Risk Identification in the Electronic Information Manufacturing Industry Using Annual Report Texts: A Comparative Analysis of LDA and BER Topic Topic Modeling Methods

Yuxin Ji*

School of Management, Jiangsu University, Zhenjiang 212013, Jiangsu, China

***Corresponding author**

DOI: https://doi.org/10.47772/IJRISS.2026.100900846

Received: 05 October 2026; Accepted: 10 October 2026; Published: 24 October 2026

ABSTRACT

With the rapid technological evolution of the electronic information manufacturing industry, intensified market competition, and increasing complexity of global industrial and supply chain environments, enterprises are facing increasingly diversified types of risks. Annual reports of listed companies contain abundant unstructured textual information related to operational risks, providing an important data foundation for industry risk identification. This study investigates risk-related texts extracted from the annual reports of listed electronic information manufacturing companies in the Shanghai and Shenzhen A-share markets during 2021–2024. A risk text corpus was constructed, and based on unified text preprocessing procedures, the Latent Dirichlet Allocation (LDA) model and BERTopic, a topic modeling approach based on pretrained language models, were employed to identify and comparatively analyze risk topics. The differences between the two models were evaluated from the perspectives of topic identification results, topic quality, and interpretability.The results show that LDA generates a relatively concentrated risk category structure and identifies seven risk topics, including macroeconomic risk, investment risk, market competition risk, supply chain risk, core talent risk, accounts receivable risk, and operational management risk, making it suitable for analyzing the overall risk structure of the industry. In contrast, BERTopic further refines risk information through contextual semantic representations and identifies 28 valid risk topics. It outperforms LDA in terms of both topic coherence and topic diversity, enabling more accurate differentiation of specific risk issues such as technological R&D risk, core talent loss risk, exchange rate fluctuation risk, and raw material price fluctuation risk. The findings demonstrate that different topic modeling approaches have distinct applicable scenarios for risk identification: LDA is more suitable for macro-level risk classification, whereas BERTopic is more effective for fine-grained risk discovery. This study provides methodological insights for risk text analysis in the electronic information manufacturing industry and offers a foundation for enterprises to establish a hierarchical risk identification framework of "overall risk categories–specific risk issues."

Keywords: Electronic Information Manufacturing Industry: Risk Identification; Annual Report Text; Topic Modeling; LDA; BERTopic

INTRODUCTION

The electronic information manufacturing industry is a fundamental sector supporting the development of the digital economy and plays a critical role in promoting the application of next-generation information technologies, industrial digital transformation, and manufacturing upgrading. With the acceleration of technological iteration, intensification of market competition, and increasing complexity of global industrial and supply chain environments, electronic information manufacturing enterprises face increasingly diverse and complex risks[Error! Unknown switch argument.]. On the one hand, enterprises need to continuously conduct technological research and development and product innovation, which may lead to internal risks associated with R&D investment, technological upgrading, loss of core talents, and intellectual property protection **Error! Unknown switch argument.**. On the other hand, external factors, including raw material price

fluctuations, supply chain disruptions, changes in market demand, exchange rate volatility, and international trade environments, may also significantly affect enterprise operations **Error! Unknown switch argument.**. Therefore, effectively identifying various types of risk information from large-scale corporate disclosure texts and establishing an interpretable risk classification framework are of practical significance for understanding industry risk structures and supporting enterprise risk management.

Annual reports of listed companies represent important information disclosure channels through which enterprises present their operational performance, financial conditions, and major risks to external stakeholders**Error! Unknown switch argument.**. Sections such as "Major Risk Factors," "Risk Factors," "Potential Risks," and "Management Discussion and Analysis" provide concentrated descriptions of managerial perceptions regarding the business environment and potential threats. Compared with risk analysis solely based on financial indicators, annual report texts contain abundant unstructured information and can complement structured data by capturing risks related to technology, markets, human resources, supply chains, and external environments that are difficult to fully characterize using quantitative indicators alone**Error! Unknown switch argument.**. As annual reports accumulate continuously, these textual resources constitute large-scale real-world datasets, providing an important foundation for identifying industry-level risk categories and analyzing the structure of risk information**Error! Unknown switch argument.****Error! Unknown switch argument.**. However, annual reports are typically lengthy and information-intensive, and different enterprises may employ diverse expressions when describing similar risks. Manual reading and classification are not only inefficient but also susceptible to researchers' subjective judgments, making them inadequate for large-scale textual analysis requirements[8]. Therefore, natural language processing (NLP) techniques are needed to improve the efficiency of risk information extraction and classification from annual report texts.

Topic modeling provides an effective technical approach for discovering latent semantic structures from large-scale unstructured texts. Among existing methods, Latent Dirichlet Allocation (LDA) is a classical probabilistic topic model that identifies latent topics by modeling probabilistic relationships among documents, topics, and words**Error! Unknown switch argument.**. Due to its clear model structure and relatively interpretable outputs, LDA has been widely applied in textual analysis scenarios involving annual reports, news articles, and policy documents. However, LDA primarily relies on word frequency and co-occurrence patterns for topic identification. Its bag-of-words representation ignores word order and contextual information, which may result in overly broad topic boundaries and the aggregation of different risk issues into the same topic when analyzing texts containing numerous professional terms or semantically ambiguous words**Error! Unknown switch argument.Error! Unknown switch argument.**. Consequently, the semantic coherence and fine-grained interpretability of the identified topics may be compromised.

In recent years, advances in pre-trained language models have introduced new approaches for topic modeling. BERTopic, a topic modeling framework based on pre-trained language models, generates topics through document semantic embeddings, dimensionality reduction, clustering algorithms, and class-based term frequency–inverse document frequency (c-TF-IDF) representations, enabling more effective modeling of semantic relationships among texts by incorporating contextual information**Error! Unknown switch argument.Error! Unknown switch argument.**. Compared with traditional topic models that mainly depend on word co-occurrence relationships, BERTopic demonstrates advantages in distinguishing texts with similar vocabulary but different semantic contexts, thereby providing new possibilities for fine-grained topic discovery**Error! Unknown switch argument.**. For risk-related texts in annual reports of electronic information manufacturing enterprises, different risk categories often share common keywords. For example, terms such as "technology," "R&D," "products," and "markets" may simultaneously appear in descriptions of technological innovation risks, market competition risks, core talent risks, and intellectual property risks. Therefore, effectively distinguishing texts with similar lexical characteristics but different risk implications through contextual semantic information is crucial for improving the accuracy and interpretability of risk topic identification.

