Identifying the Causes of Ship Collisions Accident Using Text Mining and Bayesian Networks

Jianguo Yu[1], Zhihua Wu[2], Wei Liu[3,*], Wenji Zhao[1]
[1]*Jiangxi Ganbei Waterway Affairs Center, Jiujiang, China*
[2]*Jiangxi Huitong Technology Development Co., Ltd, Nanchang, China*
[3]*School of Transportation Engineering, East China Jiaotong University, Nanchang, China*
927696194@qq.com; 2838626625@qq.com; 2022138086100078@ecjtu.edu.cn; 382500168@qq.com

***Abstract*—Under the backdrop of the robust growth of the global economy, the water transport industry is experiencing rapid development, resulting in an increase in ship collisions and a critical water traffic safety situation. This study uses text mining techniques to gather a corpus of data. The corpus includes human factors, ship factors, natural environmental factors, and management factors, which are used as target data to obtain a high-dimensional sparse original feature vector space set comprising eigenvalues and eigenvalue weight attributes. Chi-square statistics are utilised to reduce dimensionality, resulting in a final set of 33-dimensional text feature items that determine the causal factors of ship collision risk. Taking the four steps involved in the collision process as the primary focus, a Bayesian network structure for ship collision risk is constructed based on the "human-ship-environment-management" system. By incorporating existing ship collision accident/danger reports, conditional probability tables are computed for each node in the Bayesian network structure, enabling the modelling of ship collision risk. The model is validated through an example, revealing that, under relevant conditions, the probability of collision exceeds 90 %. This finding demonstrates the validity of the model and allows one to deduce the primary cause of ship collision accidents.**

***Index Terms*—Maritime safety; Text mining; Bayesian network; Causal factors; Ship collision risk.**

I. INTRODUCTION

Against the backdrop of a flourishing global economy and its increasing influence, the water transport industry in various countries is also making rapid progress [1]. However, this development has been accompanied by frequent water traffic accidents. Although relevant management departments have made concerted efforts to reduce the number of water traffic accidents, deaths and disappearances, shipwrecks, direct economic losses, and other four statistical indicators over recent years, the current water traffic safety situation still poses significant risks [2], [3]. This is due to the unpredictable nature of water traffic accidents and their potential adverse impact on society. Therefore, it is imperative for the water traffic management department and the entire water transportation industry to prioritise measures that enhance water traffic safety.

Second, improving water transportation safety can not only rely on improving infrastructure and improving ship reliability, but also needs to systematically analyse water transportation safety from human factors, machines, environment, management, and other aspects. Many scholars have studied water traffic safety from various angles and have drawn relevant conclusions on the causes and prevention methods of water traffic safety. Haapasaari, Helle, Lehikoinen, Lappalainen, and Kuikka [4] provided a comprehensive overview of the formulation of maritime safety policies in the Gulf of Finland. The aim was to establish a more suitable water traffic safety guarantee method for the Gulf of Finland based on risk identification, assessment, and shipping policies or management plans before accidents occur. The application of cutting-edge technology in maritime affairs is also helpful for studying maritime risks. Zaman [5] studied the collision risks of ships in the Strait of Malacca by extracting ship navigation data from AIS and establishing a navigation simulation model for ships in the region. Actual data were used to study the laws and causes of ship collisions, and a ship collision avoidance method suitable for the Malacca Sea area was proposed. Dominguez [6] proposed new ideas to improve the level of navigation safety and reduce operating costs by using innovative technologies such as mobile phone applications, cloud computing, and big data in maritime safety. However, in the field of water traffic accident investigation report, most of the authors analyse the standardisation of its writing and discuss a certain accident itself, but lack systematic mining and analysis.

Text mining is a branch of data mining and covers a number of research areas [7]. It is well-known that data mining can mine the potential knowledge that seems to be scarce in the explosive mass data. When the data are mined in the form of text, it can be called text mining. Of course, this process also requires the use of machine learning, information processing, pattern recognition, database and computer linguistics, and other disciplines of theory and methods [8], [9], [10]. Gupta and Lehal [11] conducted an analysis of the commonly used techniques and application

Manuscript received 6 August, 2023; accepted 21 October, 2023.

fields of text mining. Methods utilised in text mining include information extraction, topic tracking, information overview, text classification, text clustering, association analysis, information visualisation, and more. Text mining technology is mainly applied in various areas such as publishing media, telecommunications, energy, and other services, information and Internet industry, banking, insurance and financial industries, public management and policy-making, pharmaceutical research and development departments, among others. Kanya and Geetha [12] extracted information from the text and carried out an in-depth mining to discover potential rules. Using the discovered rules, they also predicted missing information from existing data. Montalvo, Martínez, Casillas, and Fresno [13] calculated the similarity of cross-lingual names, geographical names, and institutional names to perform multilingual text clustering. Wei, Yang, and Lin [14] proposed a method based on latent semantic indexing to address the language barrier problem. They constructed a multilingual indexing system using parallel corpora through the latent semantic indexing technique. The multilingual text is transformed into a latent language index space, enabling the development of a multilingual text clustering method. Anaya-Sánchez, Pons-Porrata, and Berlanga-Llavori [15] introduced a text clustering algorithm that focusses on discovering and describing topics present in a set of texts. Isa, Kallimani, and Lee [16] utilised the advantages of self-organising maps to overcome the dimension reduction caused by relying solely on Bayesian formulas. When improved ranking technology was combined with a hybrid system, the efficiency of the text classification method was further improved. Text mining can solve the processing and analysis of unstructured data, such as text files, and discover new knowledge and valuable potential information. It is currently becoming a new research hotspot and is used extensively in all walks of life.

Text mining technology enables targeted qualitative and in-depth research to be conducted. As a nascent Internet technology, text mining is only beginning to find its application in the field of transportation. Gao and Wu [17] developed a verb-based text mining method to extract the causes and outcomes of Missouri automobile traffic accidents from Web-based traffic accident reports. This approach not only facilitates the classification of traffic accidents, but also enables deeper insight into their actual causes. Nayak, Piyatrapoomi, and Weligamage [18] used big data and text mining techniques to research the prevalent issues in the construction of urban roads. By analysing accident data reports, they also highlighted the significant role of traffic facilities optimisation and route planning in reducing urban traffic accidents. Septiana, Setiowati, and Fariza [19] captured traffic-related instant messages from social networks and displayed real-time data on terminal devices. The real-time road traffic monitoring system, using data from social networks, demonstrated an accuracy rate of up to 92 % in tests, providing useful guidance to travellers. There have been limited studies on text mining for traffic accident analysis. This can be attributed to the scarcity of extensive accident text data available for research purposes. Consequently, it has become crucial to address the challenge of extracting valuable data from accident investigation reports, mitigating data loss resulting from small accident samples and identifying potential correlations among accidents.

