buildings

Article

Identification of Safety Risk Factors in Metro Shield Construction

Chao Tang [1,2], Chuxiong Shen [1], Jiaji Zhang [1,2] and Zeng Guo [1,*]

[1] School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China; 349235@whut.edu.cn (C.T.); 1378638258@whut.edu.cn (C.S.); zhangjiaji@whut.edu.cn (J.Z.)
[2] Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572025, China
* Correspondence: 12981@whut.edu.cn

Abstract: Among the construction methods for subway projects, shield method construction technology has become a more widely used construction method for urban subway construction due to the advantages of a high degree of construction mechanization, low impact of the construction process on the environment, and strong adaptability of the shield machine to the stratum, etc. However, because of the complexity of the surrounding buildings (structures) in the subway construction, coupled with the diversity of the subway shield method construction activities and the uncertainties in the construction environment, to a certain extent, it is determined that the subway construction process is very complicated. The purpose of this study is based on the text mining method, where text is mined and utilized to realize the identification, extraction, and display of safety risk factors. Thus, it guides the safety management on site and provides a basis for knowledge reuse in other metro shield construction projects. Firstly, we analyze the shortcomings of safety risk management in domestic and international metro shield construction via a literature review, especially the utilization of safety risk text data. Secondly, we collect the risk reports submitted by all parties via the "Metro Project Safety Risk Early Warning System", and manually screen the hidden danger statements with risk characterization to establish a corpus. Thirdly, we use the Jieba word separation package to extract and display the safety risk factors, so as to guide the on-site safety management. Subsequently, with the help of the Jieba word segmentation package for Chinese word segmentation, we develop a professional thesaurus to improve the effect of word segmentation; then, we use the TF-IDF parameter assignment to achieve the structural transformation of the text to extract high-frequency vocabulary; finally, from the high-frequency vocabulary to screen words containing the semantics of the risk to establish the risk of an initial set of words, we use the existing standards and norms to form the collection of safety risk factors of subway shield construction and generate the cloud diagram for visual display.

Keywords: metro construction; shield method; text mining; risk management

Citation: Tang, C.; Shen, C.; Zhang, J.; Guo, Z. Identification of Safety Risk Factors in Metro Shield Construction. *Buildings* **2024**, *14*, 492. https://doi.org/10.3390/buildings14020492

Academic Editor: Paulo Santos

Received: 19 January 2024
Revised: 5 February 2024
Accepted: 7 February 2024
Published: 9 February 2024

1. Introduction

Since the 21st century, the construction of rail transportation has been increasing, and urban rail transportation represented by the subway has the characteristics of large passenger capacity, high safety, high speed, low pollution, low energy consumption, etc., which largely relieves the traffic pressure in the city. However, the subway construction process is complex; the shield method is the main technology of subway construction, but safety accidents occur frequently. Subway shield construction safety risk management has defects and deficiencies. Risk management relies on subjective experience, the lack of mining, and the utilization of objective text data; it is difficult to meet the needs of subway shield construction safety risk management. Therefore, this paper carries out subway shield construction safety risk identification based on safety risk text data, extracts coping strategies and measures for key risks, and plays an important role in helping to improve subway shield construction site safety management. The structure of this paper is as follows: the Section 1 summarizes the current research on the risk management of metro shield construction and the practical application of text mining technology in metro construction; the Section 2 puts forward the research idea and research methodology of this paper; the Section 3 begins to process and analyze the existing data text; and the Section 4 is to sort the safety risk factors of metro shield construction and generate a cloud diagram for visualization and summary.

