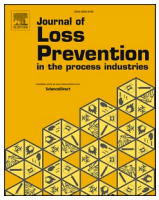
Journal of Loss Prevention in the Process Industries 80 (2022) 104911

Contents lists available at ScienceDirect

Journal of Loss Prevention in the Process Industries

journal homepage: www.elsevier.com/locate/jlp



Intelligent HAZOP analysis method based on data mining

Feng Wang *, Wunan Gu

National Foundation Research Laboratory of Fault Prevention and Control in Hazardous Chemicals Production System, Engineering Research Center of Chemical Technology Safety Ministry of Education, Beijing University of Chemical Technology, Beijing, China

ARTICLE INFO

Keywords:
HAZOP
Naive bayes
LDA
Data mining

ABSTRACT

Accidents in production equipment are occurring with increasing frequency, which has caused serious negative impacts on society and petrochemical enterprises. It is important to learn prevention and control rules and lessons from accident cases to prevent accidents and improve intrinsic safety levels of enterprises. The hazard and operability (HAZOP) analysis method has been widely used for risk identification and accident prevention in petrochemical enterprises. However, the rules contained in many completed HAZOP analysis reports have not been deeply studied, inherited, or extensively applied. The analysis of a new process, however, relies on expert experience, which is usually subjective and time-consuming. This paper proposes an intelligent HAZOP analysis method based on data mining to explore the law of accident cause mechanisms, help promote intelligent analysis, and improve the accuracy of analysis results. A HAZOP analysis data table structure is proposed. The frequencies of all words in the process parameter, guide word, cause, and consequence were calculated, and the words were sorted based on the word frequency statistics method. Words with high word frequencies were considered key in formulating the safety inspection checklists. The latent Dirichlet allocation (LDA) model and contingency table were used to explain the correlation between the process parameters, guide words, causes, and consequences. The co-occurrence probability sequence of the elements can then be acquired. Using the naive Bayes algorithm, a method for calculating the likelihoods, severities, and risk levels of accidents is proposed. The accident risk ranks can be intelligently predicted to realize an intelligent HAZOP analysis. This study developed a software to help execute the method. The intelligent HAZOP analysis method was applied based on data from 14 petrochemical units; the raw oil buffer tank in the wax oil hydrocracking unit was taken as an example to illustrate the application of the method. The method provides a technical basis and basic guarantee for risk identification, accident prevention, and rescue of petrochemical plants.

1. Introduction

The complex operation process of petrochemical plants contains a deal of unknown uncertainty, which is one of the most important reasons causing accidents resulting in unexpected injury and death, economic losses to society, and harmful impacts on the environment. Therefore, the unknown uncertainty should first be identified, analyzed, and determined, and then effective measures and means should be adopted to prevent accidents. More attention is needed to acquire abnormal reasons, predict consequences, prepare safeguards, and conduct emergency responses. Hazard and operability (HAZOP) analysis, the most commonly used risk identification and analysis method, can be employed to analyze the causes and possible adverse consequences by starting from the process parameters or guide words and can be used to grade the event likelihood and consequence severity based on the experience of the analyzers and experts. Considering all the data and information, including parameters, guide words, causes, consequences, likelihoods, and severities of the potential risk events, will contribute toward determining the risk level of the events and providing measures to prevent accidents. Because all potential scenarios of the petrochemical process could be considered and discussed in conducting the process of the method, most researchers select the HAZOP analysis method to inspect the potential abnormal situations in the production processes of petrochemical industries (Single et al., 2020; Li et al., 2021; Meng et al., 2021). Moreover, the complex accident relationships hidden in several completed HAZOP analysis reports have not been revealed, investigated, and utilized at a deeper level; therefore, it is difficult for analysts to acquire experience and knowledge in the risk analysis process from previous accidents (Suzuki et al., 2021; Nguyen et al., 2022; Zhou et al., 2018). Therefore, it is important to use the completed HAZOP analysis

* Corresponding author.
E-mail address: wangfeng991@163.com (F. Wang).

https://doi.org/10.1016/j.jlp.2022.104911
Received 18 May 2022; Received in revised form 15 September 2022; Accepted 19 October 2022
Available online 25 October 2022

report and explore the hidden rules among the factors for early warning and prevention of petrochemical accidents.

Many scholars use the completed HAZOP report and expert experience to establish an expert knowledge base and use specific reasoning mechanisms to identify the causes, consequences, and transmission paths of deviations, thus realizing the mining, determination, and sharing of accident laws to a certain extent. Wang (Wang et al., 2013) studied the HAZOP expert system framework based on real-time database technology, and established an expert knowledge base dominated by rule-based heuristic knowledge. Wang (Wang et al., 2008) proposed a computer-aided HAZOP analysis technology based on the analytic hierarchy process (AHP) to analyze the influence relationship among factors and sort the risk paths according to their relative importance. Kościelny (Kościelny et al., 2017) established a model in a process diagram and proposed an intelligent accident prevention system supporting HAZOP analysis. However, the construction of most models still depends on analysts' professional experience, which is a heavy workload and susceptible to subjective judgment.

Some scholars conduct text analysis and data mining based on accident database information, determine the causes of the accidents, and describe the accident occurrence processes qualitatively, which helps avoid the recurrence of accidents or minimize the severity of the accident consequences. Zhu (Zhu et al., 2021) used machine learning technology to analyze the key factors of accidents and evaluate the influence of different factor combinations on accident severity. Feng (Feng et al., 2021) used natural language processing technology to construct a HAZOP report event classification model to ensure consistency in the analysis. Marhavilas (Marhavilas et al., 2020) applied fuzzy AHP to the HAZOP analysis method to expand the research scope of HAZOP and rank the plant hazards. The latent Dirichlet allocation (LDA) algorithm is a three-layer Bayesian probabilistic structure model that is widely used in large-scale data clustering because of its high efficiency, speed, and influential clustering algorithm technology (Suh, 2021; Wang and Xu, 2018). Niu (Niu et al., 2019) used the LDA model to extract five accident cause themes and analyzed the major causes of accidents. Zhong (Zhong et al., 2020) used the convolutional neural network (CNN) and LDA models to automatically extract text features, classify accident narratives, examine the interdependence between causal variables, and visualize accident descriptions. However, few studies have combined text data with numerical data analysis results to comprehensively mine hazard laws in HAZOP analysis reports.

An intelligent HAZOP analysis method based on data mining is proposed to reveal the relationships among all the data, including process parameters, guide words, causes, consequences, likelihoods, severities, and risk levels of the potential risk events in the HAZOP analysis data table. The HAZOP analysis data table is based on over 5000 HAZOP event descriptions from 14 petrochemical units. The word frequencies of the process parameters, guide words, causes, and consequences can be calculated, and words with high word frequencies are considered the key attention factors. The information inspection table can be formulated to determine the inspection order. Cluster and association analyses should be conducted to acquire the co-occurrence probabilities of the information of process parameters, guide words, causes, and consequences. The naive Bayes model can establish a risk prediction model trained to calculate the likelihoods, severities, and risk levels. Using this method will be convenient and intelligent for completing a HAZOP analysis report.

2. Method of intelligent HAZOP analysis based on text data mining

This paper proposes a method based on HAZOP text data mining for the intelligent analysis of new petrochemical processes. The method can extract the topic information of causes and consequences, intelligently predict the likelihoods, severities, and risk levels of potential abnormal events, analyze and acquire the accident law, and assist the HAZOP analysis. The method includes the following three parts: (1) establishing a HAZOP analysis data table, (2) conducting text data mining based on the HAZOP analysis data table, and (3) getting the HAZOP analysis results. A flowchart of this method is shown in Fig. 1.

2.1. Establishing a HAZOP analysis data table

According to international standards, the structure and composition of a HAZOP analysis data table can be determined. Currently, there are many standards for classifying equipment, such as 'Petrochemical Enterprise's equipment management requirements' and 'CNPC Equipment Classification Standard'. In the classification standards, 9 equipment types, which were irrelevant to the production process, such as office equipment, Kitchen equipment and medical equipment, are not considered in this study. To make the risk analysis cover all the equipment types in the production process, this study classified the equipment into 29 classifications, which have been listed in Table 1, based on the classification standards, the common equipment types and the expert experience and knowledge, etc. The HAZOP analysis data table includes 13 elements: serial number, plant name, equipment name, equipment classification, process parameter, guide word, cause, consequence, likelihood, severity, risk level, safeguard, and the suggested measure.