In view of the existing water traffic accident investigation report, through text mining technology, we can analyse the potential causes of the corresponding water traffic accidents, classify them systematically, construct a suitable water traffic accident database, explore the valuable information, find new laws and accident occurrence mechanism, establish a water traffic accident decision-making model and predict the possible risks in the future in advance, to help maritime managers scientifically strengthen water traffic supervision and escort water traffic safety.

II. METHODS

A. Text Mining

Text mining refers to the utilisation of computer technology or big data technology to extract valuable information and knowledge from textual data. This technology finds application in various fields of knowledge, including statistics, natural language processing, and machine learning, thereby offering technical support for data processing and mining within the realm of big data. The specific text mining process [20] is shown in Fig. 1.

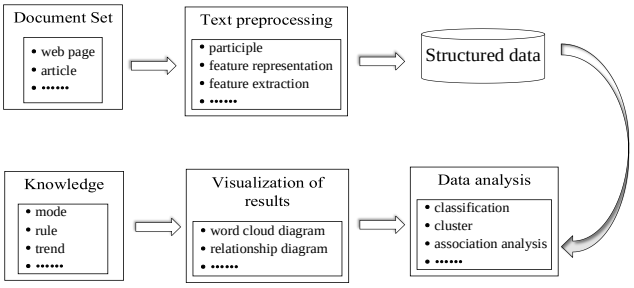

Fig. 1. Text mining process.

During the text mining process, the system will segment the text and generate a series of feature items. The space vector is composed of the feature terms and the weight terms assigned by the system to the feature term. The definition of the space vector is shown below [21].

Definition 1: In text mining, a document or a paragraph is represented by D, and all documents are recorded as N.

Definition 2: In text mining, the feature items refer to the basic language units that can represent the text material, such as characters, words, etc., represented by T_k.

Definition 3: In text mining, the feature item weight W_{ik} represents the importance D_i of the feature item to the text weight, which is calculated by $tf-idf$ as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{1}$$

where $n_{i,j}$ refers to the number of times the term appears in the text file D_i and $\sum_k n_{k,j}$ is the total number of occurrences in all entries in all files;

$$idf = \log \frac{|D|}{\left|\left\{ j : t_i \in d_j \right\}\right|}, \tag{2}$$

where $|D|$ is the total number of text files and $\left|\left\{ j : t_i \in d_j \right\}\right|$ is the file entry that contains entry statements;

$$tf - idf = tf_{i,j} \times idf_i. \tag{3}$$

Definition 4: In text mining, the constructed vector space model refers to treating feature items as high-dimensional coordinate systems, representing weights as values of high-dimensional coordinates, and the calculated vector set is the vector space model of the text.

In addition, an excessively high dimension can adversely impact the calculation speed and may lack practical significance. Therefore, it becomes necessary to reduce the dimensionality and optimise the feature items pertaining to the text of inland river ship collision accident reports. Among various dimensionality reduction methods, statistics yield better recall and precision rates. As a result, statistics are employed to reduce the dimensionality of inland ship collision accident data

$$c^2\left(t, c_i\right) = \frac{n\left(ab - cd\right)^2}{\left(a+c\right)\left(b+d\right)\left(a+b\right)\left(c+d\right)}, \tag{4}$$

where *n* denotes the number of the entire text, *a* denotes the frequency of occurrence of the text belonging to class c_i with feature *t*, *b* represents the frequency of the text that does not belong to the c_i class containing the feature item *t*, *c* denotes the frequency of occurrence of the text belonging to class c_i without feature *t*, and *d* denotes the frequency of occurrence of text that does not belong to c_i class and does not contain feature item *t*.

Then the χ^2 value of the whole text corpus is

$$\chi^2{}_{\max}\left(t\right) = \max_{i=1}^{m}\left\{\chi^2\left(t, c_i\right)\right\}. \tag{5}$$

Then, most of the meaningless text information feature items are removed by the value of *m*, so as to achieve the purpose of dimensionality reduction.

B. Bayesian Network

The Bayesian network (BN) is a probabilistic network constructed based on the Bayesian theorem [22]. It is a directed acyclic graph (DAG). The nodes and the directed edges are the two main elements of the BN, in which the nodes are variables on the network, and the nodes are connected probabilistically through the directed edges. The directed edge connects the parent node and the child node, and the direction is directed from the parent node to the child node. The strength of the relationship between each other is expressed according to the corresponding probability. A typical BN is shown in Fig. 2.

Suppose there exists a set of random variables *V*, which contains *n* variables. DAG is represented by *G*, *D* is the set of directed edges, and *C* is the set of conditional probability distributions. Then a BN model can be expressed as

$$BN = \left(G, C\right) = \left(V, D, C\right), \tag{6}$$

where:

$$G = \left(V, D\right), \tag{7}$$

$$V = \left\{V_1, V_2, ..., V_n\right\}, \tag{8}$$

$$D = \left\{\left(V_i - V_j\right) \middle| V_i, V_j \in V\right\}, \tag{9}$$

$$C = \left\{P\left(V_i \middle| V_{i-1}, V_{i-2}, ..., V_1\right), V_i \in V\right\}. \tag{10}$$

If Pa_i is the parent node of V_i, the corresponding joint probability distribution is

$$P\left(V_1, V_2, ..., V_n\right) = \prod P\left(V_i \middle| Pa_i\right). \tag{11}$$

Bayesian network learning primarily involves acquiring network knowledge through the analysis of existing data [23]. It can be categorised into structure learning and parameter learning. Structure learning aims to determine both the network structure and the associated parameters, whereas parameter learning focusses on estimating the parameters of the network given a known structure.

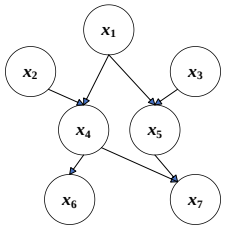

Fig. 2. An example of BN.

The process of learning the BN structure involves establishing a graphical model. Initially, the causal relationship between samples is analysed, followed by the creation of a network structure in a DAG format. During this process, it is essential to understand the degree of correlation between nodes. Knowledge and previous experience of industry experts can be utilised to judge these correlations and determine relationships between nodes, among other factors.

