Pacific International Journal, Vol. 7(3); 2024
ISSN (Print) 2663-8991, ISSN (Online) 2616-4825[1]
DOI: 10.55014/pij.v7i3.616
https://rclss.com/index.php/pij

Comparison of LDA and BERTopic in News Topic Modeling: A Case Study of *The New York Times'* Reports on China

Liu Yijia
Hunan Normal University, Changsha, Hunan, China
Email: 353161713@qq.com

Abstract: This study presents a comparative analysis of Latent Dirichlet Allocation (LDA) and BERTopic, two prominent topic modeling techniques, focusing on their efficacy in news topic modeling using The New York Times' reports on China in 2023 as a case study. The introduction underscores the significance of big data in the digital transformation era and highlights the role of machine learning, natural language processing (NLP), and topic modeling in revealing hidden patterns within vast datasets. The study aims to explore and compare the effectiveness of LDA and BERTopic in analyzing news texts related to China, analyze their strengths and weaknesses, and highlight their roles in uncovering hidden patterns within news datasets. The methodology section elaborates on data collection and preprocessing, utilizing *The New York Times* reports and Dow Jones Factiva database. The analysis sections dicusses the effectiveness, merits, and limitations of LDA and BERTopic in news topic modeling. The study concludes with findings indicating the efficacy of both techniques, offering recommendations for future research to explore hybrid approaches and enhance the generalizability of these models across various domains and languages. This study hopes is contribute to modeling techniques and their applications in news analysis and media influence studies.

Keywords:LDA; BERTopic; *The New York Times*; text analysis

1. Introduction

With the advent of the era of big data, human life is increasingly integrated into the torrent of digitization. The age of digital transformation, accompanied by the continuous growth of available datasets and concurrent leaps in computing capabilities, has brought unprecedented potential for change to the field of social science. These massive datasets, akin to an ocean of digital footprints, represent an aggregation of rich records of human activities accumulated by individuals and groups. The rise of big data in the 21st century has given rise to an urgent demand for advanced analytical techniques such as machine learning, natural language processing (NLP), and topic modeling. These techniques enable us to reveal hidden patterns and associations within data, reduce data dimensions, and thereby more accurately predict future trends.

LDA, known as the Latent Dirichlet Allocation model, is a classic topic model and probabilistic generative model proposed by Blei et al[1]. It is a three-layer Bayesian probability model with a core structure consisting of documents, topics, and words, and has long been highly regarded. In the field of text processing, the LDA topic model stands out for its excellent dimensionality reduction and modeling analysis capabilities, particularly effective in handling massive text data. This model can deeply explore the implicit topic information in the text, thus being widely applied in various disciplines and achieving significant research results. BERTopic[2] is an advanced topic modeling technique that combines the latest language models and employs the c-TF-IDF process to extract topics, further expanding the method of clustering embeddings. This technique independently handles the two processes of document clustering and generating topic representations, greatly enhancing the flexibility and practicality of the model. Although there have been previous studies comparing these two topic modeling methods[3-4], there is a relative scarcity of studies comparing their effectiveness in topic modeling on news texts.

In the vast context of exploring global communication and media influence, reports concerning China have emerged as not only a crucial channel for the international community to comprehend China, but also a pivotal factor in shaping its image and swaying international public opinion.

The study aims to achieve the following objectives:

- Explore and compare the effectiveness of LDA and BERTopic in analyzing news texts related to China from The New York Times in 2023.
- Analyze the strengths and weaknesses of LDA and BERTopic as advanced topic modeling methods.
- Highlight the roles of these models in uncovering hidden patterns and associations within vast datasets, particularly in the context of news analysis and media influence.

Through this endeavor, the study hopes to bring forth novel insights and inspirations that can contribute to the advancement of these fields.

2 Collection and Preprocessing of Corpus Data

[1] [Received 23 March 2024; Accepted 18 June 2024; Published (online) 20, June, 2024]

2.1 Collection of corpus data

Against the backdrop of studying global communication and media influence, reports involving China serve as a crucial window reflecting the international community’s understanding of China, rendering their research highly necessary. This article selects *The New York Times* as the research sample, fully considering the media’s advantages in authority, comprehensiveness, and influence. Through a thorough analysis of its China-related reports, we aim to more representatively and persuasively explore the focus of American media on China.

The Dow Jones Factiva database boasts a long history and extensive coverage, encompassing news and information in 28 languages, making it one of the largest databases globally. By setting the time range as “01/01/2023 - 12/31/2023” within this database, we successfully retrieved 28042 reports published by *The New York Times* during this period.