Existing studies have applied topic models such as LDA to enterprise annual report analysis and risk identification, while topic modeling approaches based on pre-trained language models have gradually attracted increasing attention. However, systematic empirical comparisons between LDA and BERTopic in the specific context of risk identification from annual reports of electronic information manufacturing

enterprises remain limited. In particular, these two types of models differ substantially in their text representation mechanisms and topic generation processes, which may lead to differences in the number of generated topics, topic granularity, semantic coherence, and interpretability of results[15]. When only a single model is adopted, it is difficult to determine whether the identified risk patterns originate from the inherent characteristics of industry risk texts or from the representation methods and topic generation mechanisms of the models themselves **Error! Unknown switch argument.**. Therefore, comparing the two approaches under consistent data sources and evaluation criteria can provide deeper insights into the characteristics and applicability boundaries of different topic modeling methods for industry risk identification, thereby offering guidance for model selection in practical research scenarios[17].

Accordingly, this study investigates the annual reports of listed electronic information manufacturing companies on the Shanghai and Shenzhen A-share markets from 2021 to 2024. Risk-related sections and risk description paragraphs are extracted to construct the research corpus, and LDA and BERTopic are respectively employed for topic identification after unified text preprocessing. Rather than simply determining the absolute superiority of one model over the other, this study focuses on analyzing differences between the two approaches in terms of risk information granularity, including topic identification results, topic quality, interpretability, and model applicability. Specifically, topic quality is evaluated using C_v topic coherence and topic diversity metrics, while keyword distributions and representative risk texts are further examined to assess topic interpretability. The applicability of different models in identifying overall industry risk categories and fine-grained risk issues is also discussed. This study aims to address two research questions: (1) What risk topic structures can be identified by LDA and BERTopic from annual report risk texts of electronic information manufacturing enterprises? (2) How do different text representation mechanisms and topic generation strategies influence the granularity and practical value of risk topic identification?

The contributions of this study are twofold. From a methodological perspective, this study provides empirical evidence for selecting appropriate topic modeling approaches in annual report risk analysis by comparing a traditional probabilistic topic model with a pre-trained language model-based topic modeling method using the same risk text corpus and evaluation criteria. From a practical application perspective, this study identifies both overall risk categories and specific risk issues within the electronic information manufacturing industry, and explores a hierarchical risk identification framework of “overall risk categories–specific risk issues.” The findings provide a textual analysis foundation for enterprise risk information organization, risk discovery, and subsequent risk monitoring.

METHODOLOGY

Data Sources

This study adopts a retrospective textual analysis design. The research subjects are listed companies in the electronic information manufacturing industry on the Shanghai and Shenzhen A-share markets. Industry classification is based on the Classification of National Economic Industries, specifically “Manufacture of Computers, Communication Equipment and Other Electronic Equipment (C39),” and further validated against the statistical classification criteria for the electronic information manufacturing industry issued by the Ministry of Industry and Information Technology of China. The study period covers 2021–2024, and annual reports were collected from publicly available disclosure platforms, including CNINFO, the Shanghai Stock Exchange, and the Shenzhen Stock Exchange. During the data collection process, metadata information was recorded, including stock code, company name, sub-industry classification, listing board, reporting year, report publication date, and source of annual report texts.

The inclusion criteria were as follows: (1) companies must have complete annual reports available during the study period; (2) the full text of annual reports must be successfully parsed for textual analysis; and (3) at least one risk-related section or risk description paragraph must be identifiable within the annual reports. The exclusion criteria were as follows: (1) annual report files with severe corruption or missing key pages; (2) companies whose industry classification changed substantially during the study period and could not

maintain consistent sample criteria; and (3) reports containing only a very limited amount of risk-related text that could not constitute effective document units.

Risk Text Extraction and Corpus Construction

To reduce the interference of non-risk-related descriptions in annual reports on topic modeling, this study does not use entire annual reports as modeling documents. Instead, it focuses specifically on risk-related sections and paragraphs. Section identification keywords include “Major Risk Factors,” “Risk Factors,” and “Potential Risks.” Considering variations in annual report formats across different years, a dictionary of title variants was constructed to improve the accuracy of risk section identification. For annual reports without independent risk-related sections, secondary identification was performed within the “Management Discussion and Analysis” section based on hierarchical headings and risk-triggering keywords.

The basic document unit for topic modeling is defined as a risk-related paragraph or risk item. Each risk

paragraph is represented as document did_i, while metadata information, including company_id, year, and subsector, is retained for subsequent analysis. When the same risk item contained page-break interruptions, separated headings and body text, or other formatting inconsistencies, text merging procedures were applied during the parsing process to ensure the integrity and continuity of risk-related information.

After the above processing procedures, a risk text corpus was constructed for subsequent text preprocessing and topic modeling.

Data Preprocessing

To ensure that LDA and BERTopic are compared based on consistent textual information, the extracted risk texts were subjected to unified preprocessing procedures, and the same segmentation results were used as the foundational corpus for both models. First, the extracted risk texts were cleaned by removing semantically meaningless special characters, redundant spaces, abnormal formatting elements, and other textual noise, thereby standardizing the text format.Second, the cleaned Chinese texts were segmented using Jieba, a widely used Chinese word segmentation tool. Based on a general stop-word list, domain-specific stop words were further incorporated according to the characteristics of annual report risk texts, and high-frequency words lacking semantic relevance or discriminative power for topic identification were removed. The processed segmented texts were stored in a space-separated format and used as the basic corpus for subsequent topic modeling.