2.2. Conducting text data mining based on the HAZOP analysis data table

The HAZOP analysis data table contains both textual and numerical data. The values of the process parameters, guide words, causes, and consequences are part of the text data. The numerical data mainly included the values of likelihoods, severities, and risk levels. This study employed various data-mining techniques to explore the coupling relationship between elements. The LDA model and contingency table were used to analyze and acquire the relationships among the process parameters, guide words, causes, and consequences. The naive Bayes model obtained rules hidden in the numerical data of likelihoods, severities, and risk levels. The analysis results obtained by the machine learning method can provide a technical basis for intelligent HAZOP analysis. In the process of text data mining, it is particularly important considering the large amount of the HAZOP analysis data. The accuracy and reliability of the prediction results of the cluster analysis are positively associated with the amount of the HAZOP analysis data.

2.2.1. Data preprocessing

Prior to data mining, data preprocessing was required to transform the numerical and textual data in the HAZOP analysis data table into structured text data. This process is illustrated in Fig. 2.

①Missing data imputation: Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Generally, there is no data in some cells in the HAZOP data table. Therefore, in the cluster analysis process of the data in the HAZOP data table, the blank cells were filled by using the word 'None'. This method can be helpful for cluster analysis. Using 'None' to fill the blank cells may have a certain impact on the clustering results, but the impact is small and controllable.

②Word segmentation: Word segmentation is the process of dividing written text into meaningful units, such as words or sentences. This study established a word segmentation table, including a huge number of the most frequently used phrases in chemistry and chemical engineering subjects, which can improve the accuracy of the word segmentation of the causes and consequences in the HAZOP data table.

③Stop words filtering: In the HAZOP data table, the causes and consequences mainly include adverbs, adjectives, conjunctions, and some proper nouns, which are called stop words need to be filtered out after the word segmentation.

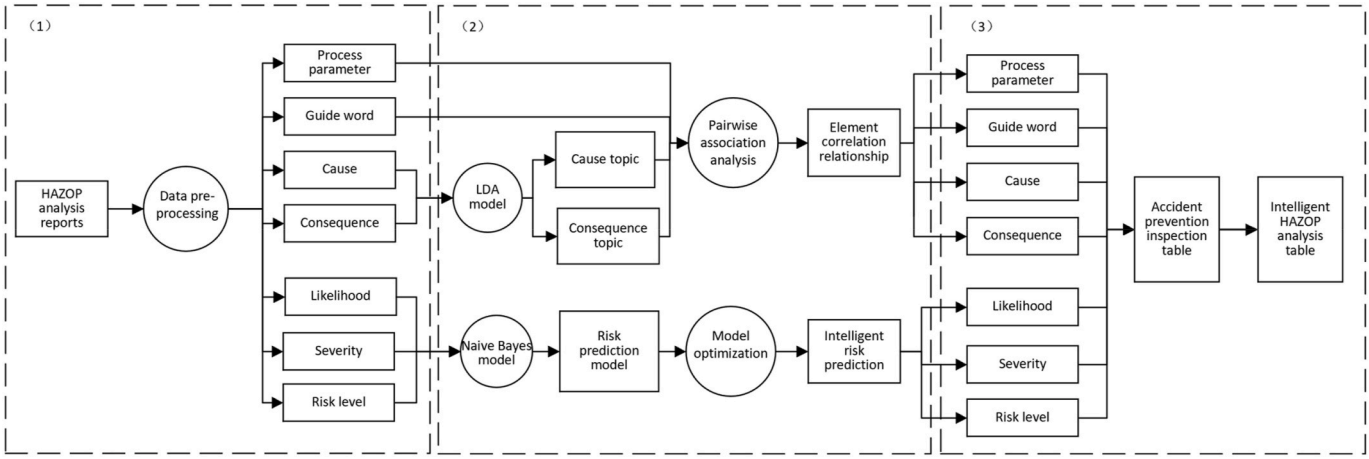


Fig. 1. Flowchart of intelligent HAZOP analysis method based on data mining.

Table 1
Equipment classification.

Equipment classification
Tank, tower, heat exchanging equipment, pump, reactor, furnace, separator, compressor, reboiler, drum, instrument, precipitator, trough, hopper, filter, desulfurization bed, pool, fan, gasification equipment, mixer, deaerator, pipeline, dryer, purging system, gas turbine, converter, valve, injector, others.

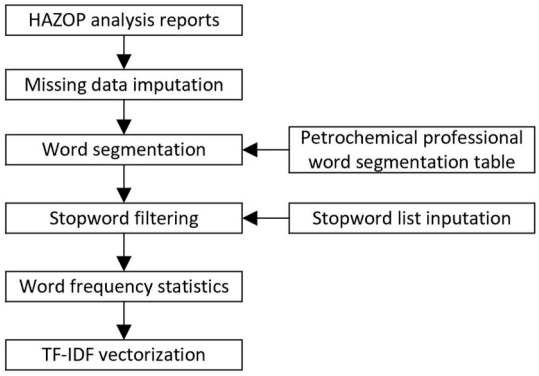


Fig. 2. Data preprocessing process.

④Word frequency statistics: The word frequencies of all words in each text element were calculated and recorded to build a word index table.

⑤TF-IDF vectorization: Term frequency-inverse document frequency (TF-IDF) is a mainstream feature item weight calculation method. The TF-IDF vectorization method can be used to process the words in cause and consequences with high frequencies to obtain the word vectors, contributing to the clustering analysis and risk prediction.

2.2.2. Text clustering analysis based on LDA model

The HAZOP data table consists of potential risk scenarios, and each scenario includes some elements, such as the cause, consequence, process parameter, and guide word. These elements in each piece of data information can be considered as an event description, that can be used to explain the details of an accident. Therefore, HAZOP analysis data tables can be seen as complex documents containing multiple event descriptions. Latent Dirichlet allocation (LDA) is an unsupervised machine learning method that can identify hidden cause and consequence topics and key information from the HAZOP analysis data table. According to the LDA method, the total number of event descriptions in the data table can be defined as M and the total number of cause/consequence topics can be defined as K. In addition, α is the Dirichlet prior parameter of the multinomial distribution of cause/consequence topics for each event description, and β is the Dirichlet prior parameter of the multinomial distribution of words for each cause/consequence topic (Wang and Xu, 2018; Niu et al., 2019). Topics should be independently generated for each event description; thus, the topic generation probability of each event description in the data table can be calculated using Equation (1).

$$p = (Z|\alpha) = \prod_{m=1}^{M} \frac{\Delta(n_m + \alpha)}{\Delta(\alpha)} \tag{1}$$

In Equation (1), $Z = (z_1, \ldots, z_M)$, z_m is the topic number corresponding to all the words in the m-th HAZOP event description; for $n_m = (n_m^{(1)}, \ldots, n_m^{(K)})$, $n_m^{(k)}$ is the number of words in the k-th topic in the m-th HAZOP event description.

Generating words for each topic is independent of each other, and the word generation probability in a topic can be calculated using Equation (2).

$$p = (W|Z, \beta) = \prod_{k=1}^{K} \frac{\Delta(n_k + \beta)}{\Delta(\beta)} \tag{2}$$

In Equation (2), $W = (w_1, \ldots, w_M)$, w_m is a word in the m-th HAZOP event description, and for $n_k = (n_k^{(1)}, \ldots, n_k^{(T)})$, $n_k^{(t)}$ is the number of t-th words produced by the k-th topic.

The above equations can calculate each event description's cause and consequence topic distribution probability, and the distribution probability of words in each topic. The topic of cause/consequence can be summarized according to words with a high probability in each topic. The input of a new event description may change the original statistical analysis results. Therefore, the model can be regularly updated and improved by adding new event descriptions.

