 6 sub-topics such as industrial innovation; (4) Financial integration covers financial cooperation and regulation, including cross-border financial services; (5) People-to-people bonds cover the inheritance of Silk Road friendly cooperation spirit, with exchanges in cultural, academic, and other fields, including 6 sub-topics such as scientific research innovation; (6) The Belt and Road Initiative's impact on China's economy focuses on China's economic changes and development after the Initiative, including 4 sub-topics such as reform and transformation; (7) Government work focuses on China's government work under the Initiative, including 3 sub-topics such as

institutional management.

The top 5 sub-topics by overall popularity are: reform and transformation, leader meetings, industrial innovation, Boao Forum for Asia, and transportation construction. Their meanings are: (1) Reform and transformation discusses the Belt and Road Initiative's impact on China's economic transformation, highlighting advantages such as improved investment and trade facilitation, optimized trade structure, and enhanced technology levels that promote liberalization, marketization, and internationalization while strengthening technology's role in the economy; (2) Leader meetings involve visits, talks, and congratulatory messages among Belt and Road leaders, describing how leaders communicate cooperation and build mutually beneficial partnerships; (3) Industrial innovation mentions the establishment and development of industrial innovation parks such as the Yangtze River Economic Belt, Yangtze River Delta urban agglomeration, and Changji New District; (4) Boao Forum for Asia is an international conference organization initiated by 25 Asian countries and Australia to enhance exchanges and cooperation, with the Belt and Road Initiative becoming an increasingly important agenda item; (5) Transportation construction involves railway development such as China-Europe Railway Express and land-water intermodal transport channels, demonstrating efforts to improve road connectivity and achieve comprehensive international logistics.

4.2.2 Topic Popularity Evolution Analysis Results The popularity of the seven topic categories across the entire time period is shown in Figure 4 [Figure 4: see original paper]. Policy coordination has the highest popularity at 0.32, accounting for nearly one-third of total popularity. Unimpeded trade ranks second at 0.20, and the Initiative's impact on China's economy ranks third at 0.17. The other four categories have relatively low popularity values, all below 0.10.

Documents were discretized by quarter from Q1 2015 to Q4 2017, totaling 12 time periods. The popularity of each topic in each period was calculated using the document-topic distribution matrix from the LDA model. The overall results show that the seven topic categories' popularity fluctuates little, but changes in official media focus can be seen through sub-topic evolution. For example, "reform and transformation" shows declining popularity while "import and export" shows rising popularity, and "transportation construction" shows fluctuating popularity.

The popularity values and trends of 30 sub-topics across periods are shown in Table 4 , where background color intensity indicates popularity level (darker = higher). Arrows indicate trends: upward-right arrows show rising popularity, downward-right arrows show declining popularity, and horizontal lines indicate stable fluctuation without clear trends.

Conclusion

This paper conducted topic extraction and popularity evolution analysis of Belt and Road news on the Chinese government website using the LDA model, examining how official media attention to different topics changes over time. Thirty Belt and Road topics were extracted, belonging to seven categories: policy coordination, facilities connectivity, unimpeded trade, financial integration, people-to-people bonds, the Initiative's impact on China's economy, and government work. Policy coordination has the most sub-topics (7), covering the richest content, while financial integration has the fewest. Across the entire time period, policy coordination, unimpeded trade, and the Initiative's economic impact have the highest popularity, accounting for about 70% of total popularity. While overall popularity of the seven categories fluctuates little, sub-topic evolution reveals changes in official media focus, with topics like "reform and transformation" declining and "import and export" rising. This probabilistic topic model-based analysis provides deeper insights into Belt and Road news content, supplementing current research. Future work could expand data sources to systematically cover various official media for more comprehensive results and explore identification of associations between news topics for deeper evolution analysis.

References

[1] Du Debin, Ma Yahua. "The Belt and Road": A geo-strategy for the rejuvenation of the Chinese nation[J]. Geographical Research, 2015, 34(6): 1005-1014.

[2] Yao Yujiao. A Study on the News Framework of "Belt and Road" Reports in People's Daily[D]. Urumqi: Xinjiang University, 2017.

[3] Zeng Runxi, Wei Feng. Research on public opinion guidance evaluation of the "Belt and Road" national strategy[J]. Journal of Intelligence, 2017, 36(5): 90-94.