Model Description and Configuration

LDA Model Configuration

LDA is a representative generative probabilistic topic model. Its fundamental idea is to represent each document as a probabilistic mixture of multiple latent topics, while representing each topic as a probability distribution over words. By modeling the three-level probabilistic relationships among documents, topics, and words, LDA can identify latent semantic structures from large-scale unstructured texts and obtain topic distributions for individual documents as well as word distributions for each topic.In this study, the LDA model was implemented using the Gensim toolkit, and topic identification was performed based on the preprocessed risk-related text corpus described above. During model training, the number of topics KK was considered a critical parameter affecting the quality of topic partitioning.In this study, candidate topic numbers were set starting from K=2, with an interval of 5, and the optimal number of topics was determined by calculating topic coherence and perplexity under different topic settings, combined with the semantic consistency and interpretability of topic keywords. The number of training iterations was set to 200, the number of corpus passes was set to 20, α\alpha was configured as auto, η\eta was set to its default value, and the random seed was fixed at 42.

BER Topic Model Configuration

BERTopic is a topic modeling approach based on semantic representations generated by pre-trained language models. Its fundamental workflow consists of three major components: document embedding, dimensionality reduction and clustering, and topic representation.

First, pre-trained language models are employed to transform textual documents into high-dimensional semantic vectors, allowing the representations to preserve contextual semantic information. Subsequently, Uniform Manifold Approximation and Projection (UMAP) is applied to reduce the dimensionality of document embeddings, and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is used to perform density-based clustering on the reduced document vectors, thereby grouping documents with similar semantic characteristics.

Finally, representative words for each topic are extracted based on class-based term frequency–inverse document frequency (c-TF-IDF) to identify and interpret topic contents.Compared with traditional topic models that primarily rely on word frequency and co-occurrence relationships, BERTopic leverages pre-trained language models to capture contextual semantic representations of texts, making it more suitable for analyzing risk-related paragraphs with strong semantic dependencies.

In this study, the BERTopic model was constructed based on the BERTopic library. During the document semantic representation stage, BAAI/bge-large-zh-v1.5 was selected as the embedding model, and the generated document embeddings were normalized before subsequent processing. In the dimensionality reduction stage, the UMAP algorithm was adopted, with n_neighbors set to 20, n_components set to 10, and min_dist set to 0.1. Cosine distance was used as the distance metric, and the random seed was fixed at 42.For the clustering stage, the HDBSCAN algorithm was employed with min_cluster_size set to 60 and min_samples set to 5.

The cluster selection method was configured as eom (excess of mass), and Euclidean distance was adopted as the distance metric. Prediction data were retained to calculate the probability distribution of topic assignments for individual documents.During the topic representation stage, term representations were constructed based on the unified word segmentation results described above, and BERTopic's c-TF-IDF mechanism was applied to extract representative keywords for each topic.

The model did not predefine the final number of topics; instead, nr_topics was set to auto to automatically merge topics based on the initial clustering results. Documents that could not be stably assigned to any topic were identified as outliers. In addition, topic probability estimation was enabled to obtain the topic assignments and corresponding probabilities of individual risk texts, providing the foundation for subsequent topic analysis and risk exposure measurement.

Model Evaluation and Comparison Methods

To systematically compare the performance of LDA and BERTopic in risk text identification within the electronic information manufacturing industry, this study evaluates the two models from three perspectives: topic quality, result interpretability, and technical characteristics.

Regarding topic quality, based on the unified preprocessed corpus, the C_v topic coherence metric is employed to evaluate the semantic association among high-weight words within the same topic, while topic diversity is used to measure the degree of keyword overlap across different topics. For both models, the Top-10 keywords of each topic are extracted, and evaluation metrics are calculated based on the same segmented corpus to ensure consistency in the evaluation criteria. Specifically, topic diversity is defined as the proportion of unique keywords among all Top-10 keywords extracted from all topics.

Regarding result interpretability, the semantic consistency within topics, the degree of differentiation among topics, and the naming feasibility of identified risk topics are analyzed by combining high-weight keywords of each topic with representative risk-related texts. Regarding technical characteristics, the final number of valid topics identified by each model is recorded. Furthermore, the applicability characteristics of the two

models for risk text identification are analyzed by considering their parameter configurations, topic generation mechanisms, and the treatment of outlier documents.

EMPIRICAL RESULTS AND ANALYSIS

Basic Characteristics of the Sample and Risk Corpus

Following the data screening, risk text extraction, and preprocessing procedures described in Chapter 2, this study ultimately obtained 1,100 annual reports from 275 listed companies and constructed a corpus containing 6,664 valid risk-related texts. The numbers of extracted risk texts in 2021, 2022, 2023, and 2024 were 1,655, 1,682, 1,663, and 1,664, respectively. Overall, each year generated a sufficient volume of risk-related textual samples, enabling comprehensive coverage of the major risk disclosure information provided by listed companies in the electronic information manufacturing industry during the study period.

After preprocessing, a standardized risk corpus was established. The corpus mainly covers risk information related to market competition, technological research and development, supply chains, raw material price fluctuations, international trade, operational management, and macroeconomic environments, providing a data foundation for subsequent topic identification. Detailed statistics are presented in Table 1.

Table 1. Distribution of Risk Text Samples from 2021 to 2024

Year	**Number of Risk Texts**	**Proportion of Total Risk-Related Texts (%)**
2021	1655	24.80%
2022	1682	25.20%
2023	1663	25.00%
2024	1664	25.00%
Total	6664	100%

LDA Topic Identification Results

Determination of the Optimal Number of Topics

The LDA model requires the number of topics to be predefined, and different topic numbers directly influence the granularity of topic partitioning and the interpretability of model outputs. To determine an appropriate number of topics for risk-related texts in the electronic information manufacturing industry, this study trained LDA models with different values of \(K\) based on the parameter settings described above. Topic coherence and perplexity were calculated separately to evaluate model performance under different topic numbers.

The variations in topic coherence and perplexity corresponding to different numbers of topics are illustrated in Figure 1. As the number of topics increased, the model gradually achieved finer-grained partitioning of risk-related texts. However, when the number of topics was too small, different types of risk information tended to be concentrated within the same topic. Conversely, when the number of topics was excessively large, semantic redundancy among topics or over-fragmentation of the same risk category could occur.

By comprehensively considering topic coherence, perplexity, and topic interpretability, K=7 exhibited relatively satisfactory performance within the candidate topic range established in this study. Therefore, seven topics were selected as the final topic number for the LDA model.