2.2.3. Risk prediction based on the naive Bayes model

The naive Bayes algorithm was used to train the classification prediction model and mine risk laws from historical analysis data to explore the influence relationship between cause and consequence in the HAZOP analysis data table, reduce the dependence of the risk matrix on analysts, and realize intelligent risk prediction. The formula is shown in Equation (3) (Dorsey et al., 2020).

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{3}$$

In Equation (3), D is the input HAZOP element, and h is the risk of the

HAZOP event description. The naive Bayes model can calculate the probability that the event description belongs to each risk classification by analyzing and considering the input HAZOP elements and returning the risk with the maximum probability as the accident data. The influence of different element combinations on risk prediction was determined. As the model in this study is a multi-classification model, the accuracy rate should be used as an indicator to evaluate the quality of the model, and the combination with the highest accuracy rate should be selected to intelligently predict the accident risk and help determine the likelihood, everity, and risk level, which is conducive to the construction of a risk matrix.

2.3. Getting the HAZOP analysis result table

Based on word frequency statistics, the word frequency of the process parameters, guide words, causes, and consequences in HAZOP data table can be counted. Based on the LDA clustering, causes and consequences can be clustered, and the cause and consequence topic ID of each analysis data can be obtained. And an information inspection table for accident prevention can be established. The items that need to be inspected can be determined from the clustering results. The inspection order for each type of equipment can be clearly defined. By summarizing and classifying the inspection items, their hierarchical structure can be clarified, and an accident prevention information checklist can be constructed based on the calculation results and expert experience.

The HAZOP data table provides the core of database functionality (for storing information). The information that the operators and analyzers are concerned with, including the equipment name, process parameter, guide word, cause, and consequence, can be combined and searched in the table. Specific HAZOP analysis information for each type of equipment can be found and proposed based on this information, providing technical guidance for accident prevention and control, and accident cause analysis.

3. Application of intelligent HAZOP analysis method based on data mining

Based on the above models and methods, this study constructed a HAZOP analysis data table containing 5503 entries of actual HAZOP analysis data collected from 14 petrochemical units for text data mining, providing technical support for intelligent risk prediction and automatic HAZOP analysis.

3.1. Overview of HAZOP analysis data table

The HAZOP analysis data table includes 13 elements already given in Section 2.1. Each piece of equipment should be classified according to the equipment classification in Table 1. The frequency of occurrence of each equipment classification can reveal the hidden dangers that the type of equipment may bring to the plants, which can be counted according to the HAZOP analysis data table. This study lists the equipment classifications and their corresponding occurrence times over 100, as shown in Fig. 3. There were seven equipment classifications whose corresponding occurrence times were more than 100 times.

The frequency of all words in the process parameter and guide word in the HAZOP data table can be calculated. The process parameters and guide words for each piece of equipment can be arranged from high to low according to the values of their word frequency. The sorted results provide the corresponding process parameters and guide words for each piece of equipment. The process parameters and guide words can constitute the deviations, and the sorted results will help analyzers consider deviations for all equipment and prevent omitting information. The word frequency of the process parameters and guide words for the equipment are shown in Figs. 4 and 5, respectively.

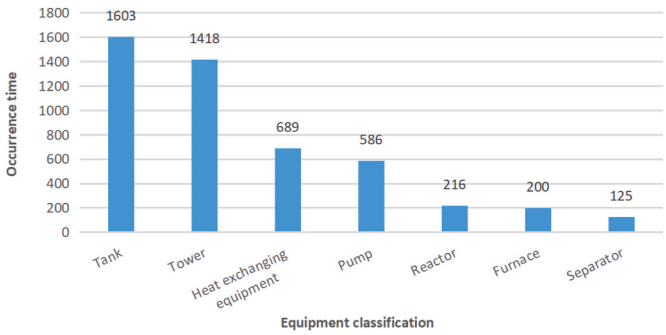


Fig. 3. Statistical results of equipment classification.

3.2. The result of text data mining based on HAZOP analysis data table

In this paper, the 'tank' with the largest occurrence times of equipment classification in the data table has been taken as an example to execute text data mining. Data mining methods, such as LDA clustering and the naive Bayes algorithm, reveal the correlation among elements, including cause, consequence, safeguard, and suggestion, in the HAZOP data table, realize intelligent risk prediction, and provide a reference for intelligent HAZOP analysis.

3.2.1. Cause and consequence word frequency statistical analysis

The word frequency statistics algorithm was used to calculate the word frequency for all words in the cause and consequence of the tank. It is suggested that defining at least 20 words with the highest word frequency as keywords improves the prominence of the analysis results in this study. The words were sorted according to their frequency, as shown in Table 2, which can be a reminder for analyzing the causes and consequences. According to the statistical results, the word frequency of 'failure' has the highest value (496 times) in the cause data of the 'tank' equipment, and the word frequency of 'loop' has 361 times. Therefore, analyzers can think and try to list certain reference causes containing the 'failure' and 'loop'. However, it is not sufficient for analyzers to completely determine the causes or consequences only by reminding them of one or two words.

3.2.2. Cause and consequence clustering results based on the LDA model

Perplexity is often used as an index to evaluate the merits and demerits of an LDA model. The smaller the perplexity, the better the modeling ability. The prediction effect of new texts and the number of topics are relatively optimal (Xu et al., 2022; Geeganage et al., 2021). The calculation formula is as follows.

$$\text{Perplexity(D)} = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{5}$$

In Equation (5), *Perplexity(D)* represents the perplexity. *D* refers to the HAZOP analysis data table and *M* represents pieces of the event description in the data table. N_d represents the number of words in each event description *d* and w_d denotes the words in the event description *d*. Here, $p(w_d)$ is the probability of word w_d generated in the event description *d*.

The classification numbers of different equipment are different, and their perplexities should be calculated for each equipment classification individually. The number of clustering topics of LDA for each equipment classification can be determined according to the perplexity. This study takes 'tank' as an example to explain the method for determining the optimal number of clustering topics for the tank. The number of clustering topics for causes and consequences should be moderate. Therefore, the value range of topic number *K* is suggested as [0,10] in this study, and the aforementioned formula is used to calculate the

