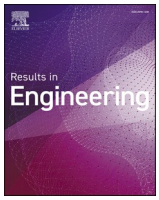
Results in Engineering 26 (2025) 104920

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Research paper

A study on the application of the latent dirichlet allocation model in production optimization

Youbin Chen *, Da Huang, Lifei Wang

School of Economics and Management, Shanghai Institute of Technology, Shanghai 200235, China

ARTICLE INFO

Keywords:
Lean six sigma
Overall equipment effectiveness
Machine learning
LDA topic modeling
DMAIC

ABSTRACT

In recent years, the global manufacturing industry has continuously undergone digital transformation amidst intense competition. The traditional experience-based lean production model is inadequate to resolve issues like declining equipment efficiency and frequent failures. This study proposes a fault diagnosis and improvement strategy that integrates the LDA topic model with the Six Sigma DMAIC methodology, carrying out in-depth analysis of failure factors in key processes on the production line and implementing a 12-week lean improvement initiative in a case enterprise. The results show that after these improvements, the production line's overall equipment effectiveness (OEE) increased significantly from 71 % to 84 %, with a marked reduction in equipment failure frequency, thereby ensuring enhanced production stability and reliability. This study overcomes the limitations of traditional lean methodologies' reliance on structured data by integrating data-driven decision-making into the conventional Lean Six Sigma framework, not only providing a quantitative foundation for equipment fault diagnosis but also offering new theoretical perspectives and practical pathways for continuous improvement and intelligent transformation in manufacturing.

1. Introduction

In recent years, against the backdrop of intensified international competition, the industrial sector has increasingly emphasized applying lean production principles to corporate operations and management, aiming to gain a competitive edge by eliminating waste and enhancing efficiency and quality. Studies have shown that green lean practices not only improve resource utilization and environmental performance but also help enterprises better meet increasingly stringent sustainability requirements [1]. Meanwhile, manufacturing business models are gradually shifting from a traditional experience-driven approach to data-driven operational decisions. By deeply integrating products and services, enterprises can achieve more accurate demand forecasting and continuous improvement [2]. However, as manufacturing systems become increasingly digitalized and information-driven, relying solely on experience-based management makes it difficult to maintain stable and efficient operations in a complex and rapidly changing market. Consequently, integrating Industry 4.0 technologies with lean methodologies has become an inevitable trend [3]. This integration not only optimizes internal production and logistics processes but also employs real-time data collection and digital decision-making to effectively address potential risks such as market fluctuations and supply chain disruptions [4]. Furthermore, when constructing and operating a lean production system, enterprises must balance flexibility and resilience, ensuring they maintain a stable production rhythm and high-level service even as external conditions shift [5]. In summary, lean production is evolving from a traditionally experience-driven model into a data analytics- and technology-driven decision process, continually driving transformation and upgrades in industrial operations.

In the context of Industry 4.0, the digital transformation of enterprises' lean production systems requires data-driven decision-making tools, which are strongly supported by the wide application of machine learning methods. Huang et al. [6] applied deep learning models to real-time storage detection of oil tanks in ports, which provides a data base for industrial port management and resilience research. Huang et al. [7] applied a machine learning approach to construct a vehicle speed predictor for the energy management problem of fuel cell vehicles. Li et al. [8] Hybrid computer vision approach for pavement crack detection applying machine learning. As machine learning methods continue to evolve, traditional lean manufacturing models also need to move from empirically dependent localized improvements to systematic innovations based on global data models. Latent Dirichlet Allocation

* Corresponding author.
E-mail address: 987154385@qq.com (Y. Chen).

https://doi.org/10.1016/j.rineng.2025.104920
Received 24 March 2025; Received in revised form 11 April 2025; Accepted 11 April 2025
Available online 12 April 2025

(LDA) is an important data-driven decision-making tool that can be flexibly applied to research in areas as diverse as technology trends [9], urban climate [10], and supply chain [11], as well as to support the digital transformation of lean manufacturing systems. By establishing a three-layer Bayesian probabilistic model encompassing documents, topics, and terms, LDA uncovers latent semantic correlations among multimodal data in manufacturing systems, including equipment sensor data, quality inspection reports, and supply chain texts. This approach not only reveals hidden efficiency losses beyond the traditional "seven wastes" but can also integrate green lean concepts to quantify the coupling relationship between resource consumption and environmental performance [3].

The current mainstream lean manufacturing model relies heavily on manual experience and structured data analysis [12,13]. Compared with the traditional lean model, the machine learning-enabled lean framework is able to simultaneously parse temporal, spatial, and unstructured data in the manufacturing system, breaking through the traditional dependence on structured data for value stream analysis. Enzhi Dong et al. (2025) proposed to train a Transformer network using equipment historical data to determine the degradation categories of mechanical systems and to dynamically obtain the probability distribution of the remaining service life of the equipment using the Kernel Density Estimation approach [14];Lubing Wang et al. (2025) used data monitored by sensors to perform predictive maintenance on aircraft engines, predicting possible system failures in advance and taking proactive maintenance measures [15];Suhwan Lee et al. (2025) Predictive Maintenance of Butterfly Valves in Ship Exhaust Systems Based on Ship Butterfly Valve Number and Exhaust Time Interval Data [16]. However, in most of the current predictive maintenance research, it needs to rely on professional analysts to conduct a large number of experiments and data analysis. The practical application of these predictive maintenance methods in an organization requires a higher level of digital competence from frontline employees, which adds a huge challenge to the operational cost as well as efficiency of the organization. Compared to quantitative analysis methods that require a certain level of statistical knowledge, LDA's subject-matter word interpretation relies more on business experience than on mathematical ability. In the manufacturing industry, the equipment maintenance logs recorded by frontline employees may contain terms, abbreviations, and colloquial expressions of their daily work, and LDA clusters these words into corresponding themes. The generated topic words are directly mapped to the employees' practical operations, avoiding the barrier of understanding caused by abstract concepts. LDA can automatically identify descriptive features related to equipment failures, extract potential topics of quality defects from unstructured data, and provide data support for constructing an enterprise knowledge graph [17,18]. Meanwhile, LDA supports incremental training, which can adjust the topics in real time according to the new operation records added by employees, and transforms abstract topics into intuitive vocabulary collections through word clouds, heat maps, and other visualization tools.

In summary, data-driven decision-making in Industry 4.0 relies on the fusion of heterogeneous data from multiple sources, while existing Industry 4.0 platforms often face problems such as semantic disconnection of multimodal data and the existence of cross-process information silos in the manufacturing site in practical applications. This study applies the LDA methodology to the lean manufacturing framework to break through the reliance on empirical as well as structured data in the traditional production maintenance system, and to provide new applications and research ideas for the study of lean manufacturing and predictive maintenance in the context of Industry 4.0.

2. Literature review

2.1. Lean six sigma

Lean Production originated from the Toyota Production System and centers on systematically identifying and eliminating non-value-adding activities to enhance operational efficiency and product quality while maximizing customer value [19].Guided by this principle, companies must not only continually refine their production processes but also emphasize effective coordination among personnel, equipment, and materials [20]. Lean manufacturing is widely used in the manufacturing industry thanks to the continuous improvement mechanism it establishes. Mojan Eskandari et al. (2022) used a lean framework to assess the performance of pharmaceutical factories [21];Darian Pearce et al. (2021) applied lean management in fruit horticulture in South Africa to achieve sustainable primary production [22];and Alessia Bilancia et al. (2025) applied the lean model to the luxury fashion industry to promote the sustainability in luxury fashion [23]. On one hand, its focus on "meeting customer needs" and "adding value to processes" enables companies to flexibly adjust operations in ever-changing market conditions; on the other hand, by minimizing waste, standardizing processes, and encouraging employee participation, organizations strike a balance among shorter delivery lead times, lower costs, and higher quality. Whether in highly automated assembly lines or workshops featuring manual or small-batch production, Lean Production helps enterprises uncover opportunities for cost reduction and efficiency gains, laying a solid foundation for subsequent efforts in sustainability and digital transformation.

Six Sigma is a data-driven method for quality improvement and process optimization aimed at reducing process variation through systematic steps, thereby enhancing overall operational performance [24]. Its core philosophy employs statistical tools and management techniques to drive continuous improvement in production or service processes, striving for near-zero defect levels [25]. In practice, Six Sigma's DMAIC (Define, Measure, Analyze, Improve, Control) framework is widely used. Lokpriya M. Gaikwad et al. (2022) Implementation of Six Sigma model to enhance the competitive advantage of firms [26];Ivana Tita Bella Widiwati et al. (2024) Implementation of Lean Six Sigma methodology to minimize wastage in food manufacturing industry [27]; Michael E. Natarus et al. (2025) Optimizing aseptic processing departments, staffing increases, and capital investments using Lean Six Sigma methodology [28]. By utilizing a scientific quantitative methodology and a series of structured improvement steps, Six Sigma not only effectively reduces process waste, but also dramatically improves quality consistency while fostering a culture of continuous improvement [29].