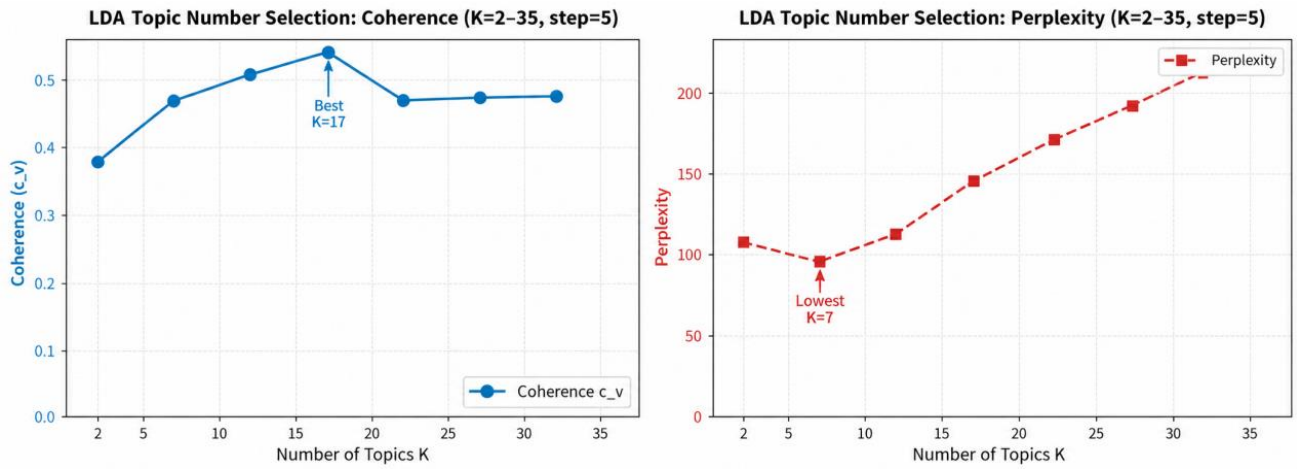


Figure 1. Variations in LDA Model Evaluation Metrics under Different Numbers of Topics

When the number of topics was set to 7, the model achieved a relatively high level of topic coherence while maintaining a relatively low perplexity value. Meanwhile, the identified topics demonstrated good semantic differentiation.

LDA Topic Identification Results

After determining the optimal number of topics as 7, this study further analyzed the topic–word probability distributions generated by the LDA model. Based on the high-probability keywords within each topic and the risk-related texts with high topic assignment probabilities under corresponding topics, the latent topics were summarized and labeled accordingly.

The seven risk topics identified by the LDA model and their representative keywords are presented in Table 2. According to the topic probability distributions of individual risk texts, the topic with the highest assignment probability was defined as the dominant topic of each text. Based on this classification, the number of texts associated with each topic and their proportions among all risk-related texts were calculated.

Table 2. Risk Topics Identified by the LDA Model

Topic ID	Risk Topic	Representative Keywords	Topic Proportion (%)
Topic 1	Macroeconomic Risk	exchange rate, economy, global	17.68%
Topic 2	Investment Risk	project, investment, environmental protection	7.64%
Topic 3	Market Competition Risk	product, industry, technology	37.66%
Topic 4	Supply Chain Risk	customer, supplier, procurement	6.48%
Topic 5	Core Talent Risk	talent, R&D, core technology	7.37%
Topic 6	Accounts Receivable Risk	accounts receivable, accounts, inventory	8.75%
Topic 7	Operational Management Risk	business, market, management	14.42%

From the perspective of topic distribution, Market Competition Risk accounted for the largest proportion, reaching 37.66%. Its representative keywords included “product,” “industry,” “technology,” “market,” and “R&D,” indicating that risks associated with market competition and technological research and development were particularly prominent in the risk texts of the sampled companies.

The next most prevalent topics were Macroeconomic Risk and Operational Management Risk, accounting for 17.68% and 14.42%, respectively. These topics mainly involved macroeconomic environments, international trade, enterprise business development, and operational management issues.

The proportions of the remaining topics were relatively lower. Specifically, Accounts Receivable Risk accounted for 8.75%, followed by Investment Risk (7.64%), Core Talent Risk (7.37%), and Supply Chain Risk (6.48%).

Overall, the risk topics identified by LDA covered the major risk categories faced by electronic information manufacturing enterprises, including market competition, external environment, operational management, finance, investment, human resources, and supply chains. Among these categories, Market Competition Risk represented the dominant risk type with the highest proportion.

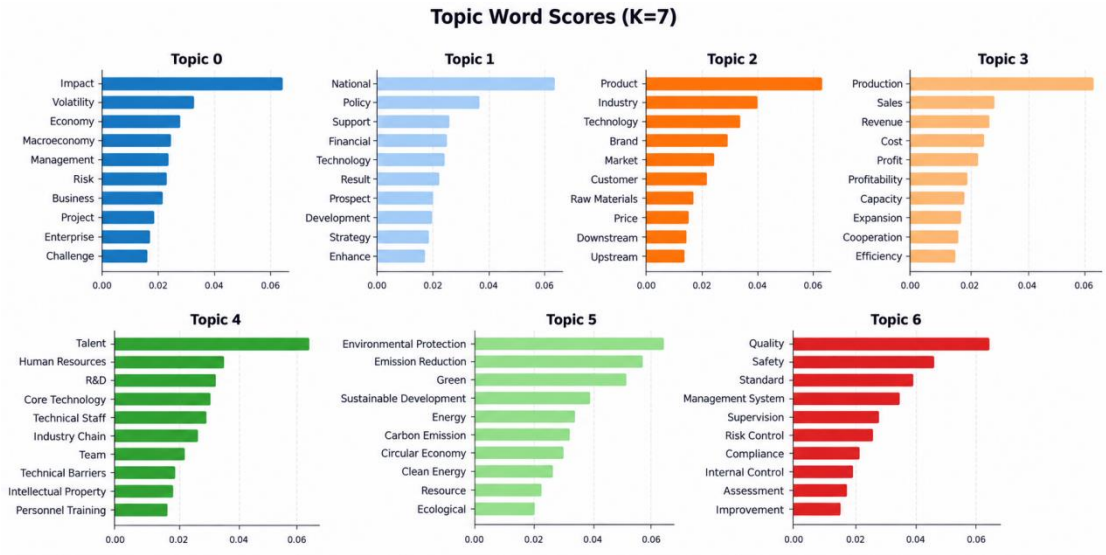


Figure 2. Distribution of High-Weight Keywords across Risk Topics Identified by LDA

Figure 2 further demonstrates that the high-weight keywords of different topics exhibit certain semantic aggregation characteristics. However, some topics still contain multiple related risk concepts, indicating that the topics identified by LDA possess a relatively strong degree of comprehensiveness.