The principle underlying BN structure learning is founded on the Bayesian theorem, with the goal of identifying a network that most closely approximates the sample data available. Assuming S^h is a random variable, it is defined as the relationship corresponding to the BN structure, where $p\left(S^h\right)$ represents the associated prior probability distribution, and the calculated posterior distribution is denoted by $p\left(S^h \middle| D\right)$

$$p\left(S^h \middle| D\right) = \frac{p\left(S^h, D\right)}{p\left(D\right)} = p\left(S^h\right)\frac{p\left(D \middle| S^h\right)}{p\left(D\right)}, \tag{12}$$

where $p\left(D\right)$ represents a normalised constant independent of the network structure and $p\left(D \middle| S^h\right)$ is the boundary likelihood distribution. Obviously, to get the posterior probability, we only need to get the entire boundary likelihood distribution. In general, the boundary likelihood distribution of the sample is the product of all (i, j).

BNs can be modelled in three main ways. The first involves creating a knowledge base, while the second involves relying

on data learning. The third method involves utilising expert knowledge for modelling purposes. In some cases, combinations of these methods are employed to improve the accuracy of the model. When modelling through data learning alone, it is essential to ensure that the structure of the corresponding model is well-formed.

BN specific learning methods can be classified into dependency test learning and search score-based learning. Dependency test learning involves evaluating the independence relationship between nodes based on a given sample set S to obtain the structure of the network. On the other hand, search score-based learning establishes the network structure by defining a unique search strategy and scoring criteria within the structural space of variables. Both methods have their own advantages. Dependency test learning is useful when the sample space is determined and can provide satisfactory results. Search score-based learning offers flexibility in defining search strategies and scoring criteria, allowing for more customised modelling. Ultimately, the choice between these methods depends on factors such as the size of the sample space and the computational resources available.

III. CASE STUDY

A. Data Sources

Data used for text mining analysis in this paper are sourced from the 2013–2017 water traffic accident danger report provided by the Yangtze River Shipping Bureau. Specifically, the analysis focusses on ship collision accidents that occurred in the lower reaches of the Yangtze River during this period. The dataset comprises a total of 419 accidents, involving 518 ships that were part of these collision incidents.

The water traffic accident danger report used records the location of the accident, the danger situation, the characteristics of the accident ship, and the weather and hydrological information at the time of the accident. The statistical items of the danger record are shown in Table I.

TABLE I. STATISTICAL PROJECT ON CHARACTERISTICS OF INLAND SHIP COLLISION ACCIDENTS.

Incident location			Dangerous situation		
Incident area	Incident mileage		Nature of risk	Risk level	
Incident time					
Year	Month		Day	Minute	
Meteorological and hydrological information					
Wind power	Wind direction	State of sea	Visibility	Temperature	Water temperature
Characteristics of accident ships					
Ship name	Wail		Nationality	Port of registry	
Ship type	Vessel age		Gross tonnage	Deadweight ton/rated passenger capacity	
Power	Captain		Breadth	Draft (front/rear)	
Number of crew members	Seating capacity		Ship owner	Nature of ship owner	
Name of goods/carrying weight	Port of departure		Port of destination	Accident number	
Human or management factors					
Preparation before sailing	Navigation alert		Collision avoidance decision-making	Manipulation execution	

B. Results and Analysis

The report on the ship collision accident primarily comprises navigation data and accident data, with the former encompassing semi-structured information within the navigation environment data. On the other hand, the accident danger data table contains certain unstructured information. During the analysis of inland ship collision accidents, it is essential to take into account human factors, ship factors, navigation environment factors, and management factors. The following section delves into the analysis of data sources and target data for text mining.

1. Navigation environment data

Navigable environment data defined by the analysis of data requirements in the database, including visibility, wind, flow, information about water level, meteorological information, maintenance depth of water, and port water regime in the ship collision accident report.

2. Ship data

The data source of unstructured data in ship data is ship collision accident report, which describes the ship collision accidents in free text, mainly including ship parameter information such as ship equipment failure or failure, ship type, ship tonnage, ship grade, etc.

3. Human factor data

The data source of unstructured data in human factor data is the ship collision accident report, which describes ship collision accidents in free text, mainly including information about human decision-making parameters such as negligence, lack of experience, failure to rest in time, and error level.

4. Management factor data

The data source of the unstructured data in the management factor data is the ship collision accident report, which describes the ship collision accident in the form of free text, mainly including the management error information such as the failure of the crew, the improper arrangement of the succession on duty, and the improper cooperation of the crew.

In this paper, we analyse the data source and target data of text mining from various factors, including human factors, ship factors, navigation environment factors, and management factors. Subsequently, the 382 reports of inland ship collision accidents are organised in text format and processed using a word segmentation programme. To enhance the accuracy of text mining and avoid the omission of professional terms and misidentification of function words, this section aims to formulate and summarise the types of vocabulary prior to word segmentation, as well as to eliminate function words within the text.

The glossary of professional terms contains relevant vocabulary from various fields, such as water traffic engineering, safety engineering, shipping, meteorology, and more. To eliminate function words, the “Modern Chinese Function Word Dictionary” is utilised. The word segmentation outcome yields a high-dimensional sparse set as the original feature item, as evidenced in Table II. Given that the ship collision accident report is the focus of text mining, words such as “collision”, “responsibility”, and “accident” frequently appear in the text. Although these words allow us to determine the attributes of the text, they do not significantly help to analyse the root cause of the accident. Therefore, these words can be disregarded during the mining process.

TABLE II. CHARACTERISTIC ITEMS IN THE TEXT SET OF THE INVESTIGATION REPORT ON INLAND SHIP COLLISION ACCIDENTS (PARTIAL).

Number	1	2	3
Characteristic	Collision	Accident	Responsibility
Number	4	5	6
Characteristic	Risk situation	Resort	Improve
Number	7	8	9
Characteristic	Shipping	Trends	Sail
Number	10	11	12
Characteristic	Violation of regulations	Inexperienced	Mis-operation
Number	13	14	15
Characteristic	Poor technical level	Crew	Inexperienced
Number	16	17	18
Characteristic	Management confusion	Organisation	Insufficient management
Number	19	20	21
Characteristic	Meet an emergency	Fault	Hydrology
Number	22	23	24
Characteristic	Meteorological phenomena	Artificial	Environment
Number	25	26	27
Characteristic	Information communication	Insufficient preventive measures	Insufficient safety checks
Number	28	29	30
Characteristic	Violation	Ship defects	Etc.