2. Literature Review

2.1. Research on Risk Management of Subway Shield

Li et al. took shield collapse accidents as the research object, analyzed the formation mechanism of construction workers' safety ability, and built a construction workers' safety ability model based on the perception and judgment of hidden dangers [1]. Chen et al. combined the triangular fuzzy number and cloud theory in the Bayesian network to build a risk analysis model for the underpass section of the shield and conducted risk assessment by taking the actual project as an example [2]. Yin et al. established the safety risk network structure of subway shield construction based on social network analysis and identified key risks with line centrality as the standard to provide a decision-making basis for risk control [3]. Taking Nanning Metro Line 3 as the background, Liu et al. divided the shield construction section to establish the shield construction structure model and identify key risk factors via matrix weight calculation [4]. Based on Bayesian networks, Chung et al. established a TBM risk analysis model for shield construction, systematically identified potential risk events of shield construction, estimated the countermeasures cost of accidents, and assessed the risk level of potential risk events [5]. According to the geological risks of subway shield construction, Nezarat et al. used a fuzzy analytic hierarchy process to sort various risk factors so as to guide the shield construction on site [6]. Yazdani et al. proposed a risk assessment model based on fuzzy set theory to evaluate risk events during subway shield construction and compared it with traditional risk assessment methods [7]. Zhou et al. used complex networks to analyze subway construction accidents and finally obtained a directed powerless network with 26 vertices and 49 edges. Via data analysis, immune strategies were adopted to reduce network efficiency and guide the safety management of subway shield construction on site [8]. Xue et al. set up the evaluation index system of excavation face stability based on the underpass river of shield tunneling, calculated the weight by the AHP-entropy weight method, and established the evaluation model of excavation face stability based on the thought point method [9]. Ren et al. set up a construction safety risk evaluation index system for buildings adjacent to shield construction in a certain section of Metro Line 3 in Xi 'an and used a fuzzy comprehensive evaluation method to evaluate the safety risk level of shield construction in the area [10]. Chen et al. combined subjective and objective methods to identify the safety risk factors of subway shield construction built an accident causative model with an interpretive structure model and analyzed the influence relationship between the factors with a decision laboratory [11]. Wang et al. took the Wuhan subway project as an example, conducted an overall analysis of the factors affecting the safety system of subway operation tunnels, and established a hierarchical structure model. On the basis of comprehensive risk evaluation, the risk grade of the tunnel shield construction section is determined by the fuzzy synthesis judgment model, maximum membership principle, and $R = P \times C$ [12]. Taking the Tianjin Metro project as an example, Pan et al. established a comprehensive index system of shield tunnel construction safety risk system based on fuzzy entropy theory. In addition, in order to quantitatively analyze the coupling degree between various factors in the safety risk system, a calculation model of coupling degree is established based on the coupling degree theory in physics [13]. Cao et al. studied a method of establishing risk analysis standards in shield tunnel construction: 3D numerical modeling using representative conditions. Risk control measures were then recommended based on the findings [14]. Huang et al. compared the TDCM evaluation method with the one-dimensional cloud model (ODCM) evaluation method and the Fuzzy Comprehensive evaluation method (FCEM) and discussed the advantages and applicability of the TDCM evaluation method [15].

2.2. Application of Text Mining in Subway Construction

Text mining is the process of obtaining interesting or useful patterns from unstructured text information. Text mining covers a variety of technologies, including information extraction, information retrieval, natural language processing, and data mining. Liu et al. applied text mining technology to tunnel engineering, established a tunnel engineering risk assessment index system with the help of R language and Jieba word segmentation, and developed a tunnel engineering risk assessment system on this basis [16]. Liu et al. collected the subway construction safety accidents after the 21st century, established a construction safety accident database, identified 48 due factors and 13 accident types, used association rules and complex networks to build a subway construction safety accident causation network, and conducted immune research on nodes [17]. Xu et al. took the safety accident report as a corpus and identified the risk factors and risk correlation relationship of subway construction safety based on text mining method, including causality and coupling relationship, and established a risk assessment model based on interpretive structure model and Bayesian network [18]. Ji et al. used a web crawler to collect subway construction safety accident cases, identified 67 risk keywords by text mining method, established a subway construction safety risk network via a complex network, and identified key risk nodes. Then, the risk probability reasoning was carried out based on the Bayesian network, and the cost control was carried out via scenario analysis [19]. Son et al. conducted text mining of bidding documents and contract documents of large-scale EPC projects in South Korea and established a schedule delay estimation model to assess the forecast schedule risk so as to determine the appropriate project duration [20]. Li et al. used association rules to find the risk correlation of subway engineering, obtained 45 subway engineering construction monitoring combinations, and proposed risk countermeasures [21].