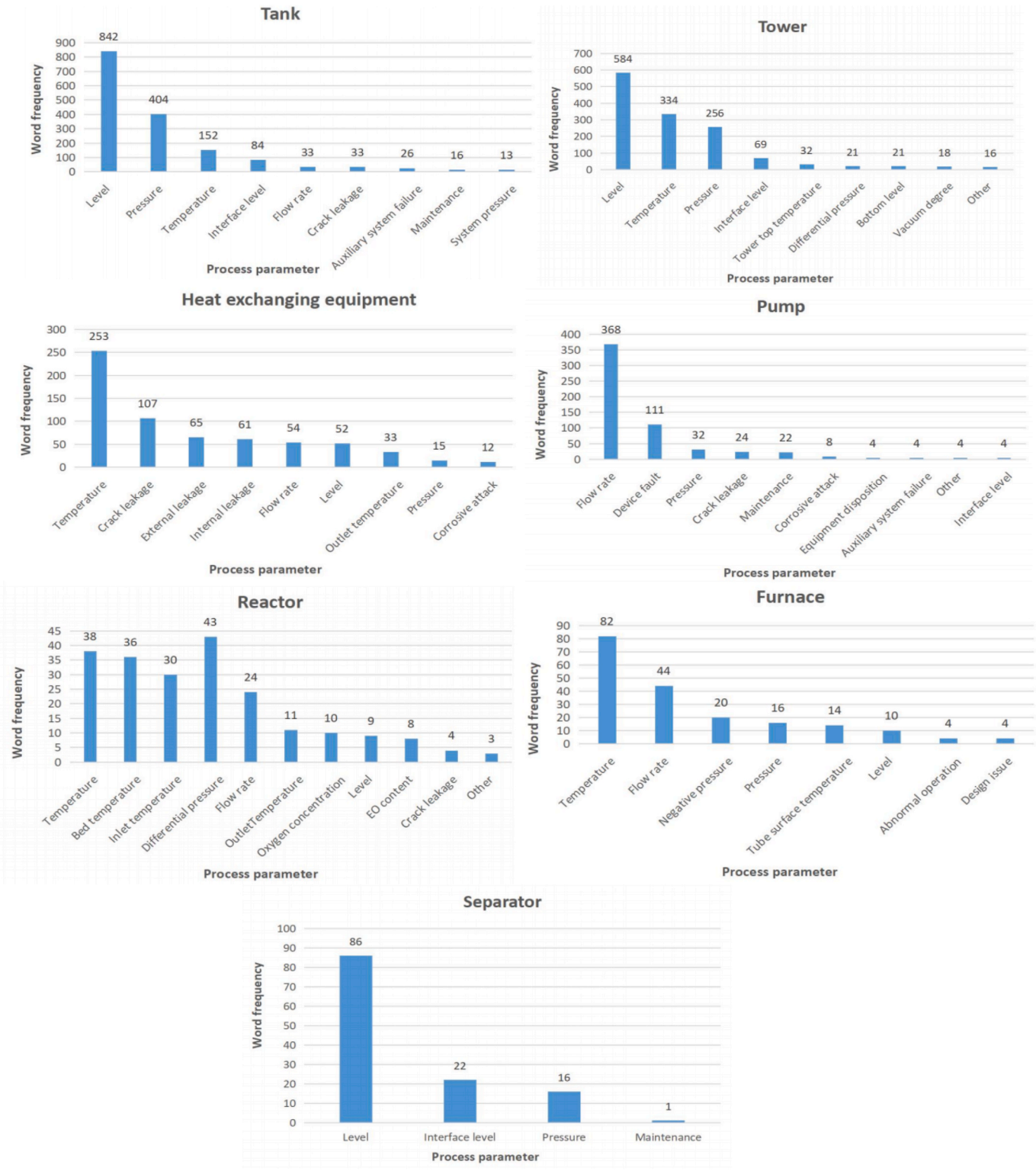


Fig. 4. Word frequency statistical results of process parameters.

perplexity corresponding to the number of topics, *K*. A curve representing the change in perplexity with the number of topics was drawn, as shown in Fig. 6. The curve shows that when the number of topics is seven, the perplexity curve is at the inflection point, the slope of the inflection point is evident, and the clustering effect is relatively optimal. Therefore, the number of clustering topics is selected to be seven.

This study employed the LDA model to cluster all the causes of the tank and count the proportion of the seven cause topics by performing the clustering analysis. All causes of the event descriptions of the 'tank' in the HAZOP data table can be classified into seven classes, and the information in each class can be determined as belonging to a cause topic. The word frequency for all the words in the cause of the 'tank' can be calculated. Defining the words with the ten highest word frequencies can represent one cause topic in this study. The ten words for each topic were sorted according to their word frequency, as shown in Table 3. For example, it can manifest equipment failure as loop failure, valve opening

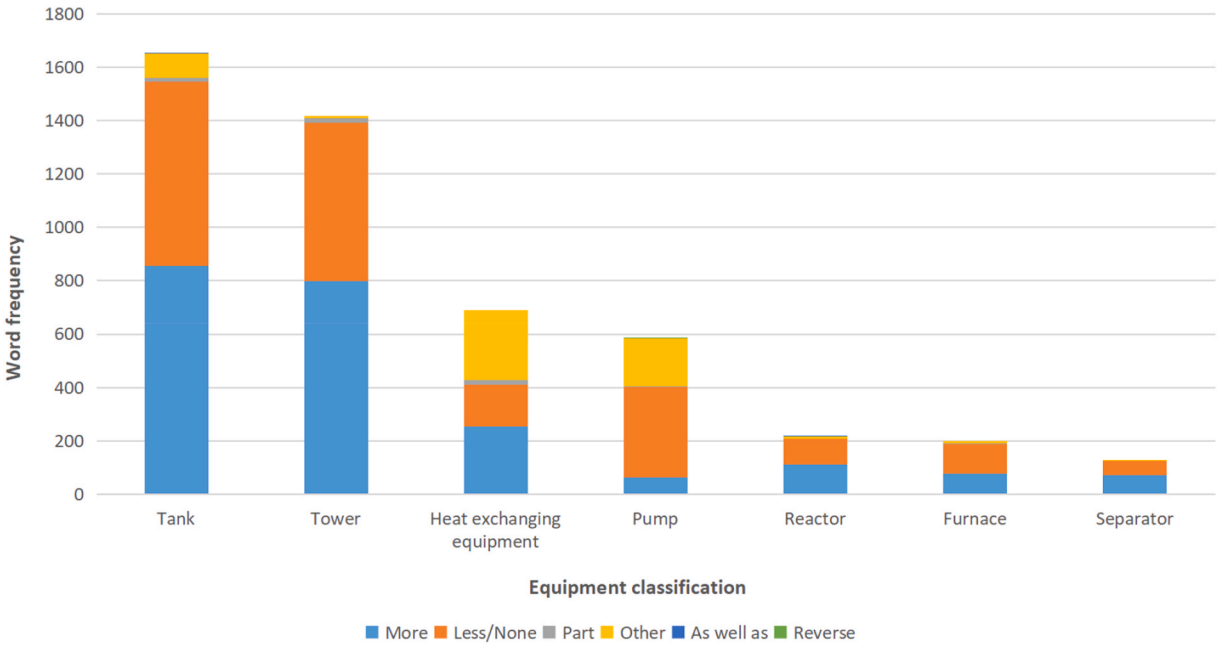


Fig. 5. Word frequency statistical results of guide words.

Table 2
Cause and consequence of 'tank' word frequency statistical results.

Element	Keywords and word frequency
Cause	(' Failure ', 496), (' Loop ', 361), (' Opening ', 273), (' False indication ', 192), (' Malfunction ', 175), (' Control ', 142), (' Incoming quantity ', 100), (' Off ', 99), (' Temperature ', 89), (' Pressure ', 84), (' Incoming material ', 79), (' Leakage ', 73), (' Internal leakage ', 73), (' Blockage ', 61), (' Valve ', 42), (' Control valve ', 41), (' Feed ', 41), (' Interruption ', 41), (' Acid ', 41), (' Steam ', 38)
Consequence	(' Elevated ', 846), (' Pressure ', 740), (' Level ', 601), (' Lowered ', 600), (' Temperature ', 343), (' Damaged ', 336), (' Affected ', 322), (' Equipment ', 237), (' Evacuated ', 212), (' System ', 191), (' Affected product quality ', 179), (' Operation ', 163), (' Descent ', 145), (' Interruption ', 130), (' Acid ', 125), (' Fluctuation ', 114), (' Overpressure ', 111), (' Full tank ', 111), (' Compressor ', 105), (' Feed ', 102)

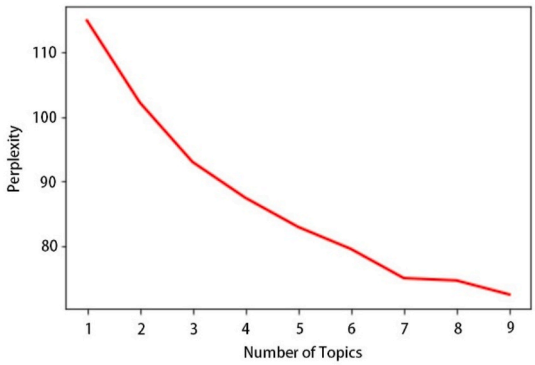


Fig. 6. Perplexity curve.

failure, etc., which can provide suggestions for HAZOP analysis of a new petrochemical plant. It can give cause topics certain sentences, clauses, or phrases to describe from the ten words representing the cause topic.

Similarly, this study used the LDA model to cluster all the consequences of the 'tank' and count the proportion of the seven consequence topics by calculating the clustering analysis. All consequences of the

Table 3
The clustering results of the 'tank' cause topics.