C. Data Processing

Following the fundamental text mining process, the preprocessing of text data is performed, resulting in 498 original features. To reduce dimensionality and streamline the text mining process, the statistical function is used for dimension reduction analysis on the feature items. As a result, four factors are identified that influence ship collisions: human factors, ship factors, environmental factors, and management factors. The text feature items are reduced to 33 dimensions, and the risk factors associated with risks of ship collision are determined, as illustrated in Table III.

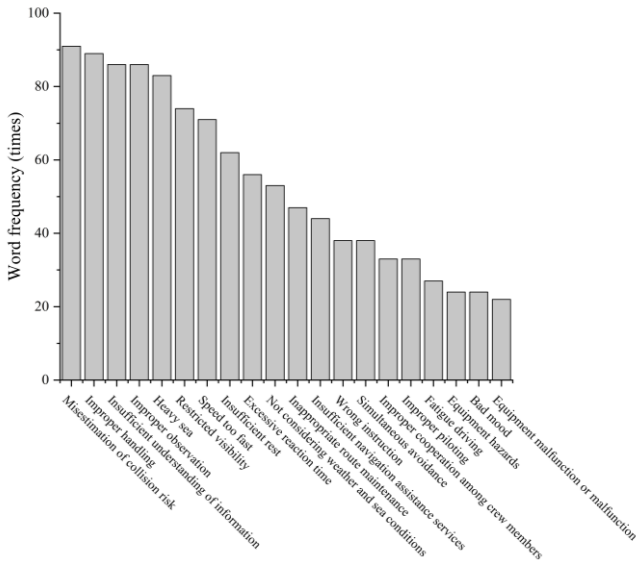

Fig. 3. Text feature item dimensionality reduction results.

Through Fig. 4, 382 accident reports that occurred at different locations and times can be correlated. The importance of the causal factor can be determined by the font size, colour, and position of the 20 factors in the word cloud graph. The larger the font, darker the colour, and closer to the centre, the more important the causal factor is.

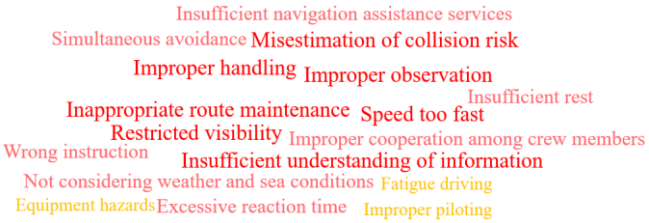

Fig. 4. Word cloud of accident-causing factors.

TABLE III. FACTORS CONTRIBUTING TO THE RISK OF COLLISION BETWEEN INLAND VESSELS.

Number	1	2	3
Characteristic	Inappropriate route maintenance	Not considering weather and sea conditions	Insufficient rest
Number	4	5	6
Characteristic	Bad mood	Improper observation	Improper piloting
Number	7	8	9
Characteristic	Inappropriate route maintenance	Fatigue driving	Insufficient understanding of information
Number	10	11	12
Characteristic	Misestimation of collision risk	Excessive reaction time	Improper handling
Number	13	14	15
Characteristic	Simultaneous avoidance	Adventure overtaking	Speed too fast
Number	16	17	18
Characteristic	Equipment hazards	Incomplete navigation information	Equipment malfunction or malfunction
Number	19	20	21
Characteristic	Host malfunction or malfunction	Auxiliary equipment malfunction or malfunction	Communication system malfunction or malfunction
Number	22	23	24
Characteristic	Navigation system malfunction or malfunction	Improper display of signal lights, light types, or signals	Heavy sea
Number	25	26	27
Characteristic	Restricted visibility	Navigation aids malfunction or malfunction	Insufficient navigation assistance services
Number	28	29	30
Characteristic	Wrong instruction	Improper cooperation among crew members	Inappropriate arrangements for taking over duty
Number	31	32	33
Characteristic	No misconduct found	No Incompetence of crew members found	No equipment hazards found

Based on the text of the inland ship collision accident report and the accident features after dimensionality reduction, the word frequency of the feature items after dimensionality reduction was analysed, and the analysis results are shown in Fig. 3.

According to the above arrangement of the word cloud of inland ship collision accidents, it can be described as the main cause factors and other cause factors. The main cause factors include mainly seven items: wrong estimation of collision risk, improper handling, insufficient understanding of

information, improper observation, strong wind, poor visibility, and too fast speed. Although the above seven main cause factors constitute most accidents, inland river ship collision accidents are not caused by single consistent cause factors, so other cause factors should also be considered. To effectively prevent and control the occurrence of accidents, it is necessary to formulate prevention and control measures from the four factors of human-ship-environment-management.

IV. BAYESIAN NETWORK MODELLING

A. Node Analysis

In line with the principles of Bayesian modelling and considering the content outlined in Section III of this paper, four factors that significantly affect ship collisions are determined: human factors, ship factors, environmental factors, and management factors. Concurrently, based on analyses of the causes and principles behind ship collision accidents, the ship collision process is divided into four distinct stages: prenavigation preparation, navigation alert, collision avoidance decision-making, and manipulation execution. Through extensive literature review and expert research, the causes of ship collision risks are identified and presented in Table IV.

Some important factors and related classifications are described below.

1. Human factors: Human factors play a key role in the transportation system; water traffic is no exception. According to relevant statistics, more than 80 % of maritime accidents are related to people. Human factors in this paper mainly represent the part of the collision avoidance system of the ship composed of individuals and the overall behaviour. Except for management personnel (such as service personnel), the relevant personnel on the shore are not included in the scope of the study.

Human factors in the prenavigation preparation stage are mainly closely related to management factors, such as improper route selection and insufficient rest. The main risks in the navigation alert stage are improper pilotage, lack of alertness, and fatigue driving, while in the collision avoidance decision-making stage, insufficient understanding of information means that the ship driver does not make correct and timely decision-making response based on sufficient information.

2. Ship factor: The ship serves as the objective subject in the collision process. Maintaining optimal ship performance is crucial to reducing the likelihood of collisions.

The evaluation criteria for the factor of "Absence of light, incorrect light type, or improper signal display" is based on the relevant provisions outlined in the fourth chapter of the "Collision Avoidance Rules", which specifically covers topics such as light types, sound signals, light signals, and appendices. Correct display of lamp types significantly influences the ability to assess the encounter situation and avoid collisions in a timely manner. Other equipment failures are relatively self-explanatory and will not be reiterated here.