Lean Six Sigma is a continuous improvement methodology developed to synthesize the benefits of Lean Manufacturing and Six Sigma. Lean Manufacturing advocates the elimination of unnecessary waste, while Six Sigma utilizes statistical methods to keep defect rates at very low levels. Lean Six Sigma combines these two approaches by not only streamlining processes through Lean principles, but also by applying Six Sigma's quantitative methods to identify the root causes of problems, resulting in both efficiency and quality improvements.

Most current Lean Six Sigma studies utilize tools such as value stream mapping, Kanban, 5S, single-minute mold change, statistical process control and hypothesis testing. Traditional Lean Six Sigma research tools rely on manual experience to identify waste and lack a data-driven dynamic assessment model. Lean manufacturing is mostly based on structured metrics and fails to effectively integrate semantic clues from textual data such as equipment logs. Meanwhile, traditional Six Sigma relies on normal distribution assumptions, which makes it difficult to handle high-dimensional data and has limitations in adapting to complex systems. Therefore, the Lean Six Sigma model needs to explore how to deeply integrate with complex dynamic production environments and unstructured data processing methods to achieve more efficient and accurate process optimization, and to promote the upgrade of Lean Six Sigma from staged improvement to continuous adaptive optimization.

2.2. Latent dirichlet allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic topic model widely employed to uncover latent topics from large-scale text data. Its core concept views each document as a mixture of multiple hidden themes, with each theme characterized by a probability distribution of words [30]. Within this model, every word in the text is assumed to be generated by a particular underlying topic governed by Dirichlet priors, thereby enabling automatic discovery of topic structures under unsupervised conditions [31].

LDA has been applied in numerous domains, including energy and environmental research on user energy consumption behavior [31], literature-based topic mapping in supply chain and operations management [32], sentiment and preference analysis of online user reviews [33], and evaluations of urban mobility and traffic policies [34]. By representing text as a "bag of words" and using probabilistic inference to distribute word frequencies among different topics, researchers can intuitively capture the main issues underlying large amounts of textual data, thus providing robust evidence for policy-making and operational decisions [30].

In the LDA model, the generation of a document is treated as a probabilistic procedure (Eq. (1)). The objective is to use the topic and word distributions to describe the probability of each word occurring in a given document.

$$P(w) = \prod_{n=1}^{N} \varphi_{z=1}^{K} P(w_n|z;\varphi)P(z|d;\theta) \tag{1}$$

Where w is the set of all words in the document, w_n represents the n-th word in the document, z is a latent (unobserved) variable denoting the topic to which w_n belongs, ϕ is the word distribution for each topic, and θ is the topic distribution for the document.

Typically, LDA modeling involves several key steps. First, the text data are collected and cleaned, followed by tokenization, removal of stop words, and other preprocessing steps to reduce noise. Next, the number of topics (k) and the hyperparameters α and β are set, and iterative sampling or optimization is performed on the corpus to obtain both the document–topic and topic–word distributions [30]. Finally, the keywords associated with each topic are qualitatively analyzed and labeled, and the model's results are interpreted and applied in conjunction with relevant domain knowledge and research context [35].

LDA demonstrates strong scalability and interpretability in the fields of natural language processing and text mining, which has gradually established it as a key tool for understanding large-scale text data. Current research on LDA in manufacturing industry mostly focuses on macro studies such as research theme analysis [36], bibliometrics [37], and exploration of sustainable manufacturing paths [38], and lacks the exploration of the application to the actual production of enterprises. Therefore, this study will combine the synergistic application of LDA methods and Lean Six Sigma in enterprise production to provide new ideas for production optimization in the context of Industry 4.0.

3. Methodology

This study employs a case study approach to investigate the fundamental causes impacting the overall equipment effectiveness (OEE) of the company's packaging production line and propose lean improvement initiatives to boost production efficiency. The case study method offers flexibility for conducting both quantitative and qualitative analyses of specific events within a company [39], and it is widely used across various fields such as industrial production, healthcare, and economics. The information and data required for this case were gathered through on-site observation at the enterprise, as well as through expert opinions from its production, engineering, and continuous improvement departments. As shown in Fig. 1, the research framework follows the five DMAIC phases of Six Sigma: Define, Measure, Analyze, Improve, and Control to ensure that the research process is systematic and scientific.

In the definition phase, the SIPOC model is used to clarify the production process boundaries and core bottlenecks, and equipment reliability is targeted as the key improvement goal in conjunction with internal and external customer requirements. In the measurement stage, the efficiency baseline is established by quantifying the comprehensive efficiency indexes of the equipment, and the time loss is categorized to identify the specific shortcomings in the three major dimensions of equipment availability, performance efficiency and quality pass rate. In the analysis stage, the LDA theme model is innovatively applied to extract the core fault themes from the equipment fault records and verify the root causes of the faults with on-site observation. In the improvement phase, 5S on-site management, error-proof design, parameter optimization and other measures are applied, and standardized maintenance processes are developed simultaneously to solidify the measures. In the control phase, real-time monitoring of equipment status, construction of fault theme intensity heat map dynamic tracking of residual risk, and through the standardization of cleaning processes, quality feedback mechanism to achieve continuous improvement. Data-driven decision-making and cross-departmental collaboration are carried out throughout the entire process, exploring the practical value of integrating traditional Six Sigma tools with text mining technology.

4. The case study

This company ranks among the world's largest fast-moving consumer goods manufacturers. In the packaging workshop of this case-study factory, each production line follows a similar workflow. For the purposes of this study, one production line that is highly representative

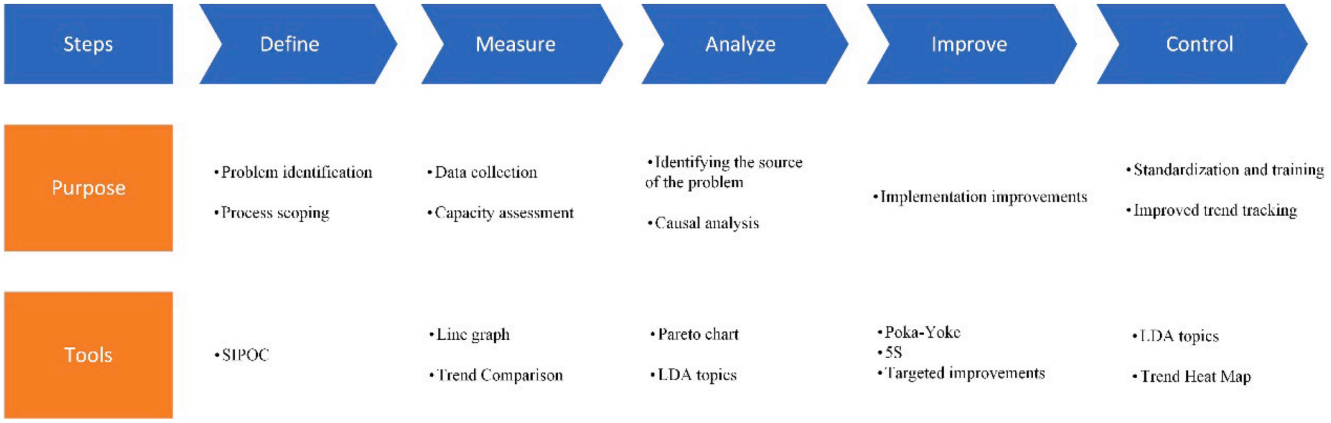


Fig. 1. Research framework.

of industry standards was selected for case analysis. The primary product processed on this line is a particular type of edible sauce, which is one of the company's top-selling items. However, the OEE of this production line has remained below the target for an extended period, resulting in financial losses and negatively impacting both customer satisfaction and production output. Consequently, the company decided to initiate a 12-week lean improvement project on this production line to ensure a marked and stable increase in OEE, reaching a high-level target. Fig. 2 presents the workflow of the selected packaging line, which includes seven major steps: bottle unloading, filling, cap placement, labeling, lane division, carton forming, boxing, sealing, and palletizing.

4.1. Define phase

In the Define phase, a SIPOC model (Supplier, Input, Process, Output, Customer) was constructed to clarify and standardize each stage of the production process, thereby identifying specific bottlenecks causing efficiency losses. The "Supplier" component refers to the pre-processing department for raw materials, which delivers sterilized and seasoned materials to the subsequent packaging stage. The "Input" comprises various materials required for the packaging line, such as raw materials, bottles, caps, labels, and cartons. The "Process" represents the main technological steps in the packaging line, while the "Output" pertains to the finished, packaged products. Finally, the "Customer" component includes the wholesalers who sell the finished product. The detailed SIPOC flow is illustrated in Fig. 3.