Zhang et al. studied the application of data mining technology in the information processing of the subway automatic data acquisition system and proposed the framework of the data mining system. Based on subway data acquisition technology, this paper studies the analysis method of subway passenger flow and travel information. By using data mining technology and statistical analysis, the metro OD matrix and traffic rate are derived from the collected data, and the travel time distribution of passengers is described in detail, which is of great significance to the scheduling and management of the metro system [22]. Hsu et al. developed a responsive passenger letter system for the Taipei Metro case study example. After random sampling of passenger letters with text types was obtained, text mining technology in text letter files was used to find customary or even new keywords to improve service quality, such as customer satisfaction [23]. Mo et al. propose a measure that uses data mining techniques to create structured data sets for subway system equipment by analyzing historical maintenance records to monitor its daily status and possible fault development trends and attempt to apply predictive dimensions before any equipment actually fails [24]. Juan et al. extracted five years of subway accident records from the document. Via text clustering, the main influencing factors of subway delay are obtained. The relationship between the influencing factors and subway delay was established by using a logit regression model [25]. Chang et al. explored data-driven security risk assessment and response models via deep learning and complex network theory. Based on key security risk factors, corresponding risk countermeasures are proposed to verify the effectiveness and applicability of the data-driven security risk management model [26].

Mou et al. studied the subway operation hazard identification algorithm based on the text mining of subway operation logs, and the research results showed that the AFP-tree algorithm could significantly improve the computing efficiency and a total of 25 types of effective critical hazard sources were excavated via the analysis, and the research results could provide an important basis for the subway operation unit to achieve "prior" accident prevention [27]. Ye Cheng et al. proposed a classification method based on an improved BERT model and a structured retrieval method based on a knowledge graph to realize text classification and efficient data retrieval of subway construction hidden dangers and provide support for the development and application of the integrated system. At the same time, this research can also provide references for text processing, data retrieval, and management in the field of architecture based on deep learning and knowledge graph technology [28]. Pan et al. proposed a text framework for automated analysis of hidden danger detection based on text mining and visualization technology, which was applied and verified in the analysis of construction safety hidden danger detection records of Wuhan Metro from 2016 to 2018. The experimental results show that the framework can effectively excavate the key points and visual information corresponding to 34 types of hidden dangers [29]. In view of the massive unstructured subway construction hidden danger text, Hei et al. proposed the idea of using text mining technology and visualization technology to analyze subway construction hidden danger so as to transform abstract text data into visual information, assist future hidden danger investigation and provide data support for it, which can be used for subway enterprises to compile hidden danger investigation yearbook and use the visual analysis results. It has practical application value in worker safety training [30].

Li et al. used R language and text mining methods to carry out word segmentation processing, feature item selection, vector space model construction, and co-occurrence rule recognition for accident reports and visualized text mining results by using word cloud and network structure graphs. Six key risk factors and 23 general risk factors of subway construction safety accidents were found [31].

2.3. Summary of Literature Research

At present, the vast majority of research is based on text mining methods to mine and utilize safety risk text or data. Because a large amount of safety risk text data is generated in the process of metro shield construction, including safety risk reports, safety inspection reports, work contact sheets, monitoring data analysis reports, risk warning sheets, etc., the conversion of text data to safety risk factors is realized via text mining-based methods, and the identification of safety risk factors and the extraction of correlation relations are realized via objective data and rarely use TF-IDF for parameterized assignment. In this article, data screening is carried out by calculating the characteristic parameters, and finally, the collection of subway shield construction safety risk factors is obtained via semantic identification; then, the indicators are screened by setting the association rules and the strong correlation relationship between the risk factors is obtained by mining based on the a priori algorithm; finally, the subway shield construction safety risk system structure is established based on the explanatory structural model, and the safety risk factors are graded, and the key risk factors are identified, which guides the on-site safety management, and provides a basis for the reuse of the knowledge to other subway shield construction projects.