ID	Cause topic	Keywords	Probability
0	Equipment failure	Failure Loop; Opening; Off; False indication; Malfunction; Pressure; Control; Level; Leakage	0.25
1	Incoming anomaly	Incoming; Pressure; Failure; Control valve; Temperature; Content; Extraction; Medium water; Leakage; Tower top	0.16
2	Control failure	Loop; Malfunction; Control; Temperature Opening Failure; Valve Backflow; Pipeline; Hot water	0.14
3	Abnormal parameters	Temperature; Internal leakage; Failure; Incoming material; Interruption; Incoming quantity; Change; Components; Poor effect; Liquefied petroleum gas (LPG)	0.10
4	Valve/pipeline leakage	Leakage Failure; False indication; Opening; Circulate; Pressure; Acid; Flash; Flow; Flange	0.13
5	Instrumentation false indication	False indication; Failure; Incoming quantity; Flow; Interruption; Interlock; Valve; Malfunction; Acid; Opening	0.13
6	Interruption unit exception	Incoming quantity; Blockage; Malfunction; Loop; Control; Condenser; Internal leakage; Steam; Filter; Extraction tower	0.09
ALL			1

event descriptions of the tank in the HAZOP data table can be classified into seven classes, and the information in every class can be determined to belong to a consequence topic. The word frequency for all words in the consequences of the 'tank' was calculated. Defining the ten words with the ten highest word frequencies represents a consequence topic in this paper. The ten words for each topic were sorted according to their word frequency, as shown in Table 4. It can give consequence topics certain sentences, clauses, or phrases to describe from the ten words representing the consequence topic.

3.2.3. *HAZOP element correlation analysis*

A contingency table is typically used to judge whether there is a

Table 4
The clustering results of the 'tank' consequence topics.

ID	Consequence topic	Keywords	Probability
0	Affect product quality	Elevated; Pressure; Acid; Damaged; Affected; Quality of products; Temperature; Level; Extraction tower; Unqualified	0.15
1	System shut-down	Lowered; Pressure; Affected; Level; Elevated; System; Temperature; Shut-down; Quality of products; Operation	0.14
2	Equipment damage	Elevated; Pressure; Level; Affected; Damaged; Lowered; Temperature; Equipment; Evacuated; Desorber	0.14
3	System with water	Lowered; Pressure; Elevated; Temperature; Level; System; With water; Material; Unqualified; Compressor	0.19
4	Interrupt operation	Elevated; Level; Pressure; Operation; Lowered; Evacuated; Interruption; System Damaged; Affected	0.12
5	Equipment evacuating/ overpressure	Elevated; Damaged; Pressure; Level; Equipment Evacuated; Affected; Overpressure; Temperature; Lowered	0.18
6	Personnel poisoning	Elevated; Level; Pressure; Lowered; Temperature; Operator; Acid; Lowered; Poisoning Shut-down	0.09
ALL			1

correlation between the two features of the same analysis data. In this study, the correlation of the four elements, including process parameters, guide words, cause topics, and consequence topics, can be analyzed in pairs by the computing and applicating of the contingency table. The co-occurrence times of every pair of elements in the HAZOP data table can be acquired and listed in the contingency table, which can represent the correlation extent of every pair of elements. Contingency tables are shown in Figs. 7–10.

The correlation analysis results, including the process parameters and guide words, process parameters and each cause topic, process parameters, and each consequence topic, are shown in Fig. 7(a)–(c).

The correlation analysis results, including the guide words and each cause topic, guide words and each consequence topic, are shown in Fig. 8(a)–(b).

The correlation analysis results, including the deviations and each cause topic, deviation and each consequence topic, are shown in Fig. 9 (a)–(b).

The correlation analysis results for the cause and consequence topics are shown in Fig. 10.

The extent of the correlation between every two elements in the contingency table can be obtained. For example, in Fig. 10, it can be observed that the number of co-occurrence times for cause topic 0 and the consequence topics are 48, 86, 50, 64, 51, 79, and 29, and the total number is 407. These values, when sorted from high to low, provides the order for the consequence topics. The discussion order of the consequence topics for cause topic 0 is consequence topic1, 5, 3, 4, 2, 0, and 6. Therefore, the most likely consequence that the cause leading to will be the consequence topic 1. Similarly, it can be seen that the co-occurrence times for consequence topic 0 and the cause topics are 48, 26, 30, 48, 29, 38, and 17, and the total number is 236. These values can be used to sort the cause topic numbers. The discussion order of the cause topics for consequence topic 0 will be cause topic 0, 3, 5, 2, 4, 1, and 6. Therefore, the most likely cause leading to the consequence will because cause topic 0. The results will contribute to inferring the causes for the consequence and predicting the consequences of the cause. This can be beneficial for potential risk identification and accident investigation.

In the HAZOP analysis process, the word frequencies suggest the process parameters and corresponding guide words. The values of the extent of correlation can suggest the cause and consequence topics and their sequences of analysis for the corresponding process parameters and guide words.

Process parameter \ Guide word	Other	As well as	Less	More	All
Pressure	0	0	197	207	404
Flow rate	0	0	19	14	33
Level	0	0	368	474	842
Temperature	0	0	61	91	152
Interface level	0	0	36	48	84
Crack leakage	33	0	0	0	33
System pressure	0	0	7	6	13
Maintenance	16	0	0	0	16
Auxiliary system failure	25	1	0	0	26
All	74	1	688	840	1603

(a). Process parameters and guide words

Process parameter \ Cause topic	0	1	2	3	4	5	6	All
Pressure	111	45	67	54	46	48	33	404
Flow rate	17	1	3	1	3	2	6	33
Level	205	142	123	69	128	132	43	842
Temperature	45	29	12	17	17	7	25	152
Interface level	8	23	6	20	3	1	23	84
Crack leakage	13	0	6	3	0	10	1	33
System pressure	0	5	1	0	3	3	1	13
Maintenance	6	1	0	1	4	4	0	16
Auxiliary system failure	2	4	1	3	3	2	11	26
All	407	250	219	168	207	209	143	1603

(b). Process parameters and cause topics

Process parameter \ Consequence topic	0	1	2	3	4	5	6	All
Pressure	65	41	60	87	51	82	18	404
Flow rate	7	3	7	4	3	2	7	33
Level	110	132	111	181	89	136	83	842
Temperature	12	34	22	10	29	33	12	152
Interface level	13	2	10	7	15	22	15	84
Crack leakage	24	3	0	2	2	2	0	33
System pressure	2	2	0	5	2	0	2	13
Maintenance	1	9	2	2	0	1	1	16
Auxiliary system failure	2	5	5	4	4	4	2	26
All	236	231	217	302	195	282	140	1603

(c). Process parameters and consequence topics

Fig. 7. (a). Process parameters and guide words (b) Process parameters and cause topics (c) Process parameters and consequence topics.

Guide word / Cause topic	Other	As well as	Less	More	All
0	21	0	172	214	407
1	5	0	113	132	250
2	7	0	97	115	219
3	7	0	68	93	168
4	7	0	73	127	207
5	16	0	106	87	209
6	11	1	59	72	143
All	74	1	688	840	1603

(a). Guide words and cause topics

Guide word / Consequence topic	Other	As well as	Less	More	All
0	27	0	91	118	236
1	16	1	76	138	231
2	7	0	80	130	217
3	8	0	88	206	302
4	6	0	47	142	195
5	7	0	223	52	282
6	3	0	83	54	140
All	74	1	688	840	1603

(b). Guide words and consequence topics

Fig. 8. (a) Guide words and cause topics (b) Guide words and consequence topics.

3.2.4. Risk prediction based on the naive Bayes model

Seventy percent of the 'tank' analysis data were used as the training set and 25% as the test set. The risk prediction model is established by data mining the training set based on the naive Bayes algorithm and is validated using the test set to verify the prediction results of the influence of different element combinations on the risk prediction results. It is difficult to implement a machine learning method because the equipment name, equipment classification, process parameters, and guide words contain less text. The five elements and the cause are combined and determined as the cause characteristics, and the consequence element is considered as the consequence characteristics. TF-IDF method was used to perform the calculations. The values calculated from the cause characteristics, consequence characteristics and text combinations of the cause and consequence characteristics are taken as the input data. The values of the likelihood, severity, and risk level were taken as the output data. A naive Bayes classification prediction model was used to conduct machine learning. The prediction results are listed in Table 5. The calculation results showed that a comprehensive consideration of all cause and consequence characteristics in the HAZOP data table is the most important aspect for accident risk prediction. Most HAZOP analyses have adopted the same risk matrix. However, there are a few exceptions, in that the risk level may be a different value when the event description has the same value of likelihood and severity. Hence, risk matrix selection usually affects the model prediction accuracy to a certain extent.