From the SIPOC model and on-site observations, it was determined that the central bottleneck in the current workflow is low production efficiency, which directly impedes the company's ability to meet customer demands. Moreover, by applying the Voice of the Customer (VOC) method, the research team captured feedback from internal customers (the pre-processing department and line operators) as well as external customers (wholesalers), thus more intuitively pinpointing the pain points arising from process bottlenecks. Wholesalers indicated that production capacity is approaching its limit, compromising responsiveness to rising demand. Meanwhile, the pre-processing department noted that frequent equipment downtime for maintenance significantly disrupts the execution of production plans. In addition, the line operators reported that the production equipment operates inconsistently and experiences frequent malfunctions, posing a serious drag on overall production efficiency. Consequently, reliability and maintenance management of the equipment emerged as the core issues constraining production efficiency, necessitating targeted measures for improvement.

4.2. Measure phase

During the Measure phase, data were gathered over a 12-week period, primarily focusing on the production line's OEE indicator and daily records of time losses. In the initial three weeks, the line's OEE was notably suboptimal; starting from the third week through the twelfth, a Lean Six Sigma improvement plan was implemented to raise the line's OEE to a stable, high-level target.

OEE was first defined by Nakajima in 1988, classifying losses according to three main dimensions—Availability (A), Performance (P), and Quality (Q)—and six forms of loss, including breakdowns, setup and adjustment losses, idling and minor stoppages, speed loss, defects and rework, and startup losses [40]. In this case study, the company refined the OEE calculation metrics by quantifying time-based waste to evaluate the comprehensive efficiency of production line equipment (as shown in Table 1) [40,41].

By collecting and measuring the company's OEE data from Weeks 1 to 3 (Fig. 4), we found the OEE to be considerably below the world-class benchmark of 85 % [39], and it showed a steadily declining trend. This posed a serious threat to the company's production efficiency, highlighting the urgent need for corrective action.

4.3. Analyze phase

In the Analysis phase, we first categorized the causes of equipment failures and downtime according to the packaging line's workflow and major machines, dividing them into the following groups: filling machine, labeling machine, boxing machine, cap placement machine, bottle unloading machine, sealing machine, forming machine, lane divider, palletizing, and conveyor chain. Using a Pareto chart, we then identified the equipment that accounted for only 20 % of downtime incidents yet produced as much as 80 % of the negative impact on overall operational efficiency (Fig. 5). The Pareto analysis revealed that the filling machine, labeling machine, and boxing machine were responsible for causing up to 80 % of the negative impact on overall efficiency.

Next, we applied the LDA model to further analyze the factors that lead to device failures and determine the optimal number of topics based on the ease of mixing and consistency measurements. In terms of stop words selection, we selected the list of commonly used stop words

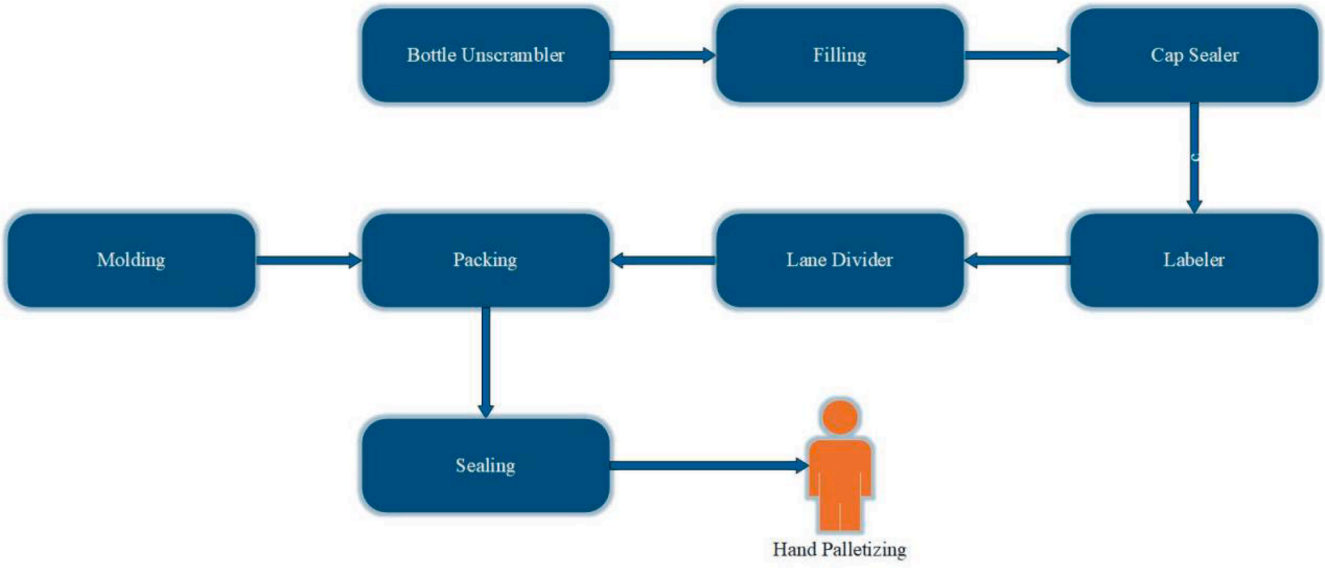


Fig. 2. The working process of the packaging line in.

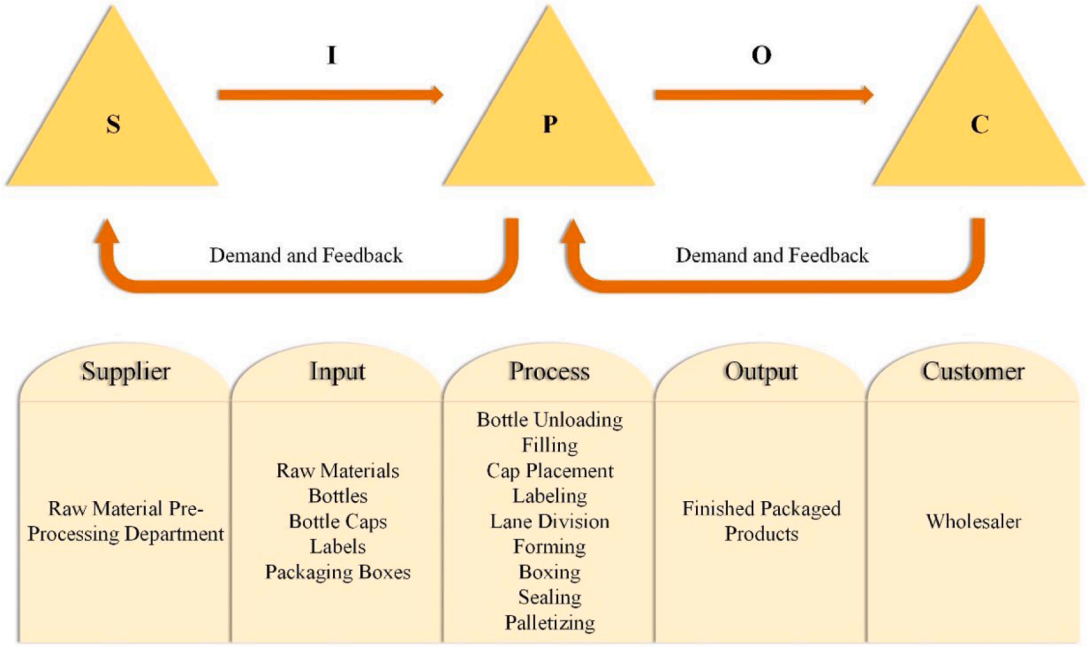


Fig. 3. SIOPC definition process.

Table 1
A comparison of these metrics is provided.