3. Research Methods

In order to find the key risk factors of subway shield construction, this study conducts analysis analyzed based on the text mining method. It is mainly divided into the following steps. First, the research problem is put forward and a theoretical analysis is conducted. Combining the domestic and international research background, the deficiencies in the safety risk management of metro shield construction are analyzed, especially the lack of utilization of safety risk text data and the traditional system modeling relying on the subjective experience of experts. Based on the research problems, the research ideas are sorted out, the system engineering theory and risky management theory are analyzed, and the subway shield construction safety risk management process are put forward to lay the theoretical foundation for the subsequent research.

Secondly, based on text mining, we identify the subway shield construction safety risk factors. We use the "subway engineering safety risk warning system" to collect safety risk reports submitted by all parties, screen risky statements to establish a corpus, use the Python language Jieba toolkit for Chinese word segmentation, develop a professional thesaurus to improve the effect of word segmentation, and use TF-IDF for parameterization. The TF-IDF is used for parameter assignment to extract high-frequency words; from the high-frequency words, the initial set of safety risks is established by screening the words with risk semantics, and the collection of 31 safety risk factors for metro shield construction is formed by comparing with the existing standards and specifications. The research framework is shown in Figure 1.

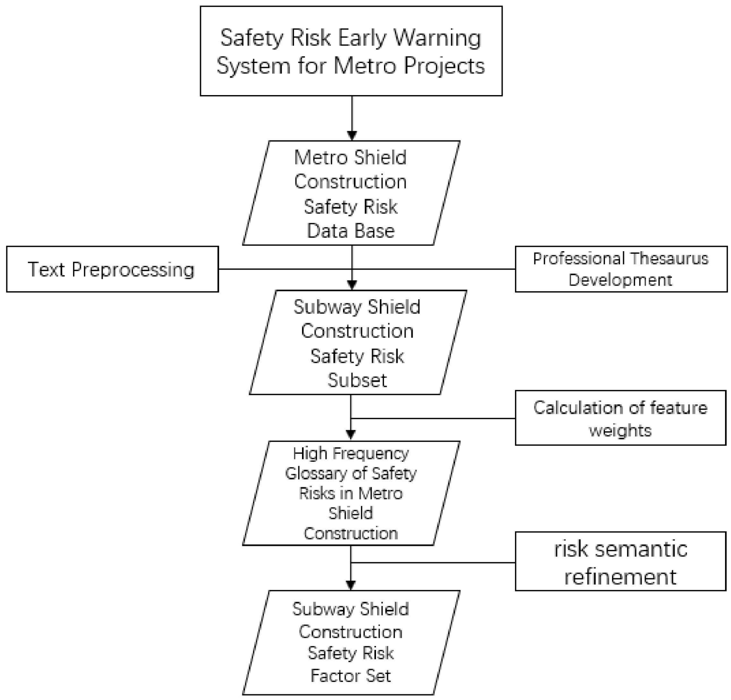


Figure 1. Research flowchart.

3.1. *Application of Text Mining in Subway Construction*

The premise of the subway shield risk text is to realize the Chinese proposed word division to choose the appropriate word division algorithm. At present, the main types of algorithms are based on dictionary, statistics, and semantic understanding of the word separation algorithm, such as the maximum matching algorithm, Markov model, long and short-term memory model, etc. Chinese proposed word separation needs to be combined with the relevant analytical tools; the current applications of the word separation tool for text mining are the Jieba word separation package, ICTCLAS system, NLPWin automatic word separator, SEG word separation system, NLPIR system, etc. Among them, the Jieba word separation package and ICTCLAS system have faster word separation speeds, the ability to perform lexical annotation, and so on. Among them, the Jieba package and ICTCLAS system are widely used because they have faster word separation speed, can perform lexical annotation, and have better word separation effects. Compared with the ICTCLAS system, the Jieba lexical package is more reasonable for risky vocabulary cutting, suitable for Python coding, and better for Chinese proposed lexical and data statistics, so this study proposes to use the Jieba lexical package for Chinese lexical.

There are three modes in the Jieba lexical package:

(1) Exact mode: cut the text into precise parts without redundant words.
(2) Full mode: all possible words in the text are scanned out with redundancy.

(3) Search engine mode: on the basis of the exact mode, the long words are cut apart again.