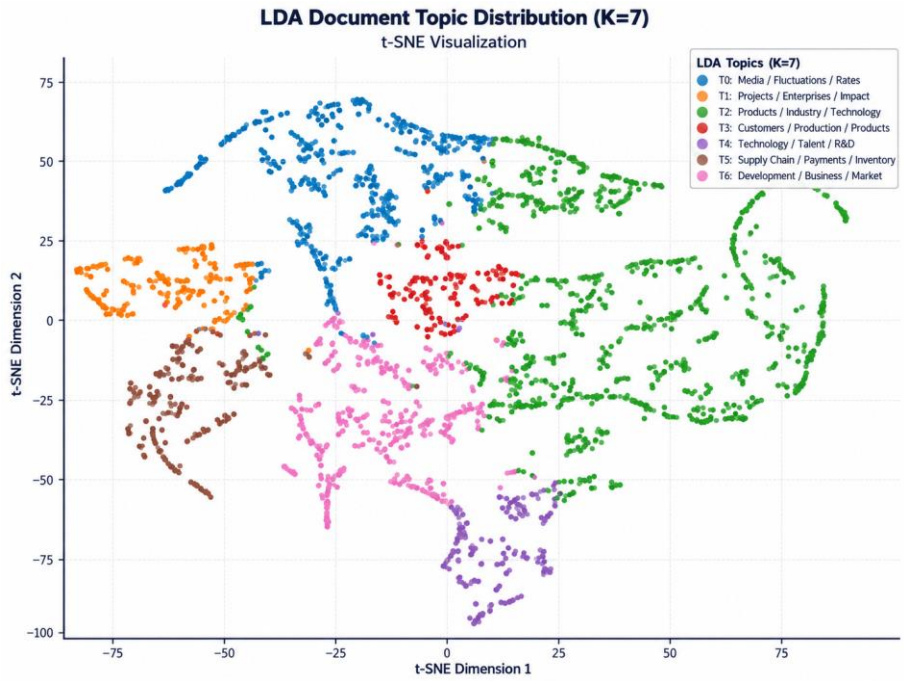


Figure 3. Visualization of Topic Distribution for LDA Risk Texts

To further investigate the distribution characteristics of LDA identification results at the document level, this study employed t-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of topic distributions of risk-related texts for visualization, with the results presented in Figure 3. Different colors represent the dominant risk topics assigned to individual documents.As shown in Figure 3, some risk topics formed relatively concentrated distribution regions in the two-dimensional space, whereas certain overlaps remained among different topics. This observation is consistent with the characteristics of LDA-based topic identification, which relies primarily on word distributions to infer latent topics. For risk-related texts sharing common vocabulary such as “product,” “technology,” and “market,” the boundaries between topics may become relatively ambiguous.Overall, these results indicate that LDA is capable of establishing a basic risk category structure; however, some risk-related texts still exhibit strong semantic associations with each other.

BER Topic Topic Identification Results

Topic Clustering and Determination of the Number of Topics

Following the model configuration described in Section 2.4.2, BERTopic was applied to perform topic modeling on the risk-related texts. The model first generated initial topics and identified texts that could not be stably assigned to other clusters as outlier documents (Topic -1). Subsequently, semantically similar initial topics were merged through the automatic topic reduction mechanism.Finally, BERTopic identified 28 valid risk topics, while 1,015 risk-related texts were classified as outlier documents, accounting for 15.23% of all risk-related texts.

Table 3. Risk Topics Identified by the BERTopic Model

Topic ID	Risk Topic	Representative Keywords	Topic Proportion (%)
Topic 1	Technological R&D Risk	technology, R&D, product	7.80%
Topic 2	Core Talent Loss Risk	technical personnel, talent, core technology	7.40%
Topic 3	Exchange Rate Fluctuation Risk	exchange rate, US dollar, foreign currency	7.11%
Topic 4	Raw Material Price Fluctuation Risk	raw materials, price, cost	6.53%
Topic 5	Accounts Receivable Risk	accounts receivable, accounts, bad debts	5.60%
Topic 6	Overseas Operation and Trade Risk	overseas, trade, foreign markets	5.42%
Topic 7	Operational Management Risk	management, scale, management capability	4.13%
Topic 8	Customer Cooperation Risk	customer, automotive, connector	3.81%
Topic 9	Consumer Electronics Market Risk	product, smart devices, customer	3.20%
Topic 10	Market Competition Risk	competition, market competition, industry	3.12%
Topic 11	Funded Project Risk	project, financing, investment project	3.11%

From the perspective of topic distribution, the risk categories identified by BERTopic exhibited a relatively dispersed pattern. Technological R&D Risk had the largest proportion, accounting for 7.80%, followed by Core Talent Loss Risk, Exchange Rate Fluctuation Risk, and Raw Material Price Fluctuation Risk, which accounted for 7.40%, 7.11%, and 6.53%, respectively. These findings indicate that technological research and development, human resources, and external economic environments represent important aspects of risk disclosure among electronic information manufacturing enterprises. Accounts Receivable Risk and Overseas Operation and Trade Risk accounted for 5.60% and 5.42%, respectively, reflecting that enterprises face both financial recovery pressures and uncertainties associated with international operations.

Furthermore, BERTopic identified several risk topics with relatively smaller proportions but clear semantic meanings, including gross profit margin decline, inventory impairment, goodwill impairment, product quality, intellectual property protection, environmental protection, and R&D failure risks. Compared with a distribution pattern dominated by a small number of topics, the proportions of BERTopic-generated topics were generally more dispersed, indicating the diversified nature of risk sources in the electronic information manufacturing industry.