3. Environmental factors: Environmental factors exert a substantial influence on ship collisions. In particular, adverse weather conditions significantly increase the likelihood of such incidents.

4. Management factors: In the literature on ship collisions, management factors are frequently overlooked, which is scientifically unsound. Although management factors exhibit a strong correlation with human factors, they should not be directly equated. Furthermore, corporate culture, the overall competence of employees, and the proficiency of senior managers constitute pivotal elements of management factors.

TABLE IV. CAUSAL FACTORS OF SHIP COLLISION RISK.

	Preparation before sailing	Navigation alert	Collision avoidance decision-making	Manipulation execution
Human factor	Improper route selection H11 Not considering weather and sea conditions H12 Insufficient rest H13 Bad mood H14	Improper observation H21 Improper piloting H22 Inappropriate route maintenance H23 Fatigue driving H24	Insufficient understanding of information H31 Misestimation of collision risk H32 Excessive reaction time H33	Improper handling H41 Simultaneous avoidance H42 Adventure overtaking H43 Speed too fast H44
Ship factors	Equipment hazards S11 Incomplete navigation information S12	Equipment malfunction or malfunction S21/Host malfunction or malfunction S22/Auxiliary equipment malfunction or malfunction S23/Communication system malfunction or malfunction S24/Navigation system malfunction or malfunction S25/Improper display of signal lights, light types, or signals S26		
Environmental factor	Heavy sea E11 Restricted visibility E12 Navigation aids malfunction or malfunction E13 Insufficient navigation assistance services E14			
Management factors	No incompetence of crew members found M11 No equipment hazards found M12	Inappropriate arrangements for taking over duty M21 No misconduct found M22	Wrong instruction M31 Improper cooperation among crew members M32	

B. Structure Analysis

1. Bayesian network structure of human factors

Human factors are an important part of the entire Bayesian network. Among them, with "inadequate rest H13" as the main incentive, combined with the existing literature, a systematic and comprehensive analysis of human factors is performed, and a Bayesian structure of human factors for the risk of collision on a ship is established as shown in Fig. 5.

1. Bayesian network structure of ship factors

According to the relevant cases ship accidents, "ship equipment failure or failure S21" includes mainly "main engine failure or failure S22" and "auxiliary engine failure or failure S23". According to the interrelationship between ship faults, the Bayesian structure of ship collision risk factors is established, as shown in Fig. 6.

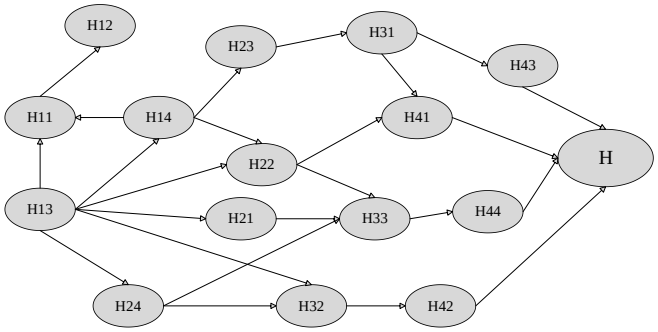

Fig. 5. Bayesian structure of human factors.

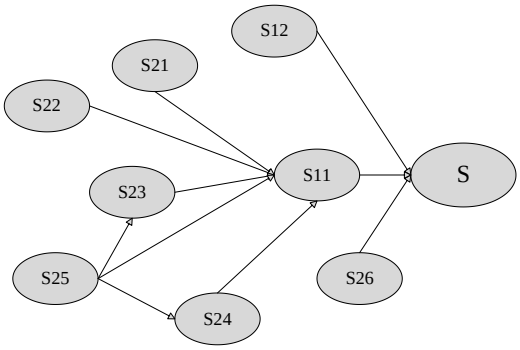

Fig. 6. Bayesian structure of ship factors.

2. Environmental factors

Because the relationship between the internal environmental factors is less, the relevant Bayesian structure diagram is not established. Environmental factors are manifested mainly in the connection with the other three factors, which are reflected in the final complete Bayesian structure.

3. Bayesian network structure of management factors

As mentioned above, management factors "not found that the crew is not competent" and "not found the hidden danger of equipment" will make the hidden dangers of human factors and ship factors ignored, resulting in serious consequences. At the same time, the internal relationship of management factors is also clear, as shown in Fig. 7.

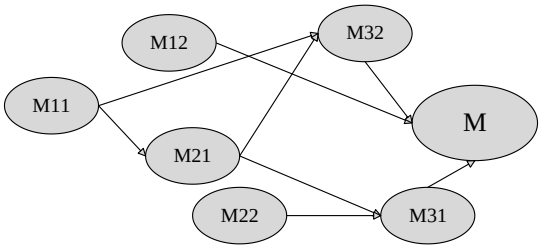

Fig. 7. Bayesian structure of management factors.

4. Bayesian structure of ship collision risk based on "Human-Ship-Environment-Management"

Upon analysing the four components of ship collisions, we explore the relationships between them from an overall perspective. Extensive literature review and expert research have facilitated the establishment of a Bayesian structure of ship collision risk based on "human-ship-environment-management", as shown in Fig. 8.