In this paper, because the existing complex text is divided into words, the precise mode via the Python3.12.0 software is chosen to call the Jieba word division package the core of the word division code mode:

```
seg_list = jieba.cut("Risk_text", cut_all = False)
```

3.2. Parameter Assignment Using TF-IDF for Structured Transformation of Text

A large number of feature terms are obtained after using the thesaurus development for the Chinese lexicon of subway shield construction safety risk identification and correlation analysis corpus; however, it still contains a large number of words not related to risk, which causes interference in the subsequent risk identification and correlation analysis; therefore, in order to transform the textual data into structured data, and to screen out the feature terms with the semantics of risk, this study carries out the structuralization of the text of the safety risk Transformation.

Using text mining methods to transform unstructured text data into structured data requires the assignment of parameters to the text, and the parameters that characterize the text are word frequency (TF), document frequency (DF), and word frequency–inverse document frequency (TF-IDF).

(1) Word Frequency (TF)

Word frequency indicates the frequency of a risk word in the document; a word that appears in the document with high frequency is called a high-frequency word. Word frequency indicates the value of vocabulary for text analysis; the higher the word frequency, the greater the value. The process of text mining can be considered to set the lower bound of word frequency to filter the mining value of low-frequency words. Word frequency is expressed in mathematical language as

$$tf_{ij} = \frac{n_i^j}{\sum\limits_{i=1}^{k} n_i^j} \tag{1}$$

where n_i^j denotes the frequency of occurrence of risk term i in document j, k denotes the number of risk terms in document j, and tf_{ij} denotes the word frequency of risk term i in document j.

(2) (Inverse) Document Frequency (DF/IDF)

Document frequency refers to the number of texts containing a risk word in the entire text set, in order to measure the importance of the risk word for the entire text set, the inverse document frequency is often used to describe. In order to measure the importance of a term to the entire text set, the inverse document frequency is often used to describe it. That is, the fewer the number of documents in which a term appears, the higher the importance of the term and the greater the mining value. Document frequency and inverse document frequency are expressed in mathematical language as

$$df_i = \sum_{i=1}^{N} X_j \quad X_j = \begin{cases} 1 & t_i \in d_j \\ 0 & else \end{cases} \tag{2}$$

$$idf_i = \log \frac{N}{df_i} \tag{3}$$

where N denotes the number of documents in the document set and is a 0–1 variable: 1 when document j contains risk term i and 0 otherwise.

(3) Word Frequency–Inverse Document Frequency (TF-IDF)

For risk feature vocabulary, it is not possible to judge whether the word is a risk feature vocabulary only based on word frequency or document frequency, assuming that the word frequency of a vocabulary is high; but when it appears repeatedly in other documents, its value decreases, and it is only when it has a high word frequency and a low document frequency that the vocabulary is a feature vocabulary with a certain degree of uniqueness. Word frequency–inverse document frequency is a weighted formula used to measure the general importance of words, expressed in mathematical language as

$$(tf - idf)_i^j = \frac{n_i^j}{\sum\limits_{i=1}^{K} n_i^j} \times \log \frac{N}{\sum\limits_{j=1}^{N} X_j} \tag{4}$$

Word frequency–inverse document frequency combines the characteristics of word frequency and document frequency and is a weighing formula used to measure the universal importance of feature items.