Nakajima		Case Company	
Availability (A)	$\frac{\text{Loading time} - \text{downtime}}{\text{Loading time}}$	Scheduled Time	Total Available Time – *Unscheduled Time*
Performance (P)	$\frac{\text{Ideal cycle time} * \text{output}}{\text{Operating time}}$	Production Time	Scheduled Time – *Planned losses*
Quality (Q)	$\frac{\text{Input} - \text{volume of quality defects}}{\text{Loading time}}$	Full Rate Time	Production Time – *Unplanned Losses*
OEE	A*P*Q	OEE	$\frac{\text{Full Rate Time}}{\text{Scheduled Time}}$

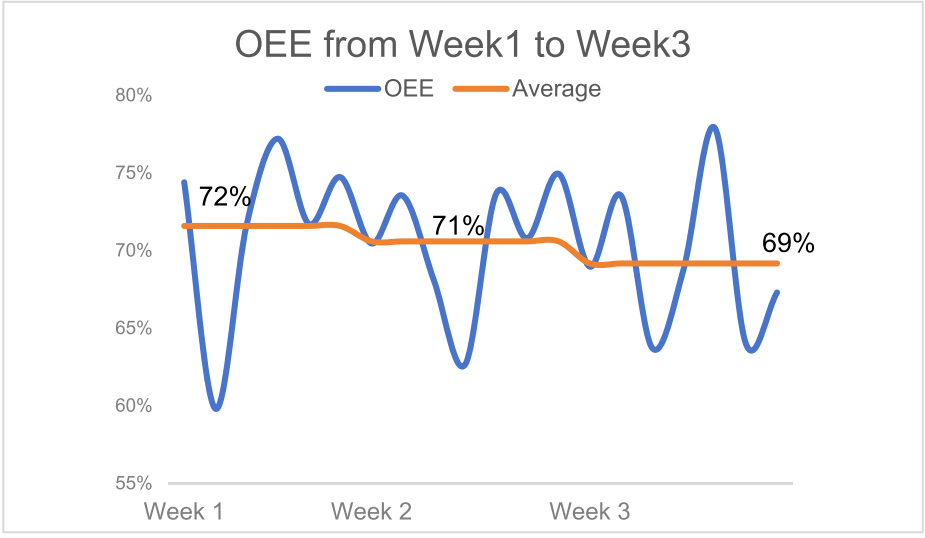


Fig. 4. The OEE trend from the 1st to the 3rd week.

published by the Natural Language Processing Laboratory of Harbin Institute of Technology according to the location of the plant, which includes such words as “the”, “is”, “in ”, “are”, “in”, “you” and so on. In terms of topic count selection, the number of LDA topics in current mainstream research needs to be determined based on one or more different metrics, such as consistency, perplexity, or differences between topics based on LDA visualization [42]. The higher the number of topics, the lower the model’s perplexity will be, but the perplexity will continue

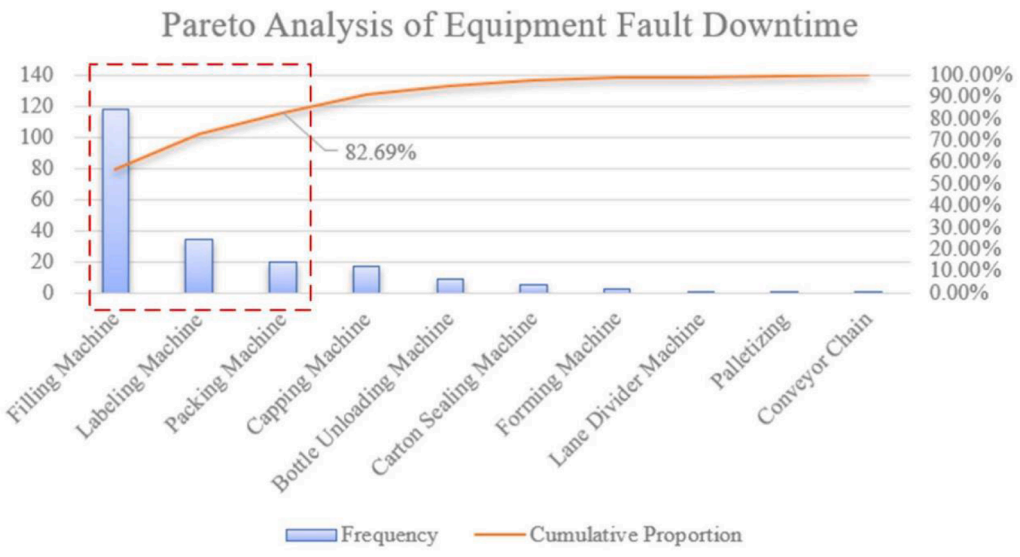


Fig. 5. Pareto analysis of equipment breakdown shutdowns.

to decline with the increase in the number of topics, when the generated model will be overfitting, so it is also necessary to combine with the consistency index to select the optimal number of topics comprehensively. In this study, we set the number of potential themes in the range of [1,10] as shown in Fig. 6, with a continuous decrease in the perplexity level while the consistency level has a localized maximum at a theme number of four. And in the visualized bubble map in Fig. 7, the bubbles are uniform in size and do not overlap, and the theme classification is largely consistent with the Pareto analysis, so the choice of 4 number of themes as the optimal theme is robust.

Table 2 presents the distribution of topic keywords identified in this study, which indicate four principal categories of equipment failures: Topic 1, failures in the filling process; Topic 2, failures in the labeling process; Topic 3, failures caused by bottle breakage in various stages; and Topic 4, failures in the boxing process.

A combination of keywords from the four subject categories and on-site observations revealed that: the main downtime issues for filling machines were clogged nozzle lines, clogged filters, jammed inner caps, jammed outer caps, and excessive foaming; the main causes of downtime for labeling machines included misaligned labels, damaged labels, and blurred print codes; and the main causes of downtime for case packers were loose or unlit sensors (photo-eyes), bottle grippers that were not in alignment, and conveyor belt slippage. Also, bottle breakage can occur at various points in the production line, which can lead to equipment downtime.

4.4. *Improve phase*

In response to major downtime issues with the filling machines, organization and tidying up was first enhanced by sorting tools and items around the equipment and introducing a Poka-yoke design, where items are painted in eye-catching colors to ensure that maintenance staff can quickly find the tools and accessories they need (Fig. 8). Next, cleaning processes were standardized, and nozzles and filters were cleaned regularly, especially in areas where dirt tends to accumulate, to prevent clogging problems. At the same time, standardized operating procedures for equipment maintenance were developed and implemented to reduce operating abnormalities caused by excessive machine foam through division of responsibilities and regular inspections. Finally, in response to the problem of cap jamming, we have largely reduced the downtime caused by cap jamming by replacing the original cap material with a cap that separates the inner cap from the outer cap and a cap that integrates

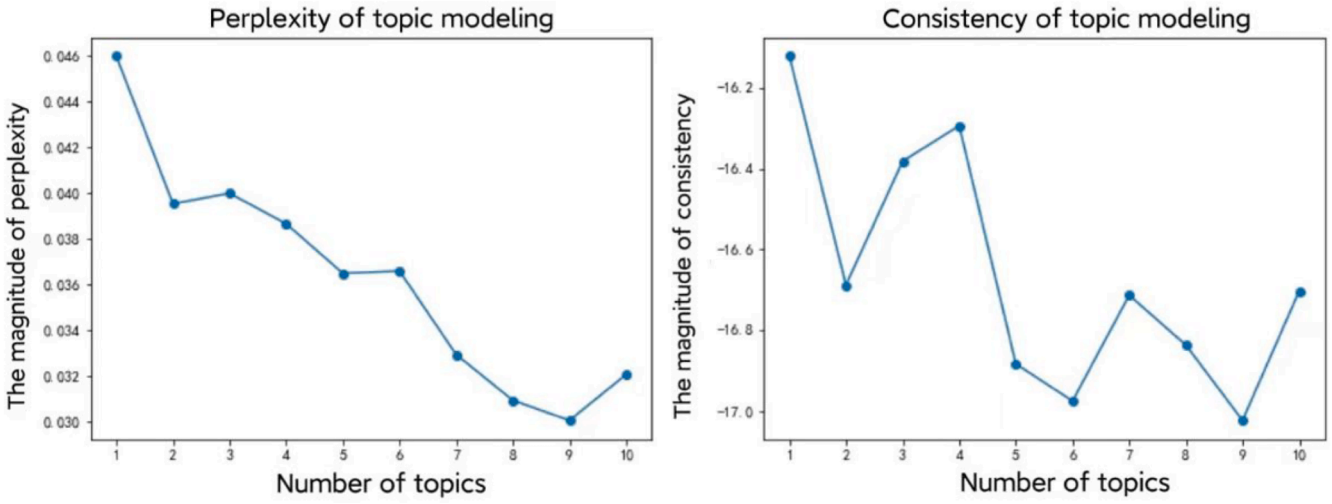


Fig. 6. The perplexity and consistency of topic modeling.