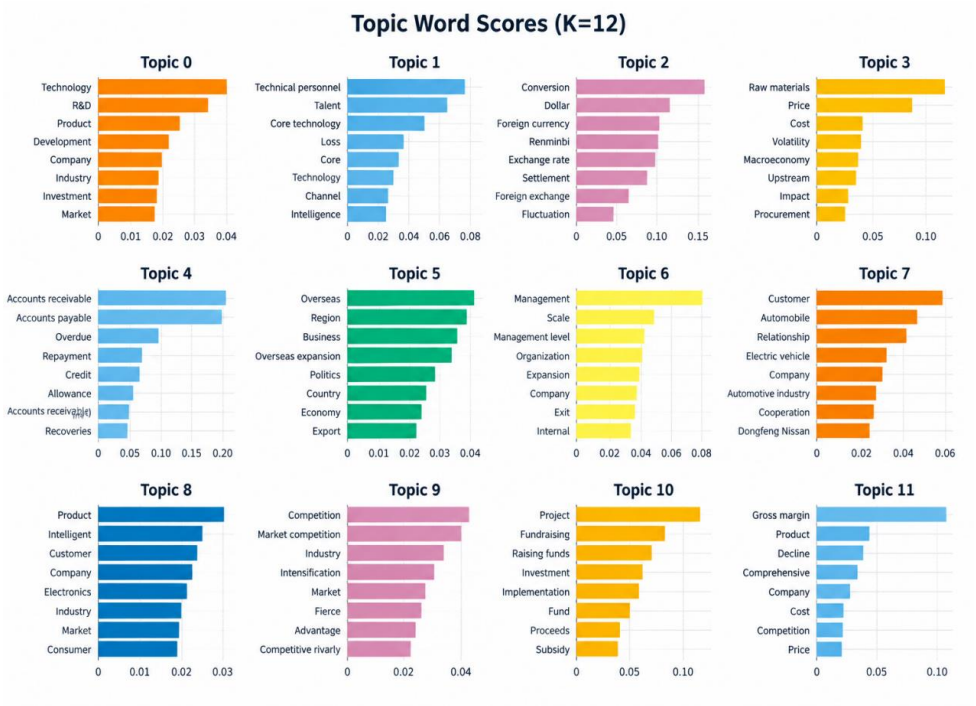


Figure 4. Distribution of Keyword Weights in Major Risk Topics Identified by BERTopic

Figure 4 shows that the high-weight keywords of BERTopic topics exhibit strong semantic concentration and provide relatively clear indications of specific risk issues, offering a more explicit semantic basis for topic labeling.

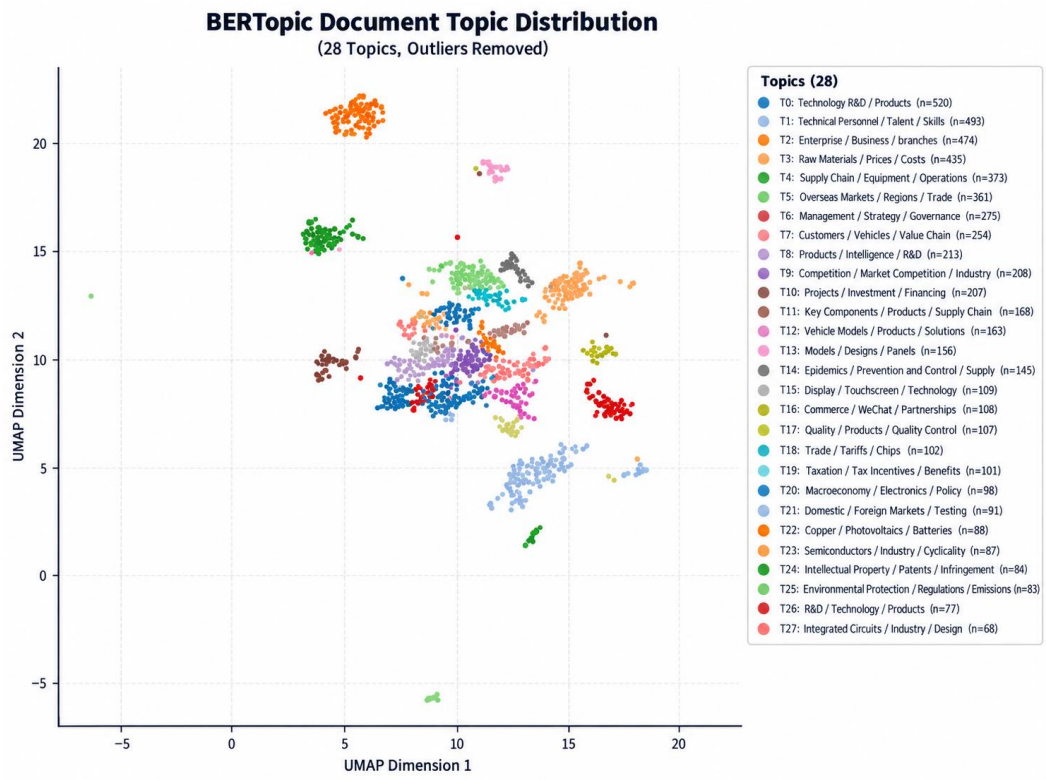


Figure 5. BERTopic-Based Topic Clustering Distribution of Risk Texts

To further demonstrate the topic clustering results of BERTopic, this study performed dimensionality reduction visualization on the semantic embeddings of risk-related texts, with the results presented in Figure 5. Different colors represent different risk topics.As shown in Figure 5, different topics formed multiple clustering regions with certain degrees of separation in the semantic space. Some topics exhibited relatively high levels of aggregation, while certain semantically similar risk topics were located close to each other in the embedding space. This indicates that different risk categories in the electronic information manufacturing industry are not completely independent but maintain certain semantic associations, while still forming relatively specific semantic categories.Combined with subsequent evaluations of topic coherence and topic diversity, the effectiveness of BERTopic in identifying different risk semantics can be further assessed.

Comparison between LDA and BER Topic Models

Comparison of Topic Identification Results

LDA identified seven risk topics, which mainly represented comprehensive risk categories, including macroeconomic risk, investment risk, market competition risk, supply chain risk, core talent risk, accounts receivable risk, and operational management risk. In contrast, BERTopic identified 28 valid topics, further generating more specific risk-related subtopics based on relatively comprehensive risk categories.