2.2 Preprocessing of corpus data

After successfully scraping all the news reports from *The New York Times* in 2023, initial preprocessing was promptly conducted on these nearly 30,000 textual corpora. Using web scraping tools, all the reports were successfully obtained in txt format, and duplicate texts were effectively removed through the cosine similarity algorithm. Subsequently, a Python program was written to filter the deduplicated texts, retaining only those mentioning the keywords “China” or “Chinese” at least three times, while excluding all others.

After the initial screening of China-related corpora, manual verification was conducted to ensure that each text centered on China. This step was crucial, as even if a text mentioned China multiple times, its main topic might still revolve around other countries. Ultimately, through meticulous manual review, the corpus texts adopted for this study were determined, consisting of 1494 China-related news reports from *The New York Times* in 2023. This process ensured the scientificity and accuracy of the research.

3 LDA Topic Modelling Analysis

In this study, to ensure the precision and scientific nature of the themes extracted through LDA modeling, it is imperative to conduct meticulous preprocessing of all sample corpora, including word segmentation and the removal of stop words, prior to officially running the model. Latent Dirichlet Allocation (LDA) is a popular topic modeling technique introduced by Blei et al. in 2001. LDA is a generative probabilistic model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. [6]Given the natural occurrence of spaces between words in English text, word segmentation can be conveniently achieved by calling the *.split(“”)* method. Additionally, it is crucial to eliminate high-frequency stop words from the corpus. Although these stop words occur frequently, they do not contribute substantially to theme extraction, such as common words like “a”, “the”, and “and”.

3.1 The effectiveness of LDA in topic modeling of news texts

Firstly, the number of topics K in LDA topic modeling is a crucial hyperparameter that needs to be manually determined. To scientifically and reasonably set the value of K, this study utilized Python tools to calculate the coherence value and conducted a visual analysis. After comprehensive consideration, the final value of K was determined to be 11, meaning that the LDA model will extract 11 topics.

After determining the number of topics, this study conducted a visual analysis of the topic modeling results using the pyLDAvis library, resulting in a dynamic HTML file. Shown by Figure 1, it can be concluded that topic 9 is about economy and trade.

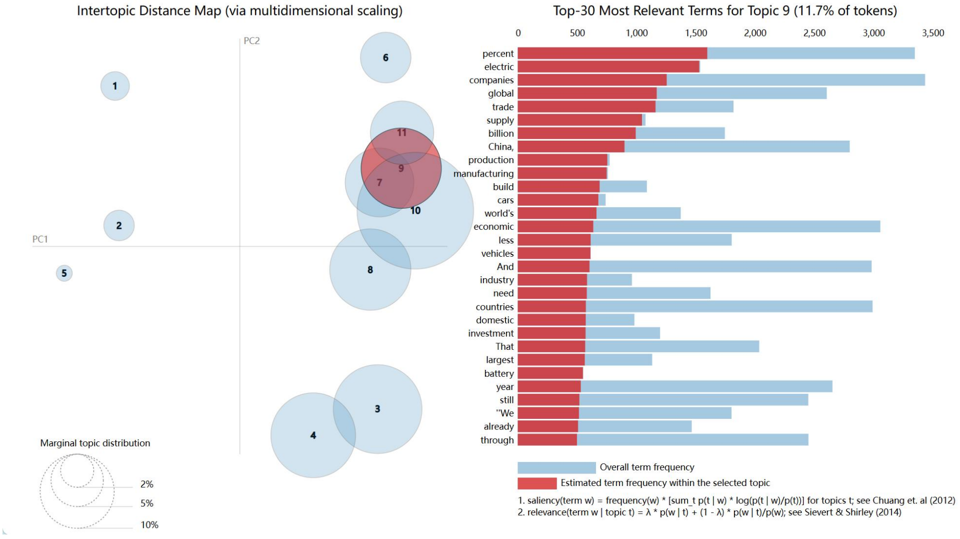

Fig.1: Visualization of LDA topic modeling

3.2 Merits of LDA

Since its establishment, the LDA model has undergone two decades of academic accumulation and development. As one of the most iconic topic modeling models in the field of natural language processing, it has always maintained a widespread influence in the social science domain, demonstrating its unique advantages in news text processing.

(1) Compared to methods based on word embeddings, LDA often extracts a more concise but core number of topics, which better aligns with the characteristics of news texts centering around limited core topics. [7]Therefore, the results of

LDA topic modeling are not only efficient but also easy to understand and interpret, enabling us to grasp the core content of news more clearly. This closely aligns with news theory, providing strong support for news analysis.