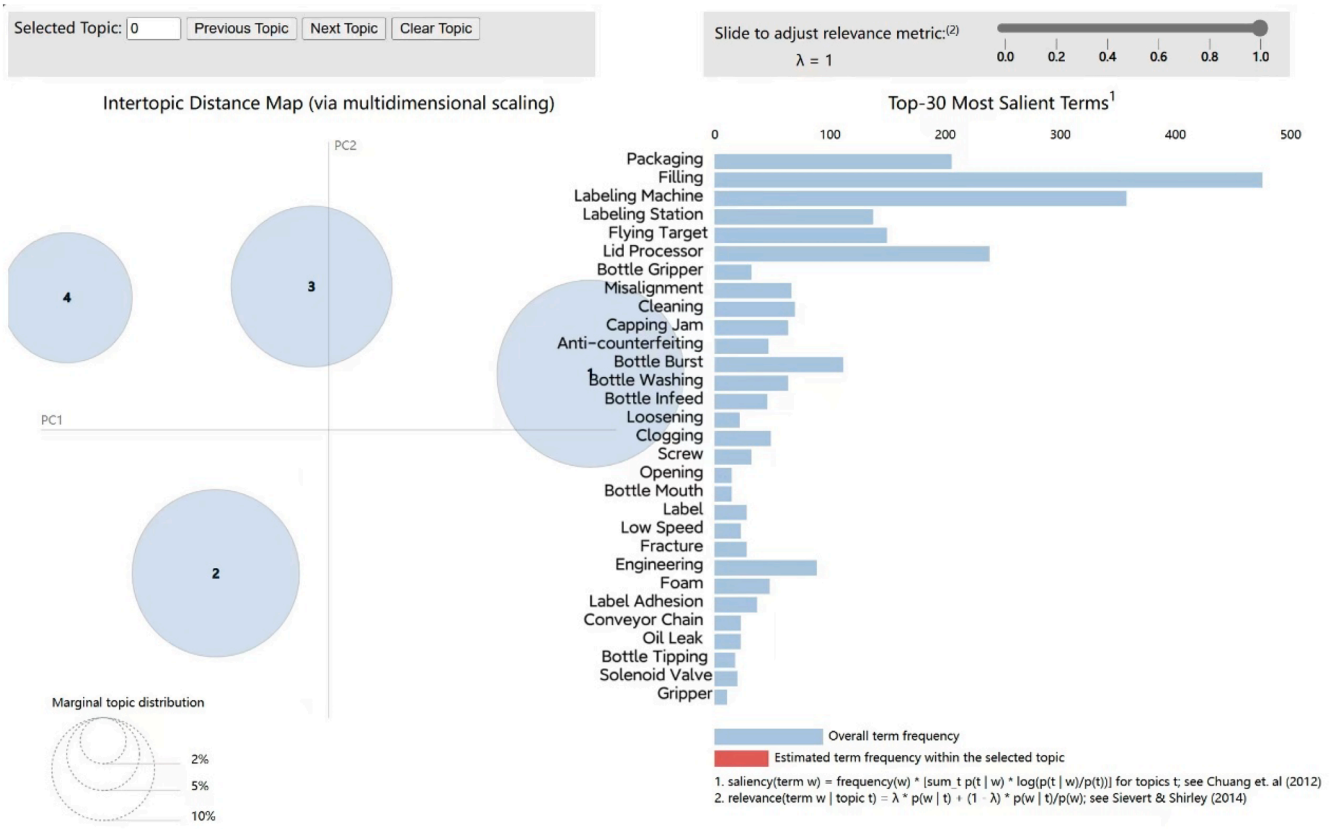


Fig. 7. Equipment Failure Topic Model.

the inner cap with the outer cap.

For the main downtime problems of the labeling machine, through the production process, constantly adjust the transmission system and positioning device of the labeling machine, to ensure the stability of the label in the conveying process, and strengthen the precise control of the labeling position. At the same time, optimize the nozzle settings of the coding equipment to ensure that the ink flow and printing distance is in the best state, and regularly clean the nozzles to prevent ink clogging affect the quality of coding.

In response to the major downtime problems of the case packer, we ensure that the photo eye remains stable during operation by tightening the mounting parts. At the same time, check the installation status regularly and deal with loose parts in time. Regularly check the power supply and wiring connections of the photo eye to avoid failure due to poor contact or aging of the wiring, and develop a maintenance plan for cleaning the surface of the photo eye to prevent dirt from affecting the detection effect.

To address the problem of bottle bursting that occurs in various segments, the star wheel parts are adjusted as the entry point. The star wheel is a key component in the conveyor system of the production line, and its main function is to guide and position the bottles so as to keep them stable during the processing such as conveying, filling and cleaning. If the gap, position or rotational speed of the star wheel is not adjusted in place, it may cause bottles to be subjected to excessive stress or collision during operation, thus increasing the risk of bottle bursting. Therefore, by constantly adjusting and polishing the star wheel components to optimize the gap, ensure that the rotational speed is matched, and avoid excessive squeezing of bottles, the possibility of uneven stress on bottles can be effectively reduced, thus reducing the probability of bottle bursting and improving the stability and safety of the production line.

In summary, as shown in Fig. 9, through the implementation of a series of targeted improvement measures, the overall equipment efficiency (OEE) of the production line improved significantly between weeks 4 and 12, rising steadily from 71 % to 84 %, an increase of 13 percentage points. Together, these measures have enhanced the stability, safety and efficiency of the production line, and the continued rise in OEE fully validates the effectiveness of the improvements.

4.5. Control phase

In the control phase, using the same steps as in the analysis phase, LDA topic modeling was performed on the failure records during the improvement phase period to identify the distribution and evolution of the topics that lead to equipment failures during the improvement process. The results of the visualization are shown in Fig. 10, and the main themes that still exist that can lead to equipment failures are: topic 1 failures in the conveying process, topic 2 failures in the cleaning process, and topic 3 failures in the coding process.

For topic 1, in the control of the conveying link, establish a perfect monitoring system to ensure that every part of the conveying link can be tracked and detected in real time. Installing monitoring equipment in key parts such as conveyor belts, sensors and drive systems to monitor their operating status in real time can effectively detect potential failure risks. In addition, the establishment of standard operating procedures and a failure warning mechanism ensures quick response and resolution when problems arise.

For topic2, in the control of the cleaning process, optimize the cleaning process and the use of tools to ensure that all cleaning steps comply with the standard operating procedures, to avoid equipment failures caused by the use of improper tools or chemicals. We also establish a cleaning quality inspection system, regularly evaluate the cleaning effect, and timely rectify the problematic cleaning links to ensure stable cleaning quality. At the same time, strengthen the operation of the staff training, improve their understanding of the importance of the cleaning process and compliance with operating standards,

Table 2
The distribution of topic terms.

	Topic	Term
Topic1	Filling-stage failures	Filling, Cap Sorter, Cleaning, Cap Jam, Engineering, Blockage, Labeling Machine, Star Wheel, Foam, Bottle Washing, Oil Leakage, Half Bottle, Solenoid Valve, Misalignment, Overload, Label Sticking, Label Fly-off, Bottle Loading, Maintenance, Boxing, Inner-Cap Channel, Waiting, Tank Change, Blocked Passage, Cap-Separation Disc, Filter, Cap Disc, Cap Feeding, Infeed, Pipeline, Soy Sauce, Conveyor Chain, Filter Mesh, Pressure, Alarm, Cause, Cap Blockage, Photo-Eye, Screw, Production, Finished Product, Sensor, Bottle Body, Machine Freeze, Unable to Open, Low Speed, Air Tube, Search, Connection Point, Frequent.
Topic2	Labeling-stage failures	Labeling Machine, Label Fly-off, Labeling Station, Cap Sorter, Bottle Breakage, Anti-counterfeiting, Engineering, Filling, Bottle Infeed, Label Sticking, Lead Screw, Maintenance, Fracture, Star Wheel, Screw, Boxing, Fastening, Barcode, Label Misalignment, Misalignment, Photo-eye, Alarm, Detachment, Rubber, Half Bottle, Open, Finished Product, Bottle Unloading, Deviation, Air Tube, Position, Tripod, Removal, Damage, Label, Bottle Washing, Foam, Bottle Feeding, Blowing Caps, Bottle Jam, Guard Plate, Gluing, Tilt, Full Bottle, Tear, Inspection, Cap Jam, Proportional Valve, Feeding, Gripper Head.
Topic3	Bottle-breakage-induced failures at various stages	Filling, Boxing, Labeling Machine, Bottle Breakage, Bottle Washing, Labeling Station, Foam, Star Wheel, Label Fly-off, Blockage, Low Speed, Engineering, Cap Sorter, Label, Fracture, Maintenance, Filter Mesh, Cleaning, Platform, Gripper Head, Misalignment, Oil Spray, Capping, Cap Press, Anti-counterfeiting, Detachment, Safety Door, Alarm, Speed, Operation, Photo-eye, Glue, Inverted Bottle, Production, Bottle Jam, Bottle Tray, Conveyor Chain, Machine Gripper, Tank Change, Start-up, Infeed Washing, Waiting, Pressure, Running, Cleaning, Bottle Infeed, Bottle Loading, Standby, Drum, Lead Screw.
Topic4	Boxing-stage failures	Boxing, Labeling Machine, Misalignment, Label Fly-off, Bottle Gripper, Labeling Station, Bottle Infeed, Loose, Bottle Breakage, Filling, Conveyor Chain, Bottle Neck, Opening, Label, Lead Screw, Tipped Bottle, Screw, Photo-eye, Machine Gripper, Cap Sorter, Case Infeed, Capping Section, Pipeline, Cap Feeding, Deviation, Belt, Air Leak, Corner Curl, Cleaning, Capping, Fastening, Platform, Cap Blockage, Jamming, Dropping, Tilt, Fixing, Glue Volume, Material Pushing, Beam, Bottle Gripper Head, Damage, Star Wheel, Safety Door, Cap Press, Tank Change, Tripod, Filter, Finished Product, Soy Sauce.

effectively reduce the risk of equipment failure caused by human factors.