From the perspective of topic structure, the topics identified by LDA were relatively concentrated, with individual topics covering broader ranges of risk-related information. In comparison, BERTopic produced a more dispersed topic structure and was able to distinguish texts with similar lexical characteristics but different risk contexts into separate topics. Therefore, the primary difference between the two models in risk identification results is that LDA tends to generate a relatively aggregated risk category structure, whereas BERTopic produces a more fine-grained risk topic structure.

Comparison of Model Evaluation Metrics

To further compare the topic identification performance of the two models, this study employed C_v topic coherence (Coherence) and topic diversity (Topic Diversity) as evaluation metrics. The results are presented in Table 4.

Table 4. Comparison of Topic Identification Performance between LDA and BERTopic

Topic Model	Number of Topics	Topic Coherence	Topic Diversity
LDA	7	0.5049	0.7143
BERTopic	28	0.6914	0.8071

The topic coherence score of BERTopic was 0.6914, which was higher than that of LDA (0.5049), indicating that the high-weight keywords within BERTopic-generated topics exhibited stronger semantic associations. The topic diversity of BERTopic was 0.8071, also exceeding that of LDA (0.7143), suggesting that the degree of keyword overlap among different topics was relatively lower.Overall, based on the evaluation metrics adopted in this study, BERTopic demonstrated higher topic coherence and topic diversity than LDA, indicating better performance in terms of topic quality.

Comparison of Model Interpretability

From the perspective of topic interpretability, both models were able to generate risk topics with certain semantic orientations; however, differences existed in the way these topics were represented. The topics generated by LDA covered relatively broader ranges, with individual topics often incorporating multiple

interconnected risk concepts. Therefore, the topic labels mainly reflected an overall summary of a category of risk information, demonstrating a relatively high level of comprehensiveness.

In contrast, the keywords within BERTopic-generated topics exhibited more concentrated semantic characteristics, resulting in more specific topic boundaries. The topic labels could more directly correspond to specific risk issues. By examining the representative keywords and associated risk texts, it can be observed that BERTopic topics possessed clearer semantic orientations, facilitating the differentiation and labeling of distinct risk issues.

Therefore, in terms of interpretative level, LDA emphasizes the categorical summarization of risk information, whereas BERTopic focuses more on the presentation of specific risk issues. The two models respectively demonstrate advantages in overall risk abstraction and fine-grained risk interpretation.

Comparison of Model Applicability

Based on the aforementioned topic identification results, evaluation metrics, and interpretability characteristics, the two models are applicable to different levels of risk analysis tasks. LDA generates a smaller number of topics with a relatively concentrated structure, allowing the overall composition of industry risks to be presented in a concise manner. Therefore, it is more suitable for industry-level risk overviews, overall risk classification, and risk structure analysis.

In contrast, BERTopic produces more fine-grained topics with higher topic coherence and topic diversity. It is capable of further identifying specific risk issues from broader risk categories, making it more suitable for fine-grained risk identification, key risk analysis, and specific risk information extraction.

Therefore, model selection should be determined according to the specific analytical objectives. When the research goal is to establish a concise overall risk framework, LDA can be applied for macro-level risk classification. When the objective is to identify specific risk issues and distinguish semantically similar risks, BERTopic is more appropriate for in-depth analysis. The two models should not be regarded as simple substitutes for each other; rather, they provide complementary perspectives on risk information at different levels of granularity.

Human Validation of Topic Identification Results

To further validate the accuracy and interpretability of the identified risk topics, a manual evaluation was conducted based on representative risk texts. All seven LDA topics and the ten most representative BERTopic topics with higher topic proportions were selected for evaluation. For each selected topic, ten representative risk-related texts were randomly sampled. Two researchers with backgrounds in risk management and industrial management independently evaluated the sampled texts according to topic relevance and topic interpretability using a five-point Likert scale.

Table 5. Human Evaluation Results of Topic Relevance and Interpretability

Model	Topic relevance	Interpretability
LDA	4.2	4.1
BERTopic	4.5	4.2

The human evaluation results are presented in Table 5. Both LDA and BERTopic achieved relatively high scores in topic relevance and interpretability, indicating that the identified topics were generally consistent with the representative risk texts and understandable to human evaluators. BERTopic achieved a higher score in topic relevance (4.5) than LDA (4.2), suggesting that its topic assignments were more closely aligned with the semantic content of risk texts. In terms of topic interpretability, BERTopic also obtained a slightly higher score (4.2) than LDA (4.1), indicating that it provided slightly clearer topic representations. However, LDA maintained satisfactory performance, reflecting its capability to summarize broader risk categories.

DISCUSSION

Multidimensional Risk Characteristics of the Electronic Information Manufacturing Industry

The results indicate that risk disclosures in the electronic information manufacturing industry exhibit clear multidimensional characteristics, covering technological innovation, market competition, core talent, supply chains, finance, and external operating environments. In particular, topics related to technological R&D, core talent loss, exchange rate fluctuations, and raw material price fluctuations received relatively high levels of textual attention, suggesting that enterprise risks originate not only from internal technological innovation and operational management but are also influenced by changes in market conditions, resource prices, and the international business environment.

This characteristic is closely associated with the industrial attributes of the electronic information manufacturing industry. The industry is characterized by rapid technological updates, high dependence on R&D investment and specialized talent, and relatively complex industrial chains, with some enterprises exhibiting a high degree of globalization. Therefore, the risk structure of this industry reflects the interaction among technological innovation, resource dependence, supply chain complexity, and external environmental uncertainty.

Differences between the Two Models Arise from Text Representation and Topic Generation Mechanisms

LDA and BERTopic exhibit different levels of granularity in risk identification, and the underlying reasons are closely related to their distinct information representation approaches. In risk-related texts of the electronic information manufacturing industry, terms such as “technology,” “product,” “market,” and “customer” demonstrate high cross-topic reuse characteristics. The same term may correspond to different risk issues under different contextual conditions. LDA primarily identifies latent topics based on word distributions and co-occurrence relationships; therefore, it tends to aggregate risk texts sharing similar vocabulary into relatively broad and comprehensive topics.