(2) News texts often involve multiple layers and perspectives, and the LDA model is able to capture this multi-topic nature effectively. For example, a news report on epidemic prevention and control may simultaneously cover the spread of the epidemic, prevention and control measures, international comparisons, and other aspects.[8] The LDA model can effectively distinguish these different thematic elements and categorize the text under corresponding multiple topics. This allows us to comprehensively understand the multi-dimensional information in news texts and delve deeper into the complexity and diversity behind the news.

(3) Another significant advantage of LDA topic modeling in news text analysis lies in its powerful dynamic visualization capabilities. [9] By utilizing LDA sequence models from third-party libraries like Gensim, we can generate dynamic HTML files that allow users to visually inspect the representative vocabulary and text distribution under each topic by clicking on different topic circles. This dynamic visualization approach not only enhances the intuitiveness and interactivity of news text topic modeling, but also aids in a deeper understanding of the associations and differences between different topics. Therefore, the combination of LDA topic modeling and dynamic visualization tools provides news analysts with a powerful and flexible tool to better explore and understand the thematic information in news texts.

(4) News text data is often vast, complex, and diverse, and the LDA model has significant advantages in handling such large-scale data. Whether it's processing thousands of news reports or millions of words of news corpora, LDA maintains high time efficiency. Additionally, for long-form news texts, LDA can effectively capture the thematic information, ensuring the accuracy and reliability of topic modeling. This makes LDA an efficient and reliable tool for processing news text data, providing powerful technical support for news analysis and mining.

3.3 Limitations of LDA

Despite the widespread application of the LDA model, the reliability and effectiveness of its results have also been questioned and criticized by some scholars[5]. With the continuous innovation and emergence of research methods, the LDA model is constantly being compared and evaluated with other advanced models in various disciplines.

(1) The LDA topic model represents text using a bag-of-words approach, which means it ignores the order of words and deeper semantic relationships. In news texts, the order of words and contextual information are crucial for understanding the development of events and the evolution of themes. However, the LDA model cannot capture these important semantic information, resulting in limited representational ability. Additionally, the soft clustering method of the LDA model may lead to overlap between topics, making it difficult to clearly distinguish certain topics. In news text analysis, this may lead to confusion and misunderstanding of events and topics.

(2) In LDA topic modeling, determining the number of topics *K* is a crucial but cumbersome task. Although numerical values such as coherence or perplexity can be calculated to assist in determining the *K* value, the process is still quite complex and time-consuming. In news text analysis, the diversity of news events and topics makes the determination of *K* even more challenging. Researchers need to repeatedly try different *K* values and evaluate the model's performance by calculating metrics such as coherence or perplexity to find the optimal number of topics. However, this calculation method often lacks clear standards, and different datasets and application scenarios may require different *K* values, making the selection of *K* more difficult and subjective. Additionally, even if a suitable *K* value is found, it cannot be guaranteed that the LDA model will accurately capture the thematic structure in news texts. Therefore, the determination of *K* not only increases the complexity of LDA topic modeling but may also affect the accuracy and reliability of the results.

(3) The LDA model primarily relies on word frequency information when extracting key topics. News texts often contain a large number of low-frequency words, which often carry significant meaning in revealing specific themes. However, the LDA model may not fully capture the contributions of these low-frequency words during topic extraction, leading to the loss of crucial information. Additionally, the LDA model is highly sensitive to word extraction results, and sometimes even adjacent words may be assigned to different topics. This can result in the fragmentation of themes in news text analysis, affecting the coherence and consistency of the themes. Therefore, when processing news texts, the LDA model may overlook the importance of low-frequency words due to its over-reliance on word frequency information, and may also lead to incoherent themes due to its sensitivity to word extraction results.

(4) Preprocessing news texts is a necessary step before running the LDA model, including word segmentation, removing stop words, etc. However, these preprocessing processes not only increase the workload of researchers but may also affect the coherence and integrity of the original text. News texts usually have specific contexts and expressions, and excessive preprocessing may destroy this important information, leading to inaccurate or distorted topic modeling results. Therefore, when using the LDA model for news text analysis, it is necessary to carefully select preprocessing methods and try to maintain the integrity and coherence of the original text.