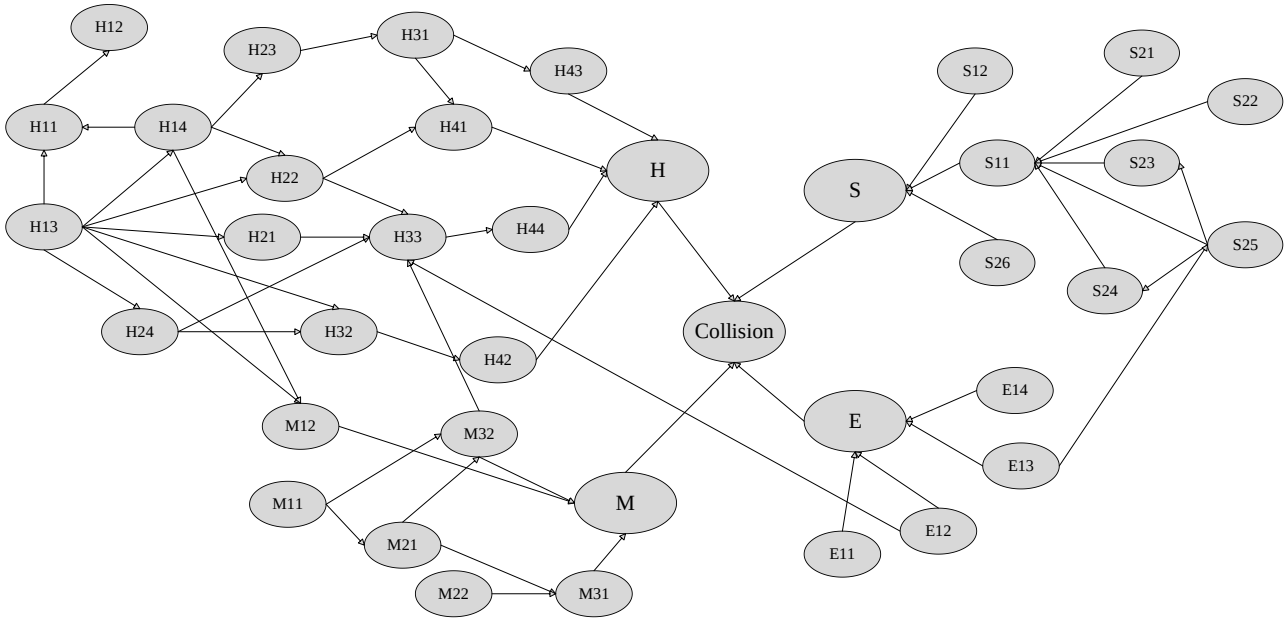

Fig. 8. Bayesian structure based on "human-ship-environment-management".

C. Bayesian Network Node Conditional Probability

Using the constructed Bayesian network structure to predict, it is necessary to determine the conditional probability of each node in the network. According to the existing ship collision/danger accident report, combined with the Bayesian network node probability calculation method, the conditional probability of each node can be obtained. In the following, the conditional probability determination process is introduced by taking "M32" as an example.

In the existing 419 ship collision/dangerous accident reports, the expressions of the causes of accidents involving improper crew cooperation mainly include "inaccurate judgment of the encounter situation of both ships", "incorrect judgment of ship dynamics", "insufficient communication", etc. The statistical results are shown in Table V.

TABLE V. STATISTICAL RESULTS OF "IMPROPER CREW COOPERATION M32".

No misconduct found M11		0		1	
Improper crew cooperation M32		0	1	0	1
0		45	88	41	48
1		54	75	24	44
Total	419	99	163	65	92

Based on the statistical results of relevant accident reports, the corresponding conditional probability table is obtained after estimating the Bayesian expectation type, as shown in Table VI.

TABLE VI. CONDITION PROBABILITY TABLE OF "IMPROPER CREW COOPERATION M32".

No misconduct found M11	0		1	
Improper crew cooperation M32	0	1	0	1
0	0.459	0.527	0.628	0.505
1	0.541	0.473	0.372	0.495

D. An Example of Ship Collision Risk Prediction Based on Bayesian Network

The data of this study come from the collision accident/danger report data of a segment of the Yangtze River Maritime Bureau in the past ten years. Through text mining technology, the accident/danger information is crawled, and the data shown in Fig. 9 are obtained. During the period 2008–2017, there were 229 ship collision accidents and 190 dangerous situations in this segment.

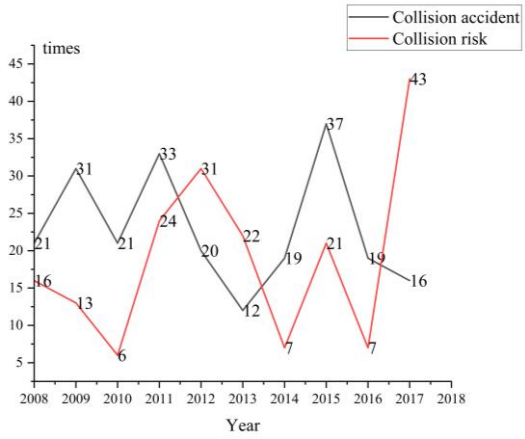

Fig. 9. Collision accidents/dangerous situations of a certain section of the Yangtze River Maritime Safety Administration from 2008 to 2017.

To validate the Bayesian model for ship collision risk constructed by our research institute, an analysis was conducted using the 2011 "Baimei # 8 collision accident" as an example. The accident overview is shown in Table VII.

TABLE VII. ACCIDENT OVERVIEW.

Accident time	May 5, 2011 at 10:00 am
Accident location	Funan Waterway # 56 floats down about 700 meters
Hydrometeor	Falling tide, southeast wind level 3-4, good visibility
Overview of the ship	
The "Baimei # 8" wheel	Chinese domestic trade bulk carrier, with a captain of 178 meters, a draft of 10.18 meters, a total tonnage of 19940, a net tonnage of 11351, and a deadweight tonnage of 33103 tons, was built in 2007 and loaded with 32063 tons of coal. It sailed from Baoshan to Zhenjiang.
Opposite vessel	The "Salt Sea Tow 98" is a heavy haul dragon boat fleet (10 barges loaded with coal).
Accident losses	Two barges of the "Yanhang Tuo 98" vessel sank, with one missing person. No losses were found in the "Baimei # 8" round.
Accident level	Major accident
Accident cause	
The main cause of the accident was a dragon boat fleet that lost control and blocked the waterway without releasing navigation information.	
The pilot did not maintain a regular lookout and did not promptly detect a dragon boat fleet in a state of loss.	
The "Baimei # 8" ship did not use a safe speed.	

According to the accident analysis, human factors such as "improper route selection H11", "improper lookout H21", "improper piloting H22", "improper route maintenance H23", "excessive reaction time H33", and "excessive speed H44" were identified. The state of "insufficient navigation assistance services E14" in environmental factors, "no misconduct detected M22" in management factors, and "improper crew cooperation M32" in management factors is set to "State1=1". Based on the actual situation and the opinions of relevant experts, other corresponding probabilities are determined. The above states are inputted into the model, and the results are shown in Fig. 10.

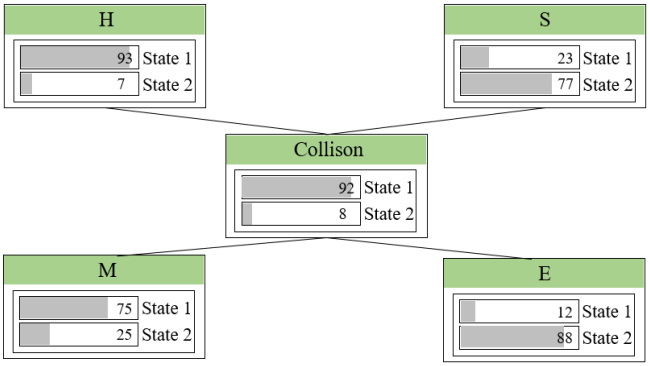

Fig. 10. Verification of the Bayesian model for ship collision risk.