In contrast, BERTopic utilizes pretrained language models to obtain contextual semantic representations of texts and performs clustering in the semantic space. As a result, it is capable of further distinguishing texts with similar vocabulary but different underlying risk meanings. For example, texts containing terms such as “technology” and “R&D” may correspond to different risk issues, including technological R&D investment risk, R&D failure risk, core technical personnel loss risk, and intellectual property risk. This indicates that for industry-specific texts characterized by abundant professional terminology and significant overlap in risk semantics, contextual semantic information can support the further refinement of risk categories.

Fine-Grained Risk Identification and Managerial Relevance

The value of fine-grained risk identification lies in transforming broad risk categories into more specific risk issues with clearer managerial implications. By providing more detailed representations of risk information, fine-grained identification helps enterprises better understand the underlying characteristics and potential sources of different risks.

This transition from broad categories to specific risk issues improves the connection between textual analysis and practical risk management. While broad risk categories help managers understand overall risk structures, specific risk issues provide more direct support for risk monitoring, early warning, and targeted response activities. Therefore, fine-grained risk identification can provide a more detailed information foundation for subsequent risk analysis and managerial decision-making.

Managerial Implications

From an application perspective, enterprises can integrate topic modeling into risk monitoring systems. A practical approach is to combine broad risk classification with detailed issue detection, where overall risk structures provide strategic understanding and specific risk topics support targeted monitoring and response.This hierarchical approach can improve the organization of risk information and provide decision support for managers.

CONCLUSION

This study investigated risk-related texts extracted from the annual reports of listed electronic information manufacturing companies in the Shanghai and Shenzhen A-share markets during 2021–2024. LDA and BERTopic were employed to identify and compare risk topics. The results indicate that both models are capable of identifying meaningful industry risk topics from unstructured annual report texts; however, they reveal different levels of risk information granularity. LDA generates a relatively concentrated topic structure and is more suitable for summarizing overall industry risk categories. In contrast, BERTopic leverages contextual semantic information to further distinguish risk texts with similar lexical characteristics and demonstrates stronger capabilities in identifying specific risk issues. The evaluation results based on topic coherence and topic diversity also show that BERTopic achieved better intra-topic semantic consistency and inter-topic differentiation under the experimental settings of this study.

The main contribution of this study lies in demonstrating that different topic modeling approaches can present risk information in the electronic information manufacturing industry at different levels of granularity. For the analysis of overall industry risk structures, relatively aggregated topic classifications can help establish a clear risk framework. For specific risk discovery, contextual semantic information is required to distinguish risk issues with different managerial implications. Therefore, in practical applications, appropriate topic granularity should be selected according to the objectives of risk identification, and topic models can serve as auxiliary tools for enterprise risk information management and risk discovery.

RESEARCH LIMITATIONS AND FUTURE DIRECTIONS

This study has several limitations. First, the data used in this study were derived from annual reports of listed companies, and the identified risk topics mainly reflect the risk information publicly disclosed by enterprises. Therefore, they may not fully capture all risks arising during actual business operations. Second, the results of topic identification may be affected by factors such as text preprocessing procedures, model parameter settings, and manual topic labeling. Thus, the stability and interpretive validity of the identified topics require further verification. Finally, this study primarily focused on comparing the risk identification performance of different topic modeling approaches and did not examine the relationships between risk topics and enterprises' financial conditions, operational performance, or actual risk events.

Future research can be extended in three directions. First, additional text sources, such as corporate announcements and regulatory inquiry letters, can be incorporated to improve the coverage of risk information. Second, annual and subsector-level information can be integrated to investigate the dynamic evolution of risk topics and industry heterogeneity. Third, risk topics can be combined with financial indicators, operational performance, or actual risk events to further evaluate the external validity of text-based risk identification results.

REFERENCES

1. SODHI M S, LEE S. An analysis of sources of risk in the consumer electronics industry[J]. Journal of the Operational Research Society, 2007, 58(11): 1430-1439.
2. LIAO Z W, TANTAI B Y, ABDUL-HAMID A Q, et al. Exploring resilience in the downstream supply chain of the semiconductor industry: The mediating roles of risk mitigation, process simplification, and flexibility[J]. International Journal of Production Economics, 2025, 281: 109530.
3. MOKTADIR M A, REN J. Global semiconductor supply chain resilience challenges and mitigation strategies: A novel integrated decomposed fuzzy set Delphi, WINGS and QFD model[J]. International Journal of Production Economics, 2024, 273: 109280.
4. CAMPBELL J L, CHEN H, DHALIWAL D S, et al. The information content of mandatory risk factor disclosures in corporate filings[J]. Review of Accounting Studies, 2014, 19(1): 396-455.
5. JIN S C, LEE H W, XIE Y A. Risk disclosures in annual reports and firms' default risk[J]. Journal of Corporate Accounting & Finance, 2024, 35(3): 192-212.
6. DYER T, LANG M, STICE-LAWRENCE L. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation[J]. Journal of Accounting and Economics, 2017, 64(2-3): 221-245.
7. ZHU X Q, WU H D, CHANG Y P, et al. Accounting fraud detection through textual risk disclosures in annual reports: From the perspective of SEC guidelines[J]. Accounting & Finance, 2025, 65(2): 1837-1862.
8. KOBEISY A N, ROBERTS R W. Corporate governance and risk disclosure: Evidence from textual analysis of annual reports. Journal of Accounting and Public Policy, 2020, 39(6): 106753.
9. BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
10. BIANCHI F, TERRAGNI S, HOVY D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, 2021: 759-766.
11. DIENG A B, RUIZ F J R, BLEI D M. Topic modeling in embedding spaces[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 439-453.
12. GROOTENDORST M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure[EB/OL]. (2022-03-11)[2026-10-06].
13. REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019: 3982-3992.
14. ANGELI G, MANANDHAR S, NEPAL S.Topic modeling using contextual embeddings: A comparative analysis of traditional and neural approaches. Information Processing & Management, 2022, 59(6): 103091.
15. GAD M, PARK G, RAWSTHORNE S, et al. When methods matter: How implementation choices shape topic discovery in financial text[J]. Accounting and Business Research, 2026.
16. ABDELRAZEK A, EID Y, GAWISH E, et al. Topic modeling algorithms and applications: A survey[J]. Information Systems, 2023, 112: 102131.
17. MISHRA M. Comparative analysis of BERTopic versus LDA: Topic modelling in marketing research[J]. Australasian Marketing Journal, 2026, 34(2): 162-173.