4 BERTopic Topic Modeling Analysis

In the operational flow of BERTopic, BERT plays a crucial role as an embedder. BERTopic skillfully integrates the UMAP dimensionality reduction technique with the HDBSCAN document clustering method, enabling efficient processing and topic extraction from text data. Given BERTopic's deep reliance on word embedding methods, maintaining the original structure of the text is crucial for the model, thus eliminating the need for complex preprocessing operations on the corpus. Compared to traditional topic models such as LDA, BERTopic exhibits unique advantages. It does not require the pre-setting of the hyperparameter *K*, or manually specifying the number of topics. Instead, BERTopic can automatically extract the number of topic words based on the text data, avoiding the complex process of determining topic words. This feature not only enhances the flexibility of BERTopic in topic modeling but also improves its practicality and accuracy.

4.1 The effectiveness of BERTopic in topic modeling of news texts

In this study, the BERTopic model was utilized to extract themes from the news reports involving China in *The New York Times* in 2023. During this process, BERTopic automatically identified and refined 46 distinct themes, which were visually presented in a bar chart for intuitive understanding. Figure 2 specifically showcases the top eight themes with the highest scores and their corresponding representative vocabulary, enabling us to gain a clearer understanding of the core content and characteristics of these themes.

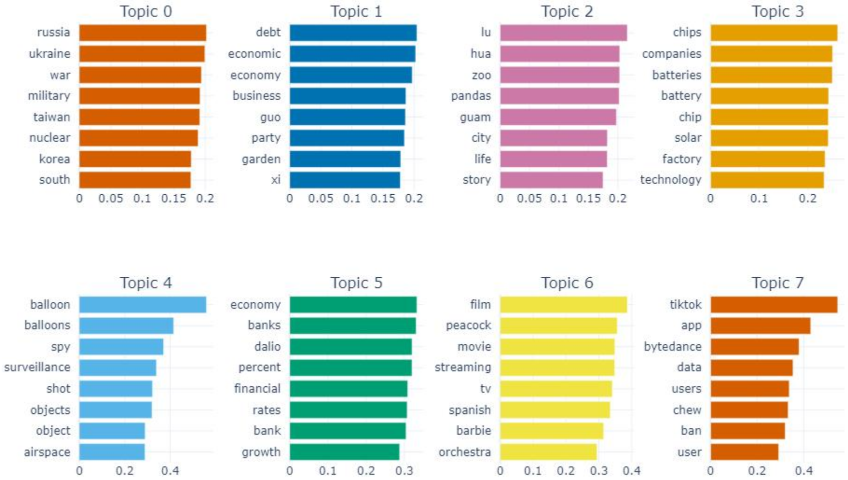

Fig.2: Top eight theme-representative term pairs extracted by BERTopic

After systematically organizing and analyzing the 46 sets of themes and their representative terms output by the model, we can further refine the secondary themes and selected and presented some representative theme contents to analyse the focal points of China-related news coverage by *The New York Times*.

4.2 Merits of BERTopic

As a novel and efficient topic modeling technique, BERTopic has attracted widespread attention from the academic community since its inception. It builds upon and innovates from previous technologies such as Top2Vec, providing a more precise and comprehensive solution for topic extraction and analysis of text data. With its gradual application in various academic fields, an increasing number of researchers are recognizing the unique advantages of BERTopic in topic modeling. In the application of topic modeling to news texts, BERTopic also demonstrates its unique strengths.

(1) Leveraging the powerful capabilities of the BERT pre-trained model, BERTopic performs deep semantic analysis on news texts to extract more meaningful and coherent themes. This helps news analysts grasp the core content of news events more accurately and understand the underlying meanings behind the news texts. Meanwhile, the application of HDBSCAN further enhances the consistency of the themes, making the extracted themes more accurate and reliable.

(2) News texts involve diverse and complex themes, and traditional topic modeling methods often require manually setting the number of topics, which brings inconvenience to researchers. However, BERTopic can automatically generate the number of topics, demonstrating wide applicability and reliability, making it suitable for extensive exploration of news text themes. Researchers can adjust parameters to reduce the number of topics if needed, making theme extraction both convenient and controllable. This intelligent approach to determining the number of topics gives BERTopic a distinct advantage in news text topic modeling.

(3) The visualization of news text topics is crucial for understanding and explaining them. BERTopic supports dynamic visual topic modeling and provides various visualization methods, enabling researchers to easily create diverse and multi-dimensional visualizations based on their research objectives. This helps researchers visually present the thematic structure of news texts, discover relationships and differences among topics, and thereby gain a deeper understanding of the inherent logic of news events.