For topic3, in the control of the coding link, ensure the regular maintenance and inspection of coding equipment, especially the cleaning and maintenance of printheads and ink systems, ensure that the coding equipment undergoes the necessary inspections and adjustments before and after use, and discover possible wear and tear or clogging problems in a timely manner. At the same time, the implementation of the coding quality monitoring system, the coding effect of each batch of random checks and feedback to ensure that the quality of coding meets the requirements.

Based on this, a heat map of theme intensity is further developed to show the change in intensity of the equipment failure themes present during the improvement phase period, helping us to identify the evolutionary trend of each theme over time. The heat map reflects the heat level of each theme within a specific time period by color shades, i.e., the frequency or intensity of the theme's occurrence within that time period. As shown in Fig. 11, the horizontal coordinate represents the time dimension and the vertical coordinate represents the theme dimension. The darker the color in the heat map means the higher the heat of the corresponding theme in that period, and vice versa. In the control phase, the problems that still existed and would lead to equipment failure were effectively controlled, and the control measures not only improved the stability and productivity of the equipment, but also enhanced the reliability of the overall production process, guaranteed the stability of product quality, and achieved the goal of continuous improvement.

Fig. 8. On-site Poka-yoke Examples.

5. Conclusions

Compared with traditional methods, this study is able to identify root causes and potential themes affecting equipment downtime more effectively by applying the LDA method to production line equipment failure analysis. In the analysis stage, when the root causes of the problems are complex and not clearly categorized, the LDA method can identify potential correlates through theme clustering and discover undefined problem clues. Traditional Lean analysis tools such as Pareto charts require manual labeling of problem clusters and reflect the surface of the problem, not revealing the complex cause and effect relationships. In the control phase, traditional lean tools such as control charts focus more on real-time responses to known indicators and identify abnormal fluctuations through statistically structured data. LDA's thematic evolution approach is more forward-looking, generating evolutionary path diagrams to capture long-term dynamic trends of implicit semantic associations and warn of potential systemic failure risks. Meanwhile, the integration of the LDA approach into the lean manufacturing system in this study also has the advantage of cross-environmental applicability because of its synergistic capability of structured problem diagnosis and data-driven logic. By parsing unstructured data such as equipment failure logs and process parameter records through text mining techniques, common failure themes in different production environments can be automatically clustered, and this semantically-associated root cause analysis breaks through the limitations of traditional empirical analysis. Whether it is a highly automated factory or a labor-intensive packaging line, the LDA method

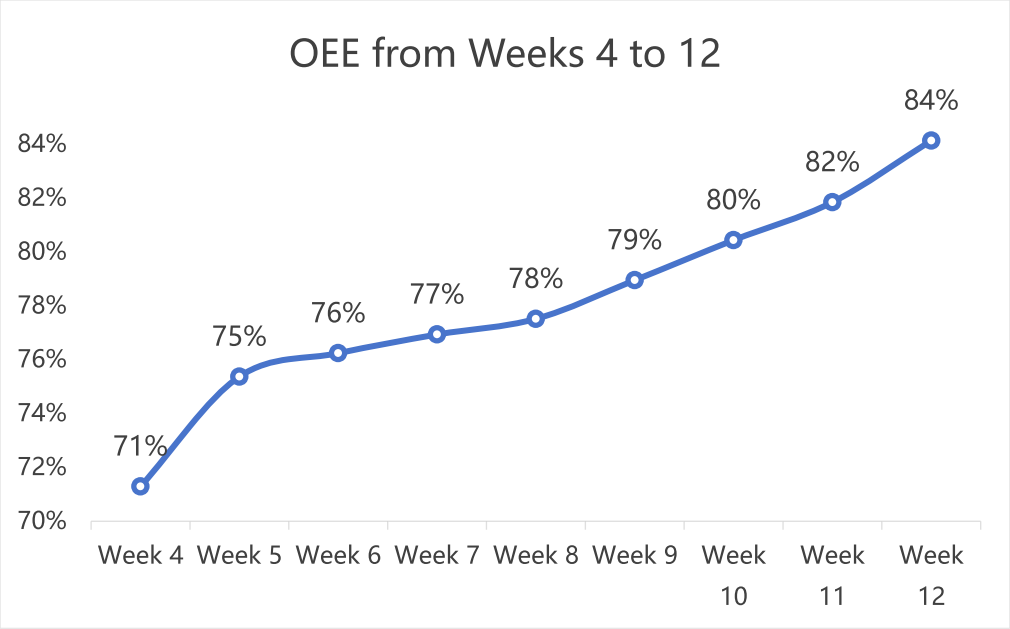


Fig. 9. The OEE trend from the 4th to the 12th week.

can extract different potential problems by analyzing text records.

In the social and environmental aspects, the integration of LDA methodology into the lean production system can systematically identify the resource waste patterns and environmental impacts implicit in the production process, thus providing a scientific basis for decision-making to optimize resource utilization, reduce carbon emissions and achieve sustainable development. By analyzing unstructured data such as production documents, process records or environmental monitoring reports, "waste themes" such as abnormal energy consumption, material redundancy and inefficient processes can be unearthed and synergized with the Lean Manufacturing's resource lifecycle management and continuous improvement mechanisms. This combination not only strengthens lean production's insight into "hidden waste" such as energy loss in non-value-added activities, but also quantifies the priority of environmental impacts through the probability distribution of themes, guiding enterprises to accurately invest resources in process reengineering, supply chain synergy, and green technological innovation. At the same time, LDA's dynamic thematic evolution analysis can track the effects of improvement measures over time, which helps companies build eco-lean systems that balance productivity, social responsibility and ecological benefits.

In terms of methodological applicability, expanding the application of LDA methodologies to non-manufacturing or automated systems requires adaptations around the core capabilities of topic modeling and data parsing. In the non-manufacturing sector, LDA can be used to optimize business processes by analyzing textual data from business processes to uncover themes of potential efficiency bottlenecks or resource wastage. In automation systems, LDA can be combined with real-time sensor data and operation logs to achieve anomaly monitoring and decision-making optimization by dynamically updating topic models. Maintenance records of automated equipment can be mapped to production parameters as "fault-condition" correlation topics, which can assist in the optimization of predictive maintenance algorithms. The key to this cross-domain migration is to transform the "value stream analysis" logic of lean manufacturing into a data-driven topic clustering task, and at the same time break through the limitations of traditional manufacturing data structures through algorithmic improvements.

In this study, in terms of the generalizability of the method, it can adapt to the data characteristics of different production scenarios, and the model can be flexibly adapted to differentiated production environments, such as discrete manufacturing and process industries, through the selection of the number of topics and the supplementation of deactivated words. At the horizontal scalability level, the LDA method can be used in intelligent manufacturing workshops by deploying edge nodes for localized topic extraction of production anomaly reports generated in real-time, and then aggregating the global topic distribution, which not only meets the demand for low-latency response but also safeguards data privacy. At the vertical extension level, it can be combined with different lean tools. When combined with value stream mapping, non-value-added links can be identified by parsing theme clustering in process documents; when combined with the On-Light system, the theme evolution trend of equipment failure logs can be mapped to the prediction of anomaly patterns in physical production lines.

In summary, through theoretical analysis and case validation, this study reveals the synergistic driving mechanism of the deep integration of LDA methodology and Lean Six Sigma framework on the operational excellence of enterprises, which breaks through the dependence of traditional Lean tools on structured data and promotes the quality improvement from passive corrective to active preventive, and provides methodological support for the construction of the data-driven sustainable operational excellence paradigm for enterprises. Support.

6. Limitations and future research

The application of LDA methods in lean manufacturing needs to rely on the structured recording of text data such as faults and production logs by enterprises, but language differences in different countries or regions may lead to limited modeling effectiveness. If the enterprise is located in a region that uses a small language or dialect, it also needs to customize and adjust the word segmentation rules and semantic database. Meanwhile, the LDA method relies on a large amount of text data for topic mining, but many traditional manufacturing enterprises have not yet established a systematic digital record system, which can lead to data fragmentation or lack of data.