It can be seen in Fig. 10 that the probability of collision occurring under relevant conditions is 92 %. At the same time, it can be seen that among the main factors causing ship collisions, human factors play a dominant role, accounting for 93 %, followed by management factors, reaching 75 %. The influence of ships and management factors can be ignored, accounting for 23 % and 12 %, respectively.

Similarly, 15 typical cases out of 419 were selected for analysis. The results obtained are shown in Table VIII.

From the results, it can be seen that except for the three collision accidents with a probability of 87 %, 88 %, and 85 %, the predicted results of other collision accidents/case studies are all above 90 %, proving the effectiveness of the model constructed by our research institute.

On the basis of proving the effectiveness of the Bayesian model for ship collision risk, the powerful inference ability of Bayesian networks is utilised to analyse the causes of ship collision. The probability of ship collision is set to 1, and probability analysis is conducted on human, machine, environmental, and management factors. The reverse inference results are shown in Table IX.

TABLE VIII. ANALYSIS RESULTS OF TYPICAL COLLISION ACCIDENTS/DANGEROUS CASES.

Accident	Collision accident of "Xinjixiang 3" ship	Collision accident of "Queen Stella" ship	Collision accident of "Alita" ship
Collision probability	97 %	90 %	94 %
Accident	Collision risk of "Zhixian 10" ship	Collision accident of "Xianhu" ship	Collision accident of "Danhai2" ship
Collision probability	87 %	91 %	98 %
Accident	Collision risk of "Piloster" ship	Collision accident of "Hexie" ship	Collision accident of

			"Jixiangwan5" ship
Collision probability	92 %	95 %	93 %
Accident	Collision accident of "Huanghaikaituo" ship	Collision risk of "Yuanwei" ship	Collision accident of "Kunlunyou003" ship
Collision probability	96 %	88 %	97 %
Accident	Collision accident of "Runze" ship	Collision risk of "Taihai2" ship	Collision risk of "Hengli" ship
Collision probability	93 %	85 %	91 %

TABLE IX. ANALYSIS RESULTS OF CAUSES OF SHIP COLLISION RISK.

Factor	H11	H12	H13	H14
Probability	11 %	7 %	65 %	47 %
Factor	H21	H22	H23	H24
Probability	72 %	63 %	44 %	37 %
Factor	H31	H32	H33	H41
Probability	47 %	75 %	69 %	77 %
Factor	H42	H43	H44	S11
Probability	21 %	17 %	83 %	3 %
Factor	S12	S21	S22	S23
Probability	5 %	1 %	1 %	1 %
Factor	S24	S25	S26	E11
Probability	1 %	0 %	17 %	4 %
Factor	E12	E13	E14	M11
Probability	55 %	1 %	7 %	11 %
Factor	M12	M21	M22	M31
Probability	1 %	23 %	37 %	44 %
Factor	M32	H	S	E
Probability	37 %	85 %	2 %	26 %
Factor	M			
Probability	43 %			

On the basis of the results, it is evident that human factors play a primary role in ship collisions. Specifically, "insufficient rest H13", "improper observation H21", "improper piloting H22", "misestimation of collision risk H32", "excessive reaction time H33", "improper handling H41", and "speed too fast H44" are the main causes of collisions attributed to human factors. The impact of ship factors is relatively minor, while management factors and environmental factors have a general influence.

V. DISCUSSION

This paper examines 382 collision accidents that occurred within the jurisdiction of the Yangtze River Maritime Bureau between 2010 and 2014. Through descriptive statistics, chart methods, and other basic statistical approaches, the study reveals the temporal distribution pattern of ship collision accidents, the distribution pattern of ship accidents in different environments, and the relationship between ship types and collision accidents. Four classifications of ship collision accidents/dangerous factors are determined, namely human factors, ship factors, environmental factors, and management factors. Based on these classifications, the research investigates water traffic accident reports, utilising text mining technology to explore the mechanism of occurrence and internal causal factors of such accidents. The factors that contribute to ship collision risks are identified, with seven main factors highlighted: wrong estimation of collision risk, improper manipulation, insufficient understanding of information, improper observation, strong wind, poor visibility, and excessive speed. When comparing the results obtained through descriptive statistics with those derived from text mining, it becomes apparent that the latter provides a more comprehensive understanding of the causes of ship collisions. Text mining supplements the limitations of intuitive statistics, offering a broader perspective on the underlying factors that contribute to these accidents.

The cause factors for ship collision accidents identified by text mining are further expanded in conjunction with previous studies. The ship collision process is divided into four steps: prenavigation preparation, navigation alert, collision avoidance decision-making, and manipulation execution, in order of priority. The expanded cause factors of ship collision accidents are treated as prior distributions and a Bayesian network structure of ship collision risk is constructed based on the "human-ship-environment-management" system. Using existing investigation reports of ship collision accidents, the conditional probabilities of each node in the Bayesian network are statistically calculated. This leads to the construction of the Bayesian network node conditional probability table and the Bayesian ship collision risk network model. To validate the effectiveness of the model, data from 15 typical collision accidents are entered, resulting in a probability of collision exceeding 90 %. This validation confirms the reliability of the model. Through further analysis using the model, the probability of collision is set to 1, allowing the conclusion to be drawn that human factors are the primary causes of ship collisions. Specifically, "insufficient rest H13", "improper observation H21", "improper piloting H22", "misestimation of collision risk H32", "excessive reaction time H33", "improper handling H41", and "speed too fast H44" are identified as the most prominent causes of collisions attributed to human factors.

VI. CONCLUSIONS

This paper mainly studies the text mining of the investigation report of ship collision accidents in the Yangtze River Basin, and uses the accident causes obtained by text mining to establish a Bayesian network model of ship collision risk based on previous studies, and applies the model to the reasoning of ship collision risk causes.

Text mining technology has emerged as a powerful tool for analysing large volumes of unstructured textual data, including accident reports. Using machine learning algorithms and natural language processing techniques, text mining can extract valuable information and insights from text data that would be difficult or impossible to obtain through manual analysis alone. In the field of accident investigation and risk analysis, text mining has great potential to improve our understanding of the causes of accidents and identify patterns and trends that may not be immediately apparent. This can help us develop more effective prevention strategies and mitigate the risks associated with accidents.