When the HAZOP analyzers encounter a new process, the TF-IDF value of all cause and consequence characteristics in the analysis data is input into trained model 3, which can realize the intelligent prediction of accident likelihood, severity, and risk level and can be adjusted appropriately according to the actual production process to assist the construction of the risk matrix.

3.3. Getting the HAZOP analysis result table

(1) Establish an information inspection table for accident prevention in tank

Based on the word frequency statistical results of the process parameters, guide words, causes, and consequences, various equipment safety inspection items can be proposed from the aspects of parameters, equipment, and instruments. An information inspection table for accident prevention is constructed, as shown in Fig. 11.

(2) Build an intelligent HAZOP analysis table for the tank

An intelligent HAZOP analysis table can be constructed by integrating the LDA clustering results and HAZOP element association analysis results. The analyst needs to determine the classification of the specific equipment. The guide words and process parameters for the specific equipment can be obtained according to the statistical results of the historical data stored in the HAZOP data table, which is shown in the left table in Fig. 12. The causes and consequences can be obtained according to the results of the clustering and correlation analysis of the HAZOP data table, which is shown in the right table in Fig. 12. The guide words, process parameters, causes, and consequences shown in the interface can provide reference suggestions for the analysts. The analysts can determine the guide words, process parameters, causes, and consequences for the specific scenario analysis according to the reference suggestions and the expert experience.

3.4. Intelligent HAZOP analysis for the plant of wax oil hydrocracking unit

3.4.1. Process introduction

A hydrocracking plant is a typical plant in petrochemical processes. The raw oil buffer tank V2001 is the main equipment in the feedstock feeding process of the plant and can be analyzed as a node. The flowchart for this process is shown in Fig. 13. The hot raw oil from the atmospheric and vacuum equipment and the cold raw oil from the tank area are mixed and filtered outside the tank, and then the oil mixture enters the raw oil buffer tank V2001. The reaction material enters the following equipment and reacts in the presence of a catalyst in the hydrotreating reactor.

3.4.2. Intelligent HAZOP analysis results of the raw oil buffer tank

According to the international standard IEC61882, the node's potential risks, including the raw oil buffer tank V2001, need to be identified. The data mining results based on the HAZOP analysis data table will contribute toward completing the analysis table.

(1) Process parameter: It can be observed from Fig. 7 that the process parameters of the raw oil buffer tank V2001 are analyzed in the order of level, pressure, temperature, interface level, flow rate, crack leakage, auxiliary system failure, maintenance, and system pressure.
(2) Guide word: It can be noted from Fig. 7 that the analysis order of the guide words of the raw oil buffer tank V2001 is more, less/none, other, and as well as.

Deviation / Cause topic	Low pressure	High pressure	Low flow	High flow	Low level	High level	Low tem-perature	High tem-perature	Low interface level	High interface level	Crack leakage	Low system pressure	High system pressure	Improper mainten-ance	Auxiliary system failure	Additional effort to maintain	All
0	52	59	9	8	85	120	22	23	4	4	13	0	0	6	2	0	407
1	19	26	1	0	67	75	12	17	11	12	0	3	2	1	4	0	250
2	37	30	1	2	54	69	2	10	3	3	6	0	1	0	1	0	219
3	24	30	1	0	35	34	4	13	4	16	3	0	0	1	3	0	168
4	20	26	1	2	44	84	3	14	3	0	0	2	1	4	3	0	207
5	34	14	1	1	66	66	4	3	0	1	10	1	2	4	2	0	209
6	11	22	5	1	17	26	14	11	11	12	1	1	0	0	10	1	143
All	197	207	19	14	368	474	61	91	36	48	33	7	6	16	25	1	1603

(a). Deviations and cause topics

Deviation / Consequence topic	Low pressure	High pressure	Low flow	High flow	Low level	High level	Low tem-perature	High tem-perature	Low interface level	High interface level	Crack leakage	Low system pressure	High system pressure	Improper mainten-ance	Auxiliary system failure	Additional effort to maintain	All
0	41	24	4	3	33	77	1	11	10	3	24	2	0	1	2	0	236
1	21	20	2	1	35	97	15	19	1	1	3	2	0	9	4	1	231
2	28	32	5	2	44	67	3	19	0	10	0	0	0	2	5	0	217
3	13	74	0	4	73	108	1	9	0	7	2	1	4	2	4	0	302
4	12	39	1	2	31	58	3	26	0	15	2	0	2	0	4	0	195
5	70	12	2	0	99	37	33	0	19	3	2	0	0	1	4	0	282
6	12	6	5	2	53	30	5	7	6	9	0	2	0	1	2	0	140
All	197	207	19	14	368	474	61	91	36	48	33	7	6	16	25	1	1603

(b). Deviations and consequence topics

Fig. 9. (a) Deviations and cause topics (b) Deviations and consequence topics.

Consequence topic / Cause topic	0	1	2	3	4	5	6	All
0	48	86	50	64	51	79	29	407
1	26	28	27	75	30	32	32	250
2	30	26	28	36	30	43	26	219
3	48	12	23	30	20	19	16	168
4	29	35	41	39	20	29	14	207
5	38	27	22	39	21	49	13	209
6	17	17	26	19	23	31	10	143
All	236	231	217	302	195	282	140	1603

Fig. 10. Cause topics and consequence topics.

Table 5
Risk prediction accuracy.

ID	Attributes	Risk consequence	Accuracy
1	Cause characteristic	Likelihood	0.798
		Severity	0.735
		Risk level	0.761
2	Consequence characteristics	Likelihood	0.811
		Severity	0.853
		Risk level	0.824
3	Cause and consequence characteristics	Likelihood	0.832
		Severity	0.845
		Risk level	0.857

(3) Deviation: As shown in Fig. 7, the analysis order of deviation of the raw material oil buffer tank V2001 is high level, low level, high pressure, low pressure, high temperature, low temperature, high interface level, low interface level, crack leakage, auxiliary system failure, low flow, improper maintenance, high flow, low system pressure, high system pressure, and additional effort to maintain.
(4) Cause: Relationships among the process parameters, guide words, and causes can be acquired. In combination with the process flowchart, it is feasible to determine the corresponding cause topics applicable to the process and expand the specific cause analysis items based on the actual process flowchart. The results are shown in Fig. 14.
(5) Consequence: The correlation analysis between the tank cause and consequence topics can be obtained from Fig. 10. In combination with the process flowchart, it is feasible to determine the consequence topics suitable for this process and expand the specific consequence analysis items based on the actual process flow. Some of the results are shown in Fig. 14.
(6) Risk prediction: The text information of the equipment names, equipment classifications, process parameters, guide words, deviations, causes, and consequences should be converted into TF-IDF values and input into the risk prediction model established. The model can automatically calculate the likelihood, severity, and risk level of event descriptions. A portion of the prediction results is shown in Fig. 14.