[4] Wang Haizao. A Comparative Study of "Belt and Road" Reports in China Daily and China Daily USA Edition[D]. Guangzhou: Guangdong University of Foreign Studies, 2017.

[5] Tian Zuoyu. A Corpus-based Study of Attitudes in Indian English Newspapers' "Belt and Road" Related News[D]. Beijing: Beijing Foreign Studies University, 2017.

[6] SALTON G, YANG C S. On the specification of term values in automatic indexing[J]. Journal of documentation, 1973, 29(4): 351-372.

[7] ALLAN J, PAPKA R, LAVRENKO V. On-line new event detection and tracking[C]//ACM SIGIR Forum. Amherst: University of Massachusetts, 1998: 37-45.

[8] Lin Nan. Research on Topic Identification and Tracking Technology Based on Web Public Opinion[D]. Fuzhou: Fuzhou University, 2014.

[9] Chen Long. Research and Application of News Hot Topic Discovery and Evolution Analysis[D]. Nanjing: Nanjing University of Science and Technology, 2017.

[10] LAVRENKO V, ALLAN J, DEGUZMAN E, et al. Relevance models for topic detection and tracking[C]//Proceedings of the second international conference on human language technology research. San Francisco: Morgan Kaufmann Publishers Inc., 2002: 115-121.

[11] ZHAI C, LAFFERTY J. A study of smoothing methods for language models applied to ad hoc information retrieval[C]//Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2001: 334-342.

[12] HOFMANN T. Probabilistic latent semantic analysis[C]//Proceedings of the fifteenth conference on uncertainty in artificial intelligence. San Francisco: Morgan Kaufmann Publishers Inc., 1999: 289-296.

[13] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3: 993-1022.

[14] BLEI D M, LAFFERTY J D. Dynamic topic models[C]//Proceedings of the 23rd international conference on machine learning. New York: ACM, 2006: 113-120.

[15] ALSUMAIT L, BARBARÁ D, DOMENICONI C. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking[C]//Proceedings of the 8th IEEE international conference on data mining. Washington: IEEE Computer Society, 2008: 3-12.

[16] Chu Keming, Li Fang. News topic evolution based on LDA model[J]. Computer Applications and Software, 2011, 28(4): 4-7.

[17] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proc. national academy of sciences, 2004, 101(1): 5228-5235.

[18] Zhou Zhenyu. A Comparative Study of Topics in Weibo and Traditional Media Based on LDA[D]. Shanghai: Shanghai Jiao Tong University, 2013.

[19] STEVENS K, KEGELMEYER P, ANDRZEJEWSKI D, et al. Exploring topic coherence over many models and many topics[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju island: Association for Computational Linguistics, 2012: 952-961.

Author Contributions

Qin Yue: Conceptualization, research framework development, writing; Wu Yaping: Revision, framework adjustment; Wang Jimin: Research design, paper revision.

An Analysis of News Topics Mining Based on LDA Model: Taking "The Belt and Road" Related News as an Example

Qin Yue[1], Wu Yaping[2], Wang Jimin[1] [1]Department of Information Management, Peking University, Beijing 100871 [2]Peking University Library, Beijing 100871

Abstract: [Purpose/significance] This paper conducted a LDA topic analysis on "the Belt and Road" related news content in official media and built a basic framework of news topic analysis using LDA model to help the public understand the dynamics and progress of the initiative and its focus in different periods. [Method/process] This paper selected "the Belt and Road" related news on the Chinese government website during 2015 to 2017, and conducted the topic extraction and heat evolution analysis using LDA model. [Result/conclusion] A total of 30 topics were extracted and summarized as seven categories called policy coordination, facilities connectivity, unimpeded trade, financial integration, people-to-people bond, economic impact and government work. Among them, the policy coordination category has the highest heat during whole time period. Unimpeded trade category and economic impact category are the second and third highest. The heat of some topics, such as "reform and transformation", decline over time, while others like "import and export" increase. These results reflect the changes in the attention of the official media to different news topics related with "the Belt and Road".

Keywords: "The Belt and Road" LDA model topic extraction heat evolution

Note: Figure translations are in progress. See original paper for figures.