However, it is important to note that text mining should be used in conjunction with other methods, such as expert interviews and site visits, to ensure the accuracy and completeness of the data. Furthermore, the quality of the text data used in the analysis can affect the reliability of the results. Therefore, it is crucial to carefully select and preprocess the data to minimise errors and biases. Despite these challenges, the application of text mining technology in accident investigation and risk analysis is a rapidly evolving

field with enormous potential.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] L. Su, X.-j. Qin, Z.-l. Liu, and Z. Zhang, "Intelligent collision avoidance decision for single ship considering ship maneuverability", in *Proc. of 2019 9th International Conference on Information Science and Technology (ICIST)*, 2019, pp. 164–168. DOI: 10.1109/ICIST.2019.8836732.

[2] Y. Xu and W. Li, "An analysis of ship collision risk parameters based on speed ship domain", in *Proc. of 2021 4th International Symposium on Traffic Transportation and Civil Architecture (ISTTCA)*, 2021, pp. 99–103. DOI: 10.1109/ISTTCA53489.2021.9654635.

[3] S. Li, J. Liu, X. Wang, H. Guan, and Q. Yan, "Optimizing the anti-collision decisions between multiple ships based on predictive risk evaluations", in *Proc. of 2019 5th International Conference on Transportation Information and Safety (ICTIS)*, 2019, pp. 1395–1399. DOI: 10.1109/ICTIS.2019.8883795.

[4] P. Haapasaari, I. Helle, A. Lehikoinen, J. Lappalainen, and S. Kuikka, "A proactive approach for maritime safety policy making for the Gulf of Finland: Seeking best practices", *Marine Policy*, vol. 60, pp. 107–118, 2015. DOI: 10.1016/j.marpol.2015.06.003.

[5] M. B. Zaman, "Study on safety of navigation using automatic identification system for marine traffic area case study: Malacca Straits", *Int. J. of Marine Engineering Innovation and Research*, vol. 1, no. 1, pp. 26–30, 2016. DOI: 10.12962/j25481479.v1i1.1462.

[6] A. G. Dominguez, ""Smart ships": Mobile applications, cloud and bigdata on marine traffic for increased safety and optimized costs operations", in *Proc. of 2014 2nd International Conference on Artificial Intelligence, Modelling and Simulation*, 2014, pp. 303–308. DOI: 10.1109/AIMS.2014.39.

[7] A.-H. Tan, "Text mining: The state of the art and the challenges", in *Proc. of the PAKDD Workshop on Knowledge Discovery from Advanced Databases*, 1999, pp. 65–70.

[8] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems*. The Benjamin/Cummings Publishing, 1989.

[9] T. Zhou, W. Liu, M. Zhang, and J. Jia, "Optimization of AEB decision system based on unsafe control behavior analysis and improved ABAS algorithm", *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2023. DOI: 10.1109/TITS.2023.3317095.

[10] L. Yan, L. Jia, S. Lu, L. Peng, and Y. He, "LSTM‐based deep learning framework for adaptive identifying eco‐driving on intelligent vehicle multivariate time‐series data", *IET Intelligent Transport Systems*, pp. 1–17, 2023. DOI: 10.1049/itr2.12443.

[11] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications", *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, pp. 60–76, 2009. DOI: 10.4304/jetwi.1.1.60-76.

[12] N. Kanya and S. Geetha, "Information extraction - A text mining approach", in *Proc. of IET-UK International Conference on Information and Communication Technology in Electrical Sciences*, 2007, pp. 1111–1118. DOI: 10.1049/ic:20070576.

[13] S. Montalvo, R. Martínez, A. Casillas, and V. Fresno, "Bilingual news clustering using named entities and fuzzy similarity", in *Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science()*, vol. 4629. Springer, Berlin, Heidelberg, 2007, pp. 107–114. DOI: 10.1007/978-3-540-74628-7_16.

[14] C.-P. Wei, C. C. Yang, and C. M. Lin, "A Latent Semantic Indexing-based approach to multilingual document clustering", *Decision Support Systems*, vol. 45, no. 3, pp. 606–620, 2008. DOI: 10.1016/j.dss.2007.07.008.

[15] H. Anaya-Sánchez, A. Pons-Porrata, and R. Berlanga-Llavori, "A document clustering algorithm for discovering and describing topics", *Pattern Recognition Letters*, vol. 31, no. 6, pp. 502–510, 2010. DOI: 10.1016/j.patrec.2009.11.013.

[16] D. Isa, V. P. Kallimani, and L. H. Lee, "Using the self organizing map for clustering of text documents", *Expert Systems with Applications*, vol. 36, no. 5, pp. 9584–9591, 2009. DOI: 10.1016/j.eswa.2008.07.082.

[17] L. Gao and H. Wu, "Verb-based text mining of road crash report", in *Proc. of Transportation Research Board 92nd Annual Meeting*, 2013.

[18] R. Nayak, N. Piyatrapoomi, and J. Weligamage, "Application of text mining in analysing road crashes for road asset management", in *Engineering Asset Lifecycle Management*. Springer, London, 2010, pp. 49–58. DOI: 10.1007/978-0-85729-320-6_7.

[19] I. Septiana, Y. Setiowati, and A. Fariza, "Road condition monitoring application based on social media with text mining system: Case study: East Java", in *Proc. of 2016 International Electronics Symposium (IES)*, 2016, pp. 148–153. DOI: 10.1109/ELECSYM.2016.7860992.

[20] O. Kovalchuk, S. Banakh, M. Masonkova, K. Berezka, S. Mokhun, and O. Fedchyshyn, "Text mining for the analysis of legal texts", in *Proc. of 2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, 2022, pp. 502–505. DOI: 10.1109/ACIT54803.2022.9913169.

[21] Y. Huang and H. Tao, "Research on text naming recognition algorithm based on text mining", in *Proc. of 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, 2022, pp. 431–435. DOI: 10.1109/ICDSCA56264.2022.9988337.

[22] Y. Hata, N. Hayashi, Y. Makino, A. Takada, and K. Yamagoe, "Alarm correlation method using Bayesian network in telecommunications networks", in *Proc. of 2022 23rd Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2022, pp. 1–4. 10.23919/APNOMS56106.2022.9919924.

[23] V. Senthilkumar and V. Jayalakshmi, "Radial Basis Function Networks for image restoration with stochastic normalizations as Bayesian learning in deep conventional neural network", in *Proc. of 2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 2022, pp. 1298–1302. DOI: 10.1109/ICCMC53470.2022.9753702.