Future research can combine pre-trained language models to optimize the ability of participle and semantic understanding, reduce the dependence on manually customized deactivated word lists, develop LDA enhancement tools that support multiple languages, adapt industrial terminology from different regions, and reduce the technical

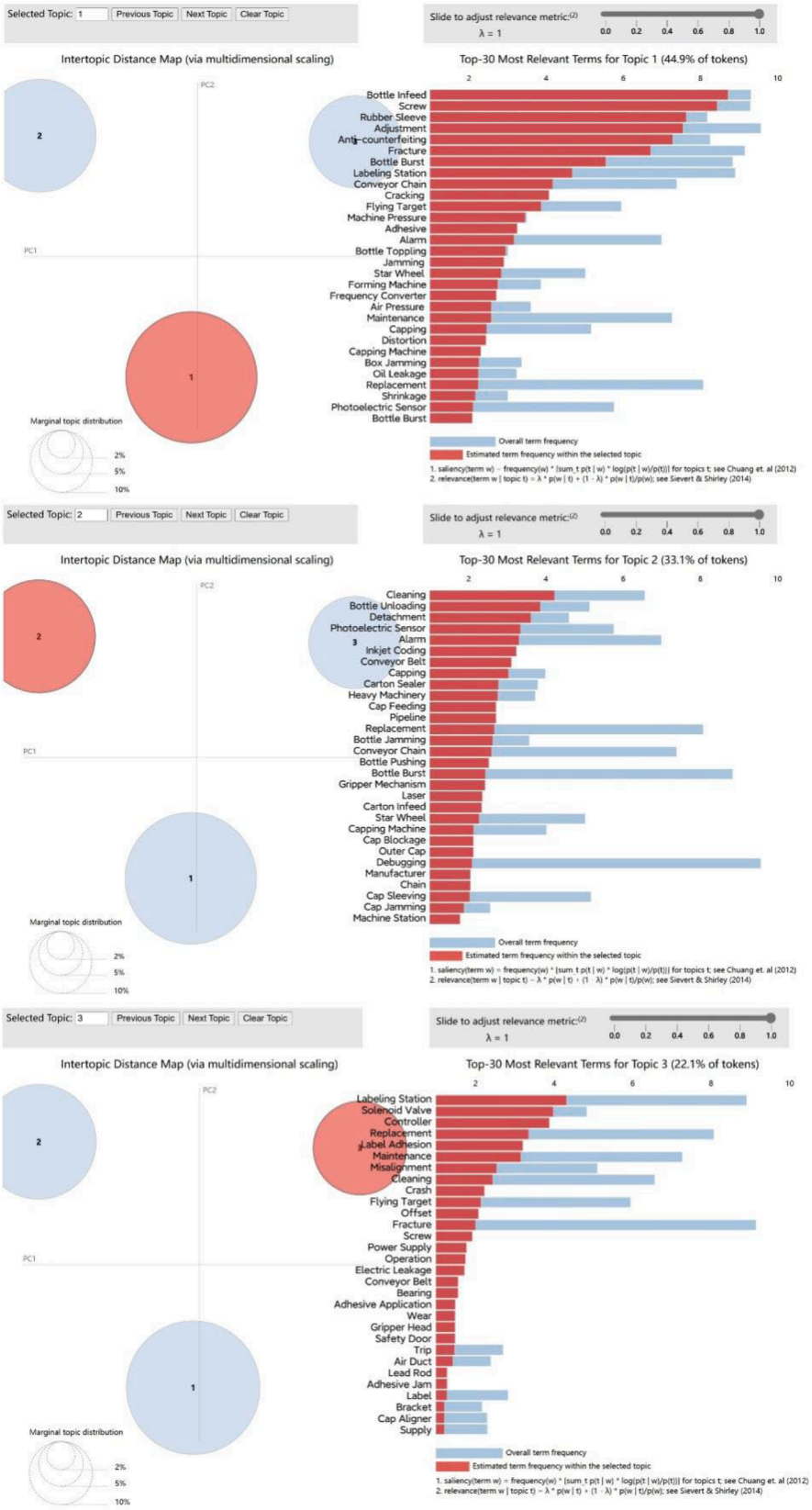


Fig. 10. The topics of equipment failures that still exist during the improvement process.

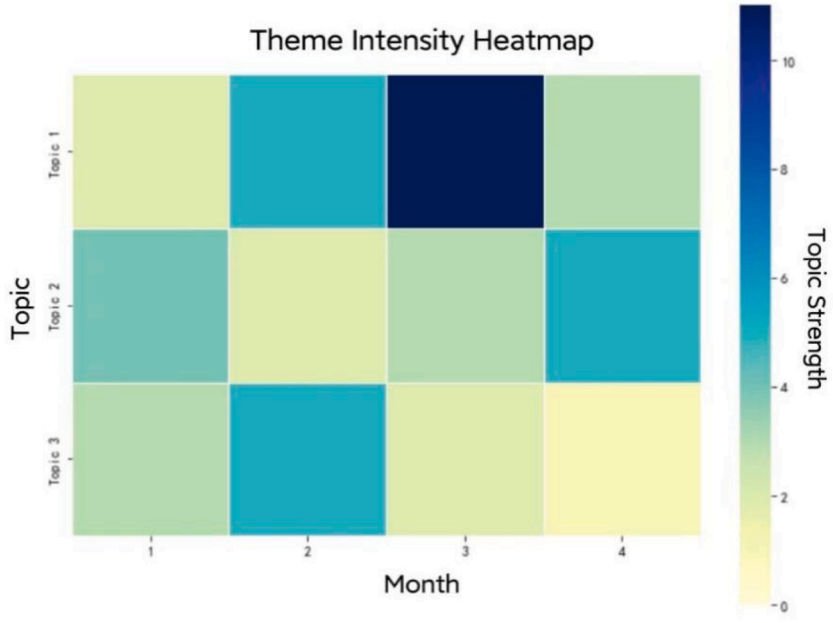


Fig. 11. Topic heat maps in different periods.

threshold. At the same time, we build an interactive LDA analysis platform that allows frontline employees to provide feedback on problems through natural language, and allows the model to generate optimization suggestions in real time and feed them back to the production management system, forming a closed loop of "problem discovery-improvement implementation". Although the current application of LDA method in lean production is limited by language, data base and other factors, with the advancement of technology and the deepening of the digital transformation of enterprises, its potential will be gradually released to become one of the core tools to promote intelligent manufacturing and continuous improvement.

Consent for publication

All authors of this manuscript have provided their consent for the publication of this research.

CRediT authorship contribution statement

Youbin Chen: Writing – review & editing, Writing – original draft, Methodology, Data curation. **Da Huang:** Writing – original draft. **Lifei Wang:** Data curation.

Declaration of competing interest

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

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The data that has been used is confidential.

References

[1] M. Kurdve, M. Bellgran, Green lean operationalisation of the circular economy concept on production shop floor level, J. Clean. Prod. 278 (2021) 123223.
[2] Z. Hao, C. Liu, M. Goh, Determining the effects of lean production and servitization of manufacturing on sustainable performance, Sustain. Prod. Consum. 25 (2021) 374–389.
[3] F. Dillinger, B. Tropschuh, M.Y. Dervis, G. Reinhart, A systematic approach to identify the interdependencies of lean production and industry 4.0 elements, Procedia CIRP. 112 (2022) 85–90.
[4] V. Saddikutti, P.K. Gudavalleti, M.P. Singh, Lean and Legacy supply chains for coordinated demand driven production to handle disruptions, IFAC-PapersOnLine 55 (10) (2022) 2846–2851.
[5] L. Potthoff, L. Gunnemann, Resilience of lean production systems: a systematic literature review, Procedia CIRP. 120 (2023) 1315–1320.
[6] H. Huang, et al., Spatial classification model of port facilities and energy reserve prediction based on deep learning for port management–A case study of Ningbo, Ocean. Coast. Manage. 258 (2024) 107413, https://doi.org/10.1016/j.ocecoaman.2024.107413, 2024/11/01/.
[7] X. Huang, et al., Deep reinforcement learning-based health-conscious energy management for fuel cell hybrid electric vehicles in model predictive control framework, Energy 304 (2024) 131769, https://doi.org/10.1016/j.energy.2024.131769, 2024/09/30/.
[8] J. Li, C. Yuan, X. Wang, G. Chen, G. Ma, Semi-supervised crack detection using segment anything model and deep transfer learning, Autom. Constr. 170 (2025) 105899, https://doi.org/10.1016/j.autcon.2024.105899, 2025/02/01/.
[9] D. Mu, C. Yue, A. Chen, Are we working on the safety of UAVs? An LDA-based study of UAV safety technology trends, Saf. Sci. 152 (2022) 105767, https://doi.org/10.1016/j.ssci.2022.105767, 2022/08/01/.
[10] S. Jin, G. Stokes, C. Hamilton, Empirical evidence of urban climate adaptation alignment with sustainable development: application of LDA, Cities. 136 (2023) 104254, https://doi.org/10.1016/j.cities.2023.104254, 2023/05/01/.
[11] N. Li, S. Li, Research on the LDA-ECD based support policy for China's agricultural cold chain logistics, Sustain. Futures 9 (2025) 100460, https://doi.org/10.1016/j.sftr.2025.100460, 2025/06/01/.
[12] A. da Silva, A. Dionísio, L. Coelho, Flexible-lean processes optimization: a case study in stone sector, Results Eng. 6 (2020) 100129, https://doi.org/10.1016/j.rineng.2020.100129, 2020/06/01/.
[13] M.A. Habib, R. Rizvan, S. Ahmed, Implementing lean manufacturing for improvement of operational performance in a labeling and packaging plant: a case study in Bangladesh, Results Eng. 17 (2023) 100818, https://doi.org/10.1016/j.rineng.2022.100818, 2023/03/01/.
[14] E. Dong, et al., A data-driven intelligent predictive maintenance decision framework for mechanical systems integrating transformer and kernel density estimation, Comput. Ind. Eng. 201 (2025) 110868, https://doi.org/10.1016/j.cie.2025.110868, 2025/03/01/.
[15] L. Wang, B. Li, X. Zhao, Multi-objective predictive maintenance scheduling models integrating remaining useful life prediction and maintenance decisions, Comput. Ind. Eng. 197 (2024) 110581, https://doi.org/10.1016/j.cie.2024.110581, 2024/11/01/.