According to the suggestions in the accident prevention information diagnostic form, safety management personnel should identify all

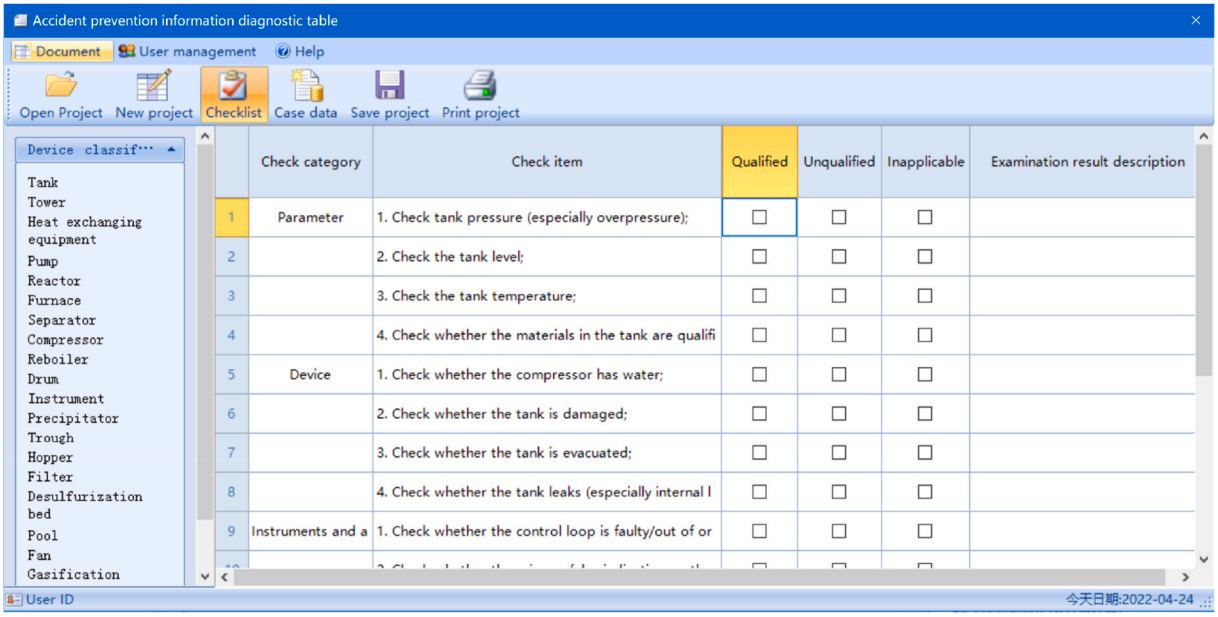


Fig. 11. Tank accident prevention information diagnostic form.

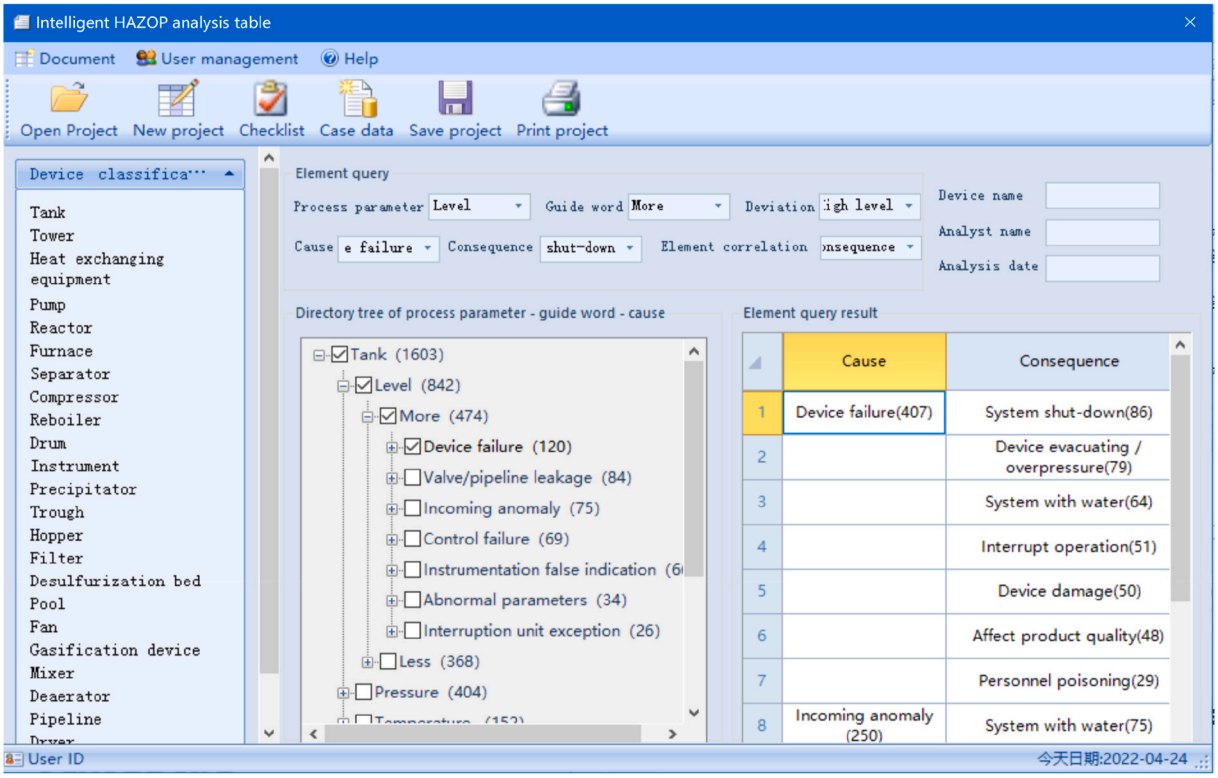


Fig. 12. Intelligent HAZOP analysis table of tank.

potential risks and record specific problems in the diagnosis process. By comparing the risk level of each analysis data in the intelligent HAZOP analysis table combined with the key attention factors of the tank, the key prevention and control objects in accident prevention can be defined to take effective measures to avoid the occurrence of accidents.

4. Conclusions

This study presents an intelligent HAZOP analysis method based on data mining. A HAZOP analysis data table was constructed, and a word frequency statistical algorithm was used to identify the key concerns of each type of equipment and develop an information inspection table. The LDA clustering models were employed to explore the topics of potential accident causes and consequences of 29 types of equipment individually. The correlation between HAZOP elements was confirmed, providing HAZOP analysis information and sequences for safety management personnel to prevent and control accidents in petrochemical production plants. Comprehensive analysis of HAZOP analysis data in

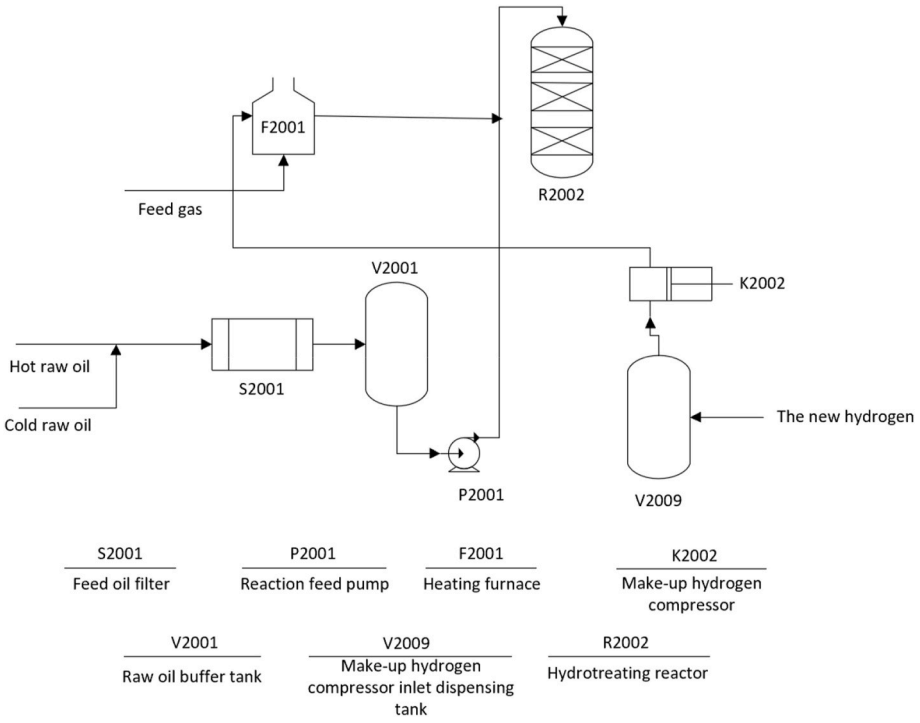


Fig. 13. Flowchart for feeding process of raw material oil.