(4) BERTopic supports multi-language models, adapting to the multilingual nature of news texts. News texts often involve multiple languages, and there may be differences in topic modeling for texts in different languages. BERTopic is suitable for over 50 language models, allowing researchers to flexibly select available models based on their actual needs. This makes BERTopic more adaptable when processing multilingual news texts, enabling it to more accurately extract thematic information from texts in different languages and providing powerful support for cross-language news analysis.

4.3 Merits of BERTopic

Although BERTopic has shown significant advantages in topic modeling, it still inevitably has some limitations in practical applications.

(1) BERTopic can generate more coherent topics in news text topic modeling, but it may produce outliers (with a value of -1) during the process. These outlier topics need to be excluded in subsequent analysis, otherwise they will affect the accurate understanding and analysis of news text topics. This adds complexity to data processing and may reduce the efficiency and accuracy of topic modeling.

(2) When extracting topics from news texts, BERTopic does not fully consider the situation where a news report may involve multiple topics simultaneously. For example, a news report about a Chinese family celebrating the Mid-Autumn

Festival in a Chinese restaurant covers both the traditional customs of the Mid-Autumn Festival and the culinary culture of the restaurant. However, BERTopic may only classify this news report under a single topic such as “Festival Culture” or “Traditional Food,” unable to reflect its multiple thematic characteristics. This limits the application of BERTopic in the analysis of complex news text topics.

(3) When performing topic modeling on news texts, BERTopic takes into account the contextual semantics of the documents, but the topic words are still based on the bags-of-words model. This may lead to the presence of highly similar and somewhat redundant words in the extracted topics, such as “car” and “cars” appearing in the same topic. Although the number of representative words can be adjusted by setting the n_words parameter, such redundancy may affect the accurate understanding and interpretation of the topic content in news text analysis.

(4) BERTopic may perform poorly when dealing with small datasets or corpora with long text lengths in news text topic modeling. In the field of news, with the continuous development of news media, the number of news texts is constantly increasing, and some important news reports may be quite lengthy. If BERTopic is unable to effectively handle these long texts or corpora with small data volumes, the effectiveness of its topic modeling may be affected, thus limiting its application scope in news text analysis.

5 Conclusion

This study conducted a comparative analysis of LDA and BERTopic in news topic modeling, specifically focusing on *The New York Times*' reports on China. The findings indicate that while LDA remains effective for dimensionality reduction and modeling analysis, BERTopic excels in extracting nuanced and contextually relevant topics. This comparison offers a deeper understanding of the strengths and limitations of both models in news topic modeling.

The field of topic modeling is poised to benefit from the ongoing advancements in NLP techniques and the integration of cutting-edge machine learning algorithms. Future research endeavors should explore the potential of combining LDA and BERTopic or developing hybrid approaches to leverage the unique strengths of each model. Additionally, there is a need to investigate the generalizability of these models across various domains and languages, with an emphasis on enhancing interpretability and scalability. As the volume and complexity of digital data continue to proliferate, the development of more robust and efficient topic modeling techniques will be paramount for extracting meaningful insights and informing decision-making in diverse fields.

References

[1] Blei M, Ng Y, Jordan I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003, 3(4/5): 993-1022.

[2] Grootendorst M. BERTopic: Neural Topic Modeling With a Class-Based TF-IDF Procedure. arXiv:2203.05794v0571. Available online at: https://arxiv.org/pdf/2203.05794.pdf (accessed March 15, 2022).

[3] Egger R, Yu J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Front. Sociol. 2022, 7: 886498.

[4] Ogunleye B, Maswera T, Hirsch L, Gaudoin J, Brunsdon T. Comparison of Topic Modelling Approaches in the Banking Context. Applied Sciences. 2023, 13(2): 797.

[5] Arefieva V, Egger R, Yu J. A machine learning approach to cluster destination image on Instagram. Tourism Management. 2021, 85: 104318.

[6] David M. Blei; Andrew Y. Ng; Michael I. Jordan; "Latent Dirichlet Allocation", NIPS, 2001.

[7] Yee W. Teh; David Newman; Max Welling; "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation", NIPS, 2006.

[8] Yezheng Liu; Fei Du; Jianshan Sun; Yuanchun Jiang; "ILDA: An Interactive Latent Dirichlet Allocation Model to Improve Topic Quality", JOURNAL OF INFORMATION SCIENCE, 2020.

[9] Chetan Sharma; Shamneesh Sharma; S Sakshi; "Latent DIRICHLET Allocation (LDA) Based Information Modelling on BLOCKCHAIN Technology: A Review of Trends and Research Patterns Used in Integration", MULTIMEDIA TOOLS AND APPLICATIONS, 2022.