[16] S. Lee, D. Kim, E. Yeom, Predictive maintenance using estimation from time interval for butterfly valves, Results Eng. 26 (2025) 104609, https://doi.org/10.1016/j.rineng.2025.104609, 2025/06/01/.
[17] V.M. Nesro, T. Fekete, H. Wicaksono, Leveraging causal machine learning for sustainable automotive industry: analyzing factors influencing CO2 emissions, Procedia CIRP. 130 (2024) 161–166.
[18] V. Sudarshan, W.D. Seider, Advancing machine learning in industry 4.0: benchmark framework for rare-event prediction in chemical processes, Comput. Chem. Eng. 194 (2025) 108929.
[19] X. Xue, N. Hu, F. Qiu, G. Li, Chief operating officers' long-term orientation and corporate lean production, J. Bus. Res. 191 (2025), https://doi.org/10.1016/j.jbusres.2025.115247.
[20] M.K. Lim, M. Lai, C. Wang, S.Y. Lee, Circular economy to ensure production operational sustainability: a green-lean approach, Sustain. Prod. Consum. 30 (2022) 130–144, https://doi.org/10.1016/j.spc.2021.12.001.
[21] M. Eskandari, M. Hamid, M. Masoudian, M. Rabbani, An integrated lean production-sustainability framework for evaluation and improvement of the performance of pharmaceutical factory, J. Clean. Prod. 376 (2022) 134132, https://doi.org/10.1016/j.jclepro.2022.134132, 2022/11/20/.
[22] D. Pearce, M. Dora, J. Wesana, X. Gellynck, Toward sustainable primary production through the application of lean management in South African fruit horticulture, J. Clean. Prod. 313 (2021) 127815, https://doi.org/10.1016/j.jclepro.2021.127815, 2021/09/01/.
[23] A. Bilancia, F. Costa, A.P. Staudacher, Achieving sustainability and circular economy in the luxury fashion industry through lean practices: a systematic literature review, Comput. Ind. Eng. (2025) 111107, https://doi.org/10.1016/j.cie.2025.111107, 2025/04/11/.
[24] A. Mittal, P. Gupta, V. Kumar, A. Al Owad, S. Mahlawat, S. Singh, The performance improvement analysis using six Sigma DMAIC methodology: a case study on Indian manufacturing company, Heliyon 9 (3) (2023) e14625, https://doi.org/10.1016/j.heliyon.2023.e14625. Mar.
[25] I. Vicente, R. Godina, A.Teresa Gabriel, Applications and future perspectives of integrating Lean six Sigma and ergonomics, Saf. Sci. 172 (2024), https://doi.org/10.1016/j.ssci.2024.106418.
[26] L.M. Gaikwad, U. Bhushi, S.N. Teli, Implementation of Six Sigma methodologies to gain a competitive advantage: a Case Study approach, in: *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, 21-24 Feb. 2022, 2022, pp. 1–4, https://doi.org/10.1109/ASET53988.2022.9735103.
[27] I.T.B. Widiwati, S.D. Liman, F. Nurprihatin, The implementation of Lean Six Sigma approach to minimize waste at a food manufacturing industry, J. Eng. Res. (2024), https://doi.org/10.1016/j.jer.2024.01.022, 2024/02/08/.
[28] M.E. Natarus, et al., Optimization of a sterile processing department using lean six sigma methodology, staffing enhancement, and capital investment, Joint Commission J. Qual. Patient Saf. 51 (1) (2025) 33–45, https://doi.org/10.1016/j.jcjq.2024.10.006, 2025/01/01/.
[29] D.M. Utama, M. Abirfatin, Sustainable Lean six-sigma: a new framework for improve sustainable manufacturing performance, Clean. Eng. Technol. 17 (2023), https://doi.org/10.1016/j.clet.2023.100700.
[30] J. Zimmermann, L.E. Champagne, J.M. Dickens, B.T. Hazen, Approaches to improve preprocessing for Latent Dirichlet Allocation topic modeling, Decis. Support. Syst. 185 (2024), https://doi.org/10.1016/j.dss.2024.114310.
[31] X. Chen, C. Zanocco, J. Flora, R. Rajagopal, Constructing dynamic residential energy lifestyles using Latent Dirichlet Allocation, Appl. Energy 318 (2022), https://doi.org/10.1016/j.apenergy.2022.119109.
[32] P. Madzík, L. Falát, D. Zimon, Supply chain research overview from the early eighties to Covid era – Big data approach based on Latent Dirichlet Allocation, Comput. Ind. Eng. 183 (2023), https://doi.org/10.1016/j.cie.2023.109520.
[33] Y. Guo, F. Wang, C. Xing, X. Lu, Mining multi-brand characteristics from online reviews for competitive analysis: a brand joint model using latent Dirichlet allocation, Electron. Commer. Res. Appl. 53 (2022), https://doi.org/10.1016/j.elerap.2022.101141.
[34] M.Motta Queiroz, C. Roque, F. Moura, J. Marôco, Understanding the expectations of parents regarding their children's school commuting by public transport using latent Dirichlet Allocation, Transport. Res. Part A 181 (2024), https://doi.org/10.1016/j.tra.2024.103986.
[35] J. Kim, W. Shin, S. Han, S. Moon, J.-J. Kim, Enhancing occupant experience in defect repair services through text mining-based latent dirichlet allocation metric identification, Develop. Built Environ. 17 (2024), https://doi.org/10.1016/j.dibe.2024.100354.
[36] H. Xiong, Y. Cheng, W. Zhao, J. Liu, Analyzing scientific research topics in manufacturing field using a topic model, Comput. Ind. Eng. 135 (2019) 333–347, https://doi.org/10.1016/j.cie.2019.06.010, 2019/09/01/.
[37] C.-H. Lee, C.-L. Liu, A.J.C. Trappey, J.P.T. Mo, K.C. Desouza, Understanding digital transformation in advanced manufacturing and engineering: a bibliometric analysis, topic modeling and research trend discovery, Adv. Eng. Info. 50 (2021) 101428, https://doi.org/10.1016/j.aei.2021.101428, 2021/10/01/.
[38] P. Madzík, L. Falát, N. Yadav, F.L. Lizarelli, K. Čarnogurský, Exploring uncharted territories of sustainable manufacturing: a cutting-edge AI approach to uncover hidden research avenues in green innovations, J. Innov. Knowl. 9 (3) (2024) 100498, https://doi.org/10.1016/j.jik.2024.100498, 2024/07/01/.
[39] P. Tsarouhas, N. Sidiropoulou, Application of Six Sigma methodology using DMAIC approach for a packaging olives production system: a case study, Int. J. Lean Six Sigma 15 (2) (2023) 247–273.
[40] S. Nakajima, Introduction to TPM: Total Productive Maintenance.(Translation), Productivity Press, Inc., 1988, p. 129, 1988.
[41] A. Muñoz-Villamizar, J. Santos, J.R. Montoya-Torres, C. Jaca, Using OEE to evaluate the effectiveness of urban freight transportation systems: a case study, Int. J. Prod. Econ. 197 (2018) 232–242.
[42] Y. Guo, B. Ayoun, S. Zhao, Is someone listening to me? A topic modeling approach using guided LDA for exploring hospitality value proposition through online employee reviews, Int. J. Hosp. Manage. 127 (2025) 104114, https://doi.org/10.1016/j.ijhm.2025.104114, 2025/05/01/.