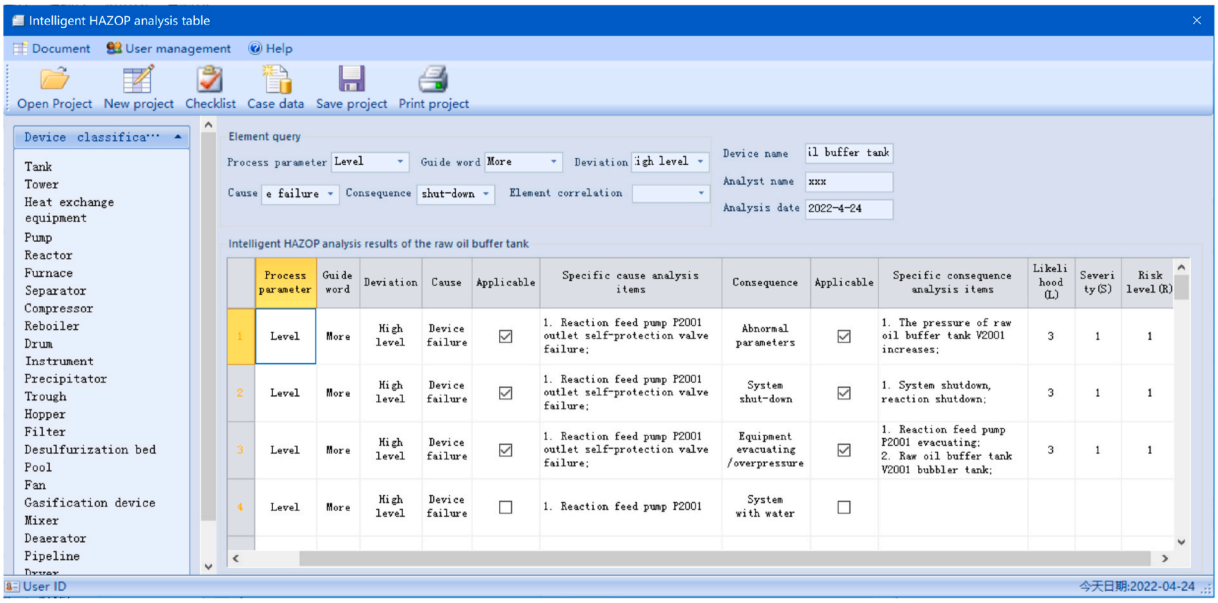


Fig. 14. Intelligent HAZOP analysis results of the raw oil buffer tank.

the table data, building the accident scene. Using the naive Bayes algorithm to calculate the likelihood, severity, and risk level of the accident scene, and the prediction accuracy of each value is above 80%. Which can verify the causation of the accident, and realize the intelligent prediction of the risk can help to solve the mutual independence between HAZOP reports, and comprehensively identify accident problems. The realization of intelligent HAZOP analysis can provide key monitoring objects, and prevention and control measures for safety management personnel to prevent and control potential accidents in petrochemical production plants and effectively avoid omissions in risk analysis, which is of great significance in improving the safety and reliability of petrochemical production processes.

Author contribution

Feng Wang: Conceptualization, Methodology, Data curation, Software, Supervision, Project administration. **Wunan Gu**: Investigation, Writing – original draft, Supervision, Writing- Reviewing and Editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. JD2217 and XK2020-04), high-performance computing platform of BUCT, and CNOOC Technical Cooperation Project.

References

Dorsey, L.T.C., Wang, B., Grabowski, M., Merrick, J., Harrald, J.R., 2020. Self healing databases for predictive risk analytics in safety-critical systems. J. Loss Prev. Process. Ind. 63, 104014 https://doi.org/10.1016/j.jlp.2019.104014.

Feng, X.Y., Dai, Y.Y., Ji, X., Zhou, L., Dang, Y.G., 2021. Application of natural language processing in HAZOP reports. Process Saf. Environ. Protect. 155, 41–48. https://doi.org/10.1016/j.psep.2021.09.001.

Geeganage, D.T.K., Xu, Y., Li, Y.F., 2021. Semantic-based topic representation using frequent semantic patterns. Knowl-Based Syst. 216, 106808 https://doi.org/10.1016/j.knosys.2021.106808.

Kościelny, J.M., Syfert, M., Fajdek, B., Kozak, A., 2017. The application of a graph of a process in HAZOP analysis in accident prevention system. J. Loss Prev. Process. Ind. 50, 55–66. https://doi.org/10.1016/j.jlp.2017.09.003.

Li, F.G., Zhang, B.K., Gao, D., 2021. Construction method of HAZOP knowledge graph. Chem. Ind. Eng. Prog. 40 (8), 4666–4677. https://doi.org/10.16085/j.issn.1000-6613.2020-2004.

Marhavilas, P.K., Filippidis, M., Koulinas, G.K., Koulouriotis, D.E., 2020. An expanded HAZOP-study with fuzzy-AHP (XPA-HAZOP technique): application in a sour crude-oil processing plant. Saf. Sci. 124, 104590 https://doi.org/10.1016/j.ssci.2019.104590.

Meng, Y.F., Song, X.M., Zhao, D.F., Liu, Q.L., 2021. Alarm management optimization in chemical installations based on adapted HAZOP reports. J. Loss Prev. Process. Ind. 72, 104578 https://doi.org/10.1016/j.jlp.2021.104578.

Nguyen, H.T., Safder, U., Kim, J., Heo, S.K., Yoo, C., 2022. An adaptive safety-risk mitigation plan at process-level for sustainable production in chemical industries: an integrated fuzzy-HAZOP-best-worst approach. J. Clean. Prod. 339, 130780 https://doi.org/10.1016/j.jclepro.2022.130780.

Niu, Y., Fan, Y.X., Gao, Y., 2019. Topic extraction on causes of chemical production accidents based on data mining. J. Safety Sci. Technol. 15 (10), 165–170, 10. 11731/j. issn. 1673-193x. 2019. 10. 026.

Single, J.I., Schmidt, J., Denecke, J., 2020. Ontology-based computer aid for the automation of HAZOP studies. J. Loss Prev. Process. Ind. 68, 104321 https://doi.org/10.1016/j.jlp.2020.104321.

Suh, Y., 2021. Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database. Saf. Sci. 142, 105363 https://doi.org/10.1016/j.ssci.2021.105363.

Suzuki, T., Izato, Y., Miyake, A., 2021. Identification of accident scenarios caused by internal factors using HAZOP to assess an organic hydride hydrogen refueling station involving methylcyclohexane. J. Loss Prev. Process. Ind. 71, 104479 https://doi.org/10.1016/j.jlp.2021.104479.

Wang, F., Gao, J.J., Zhang, B.K., Zhang, X., 2008. Computer aided HAZOP analysis technology based on AHP. Chem. Ind. Eng. Prog. 27 (12), 2013–2018, 1000–6613 (2008) 12–2013–06.

Wang, H.S., Zhao, D.F., Liu, Y., 2013. The design of HAZOP expert system framework based on the real-time database. J. Safety Sci. Technol. 9 (4), 82–86, 10. 11731/j. issn. 1673-193x. 2013. 04. 015.

Wang, Y.B., Xu, W., 2018. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis. Support Syst. 105, 87–95. https://doi.org/10.1016/j.dss.2017.11.001.

Xu, H., Liu, Y., Shu, C.M., Bai, M.Q., Motalifu, M., He, Z.X., Wu, S.C., Zhou, P.G., Li, B., 2022. Cause analysis of hot work accidents based on text mining and deep learning. J. Loss Prev. Process. Ind. 76, 104747 https://doi.org/10.1016/j.jlp.2022.104747.

Zhong, B.T., Pan, X., Love, P.E.D., Sun, J., Tao, C.J., 2020. Hazard analysis: a deep learning and text mining framework for accident prevention. Adv. Eng. Inf. 46, 101152 https://doi.org/10.1016/j.aei.2020.101152.

Zhou, G.W., Yang, X., Zheng, S.Q., 2018. Research progress of intelligent HAZOP analysis system. Chem. Ind. Eng. Prog. 37 (3), 815–821. https://doi.org/10.16085/j.issn.1000-6613.2017-1061.

Zhu, R.C., Hu, X.F., Hou, J.Q., Li, X., 2021. Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Saf. Environ. Protect. 145, 293–302. https://doi.org/10.1016/j.psep.2020.08.006.