4. Results and Analysis

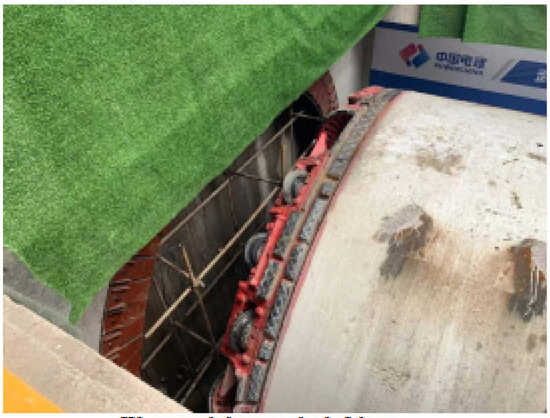
4.1. Building a Corpus

The Metro Group requires all units involved in the construction to conduct regular safety inspections, while the risk assessment unit organizes daily safety inspections of the shield construction zone. The results of the safety risk assessment and site monitoring reports are uploaded to the Metro Engineering Early Warning System on a weekly basis. So far, the risk assessment unit has collected and uploaded a total of 308 safety risk reports. The risk text data are initially organized with the following characteristics: First, the text data are large in size, and all the monitoring data, such as settlement, offset, and axial force, are recorded during the construction of the metro shield structure. Second, the text data are many in the process of subway shield construction safety risk management, mainly including two categories of data, namely the numerical data of sensing monitoring and the text data of risk analysis by management personnel, and the numerical data and the text data can be divided into a variety of categories according to the different objects of analysis. Thirdly, data; data production speed is fast. Fourth, the data authenticity is high; all the inspection record data are obtained from the real record data at the scene for the first time, and the use of sensors to obtain the monitoring data is also recorded for the first time via the system. Fifth, low data value density; due to the huge scale and fast production speed of text data extracted from safety risk records related to subway shield construction, the extraction of higher degrees of risk factors and the relationship between risk factors becomes more difficult, resulting in the whole data value density being low.

The following is a part of the subway shield construction safety risk text display, as shown in Figure 2.

The construction safety risk texts of metro shield construction for many lines of a certain metro in recent years from 2020 to 2023 was collected, such as the existing Line 3, Line 4, Line 5, Line 6, Line 8 Phase 1, the planned Line 10, Line 11 Phase 3, Line 14, Line 15, Line 16, Line 21, Line 24, Line 26, Line 28, and so on, and this was used as a text database. From the above report, it can be found that although risk information is recorded in the construction safety risk text, there exists a large amount of redundant information unrelated to text mining; therefore, risk statements are extracted from the construction safety risk text in accordance with the recording time, construction location, risk description, etc., and at the same time, the misspellings and non-standardized terms in the report are corrected, and the preliminary text data of 13,129 texts are sorted out as the basis for the construction. The corpus of subway shield construction safety risk identification and correlation analysis is shown in Table 1 below for some text data due to space limitations.

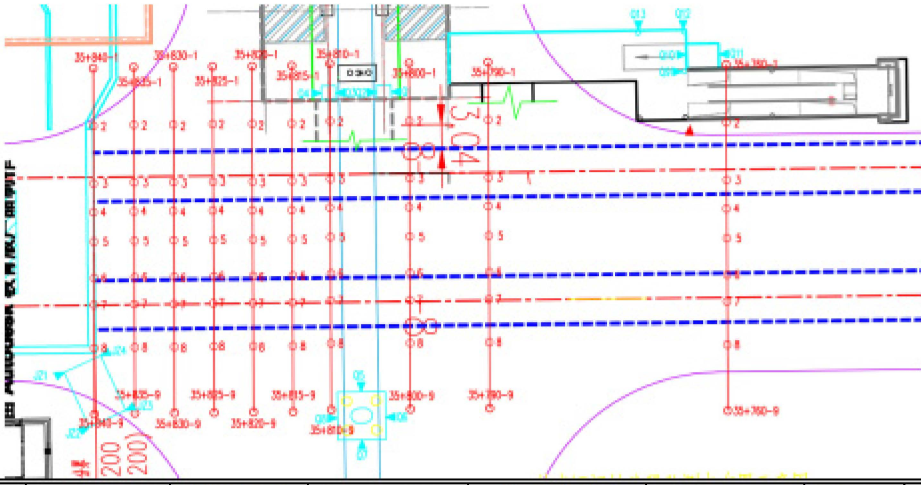
Shield machine debugging



Wet tunnel door on the left line

（2）Inspection situation: the two-line shield machine is powered on and debuted, and the water dripping in the probe hole of the double-line starting tunnel door does not increase significantly.

（3）Monitoring data analysis this week: At present, the double-line shield has not yet started, only the pier of Line 5 is monitored, and the monitoring data of the construction party is selected for analysis. The monitoring data this week is stable, and the detailed data are shown in the following table.



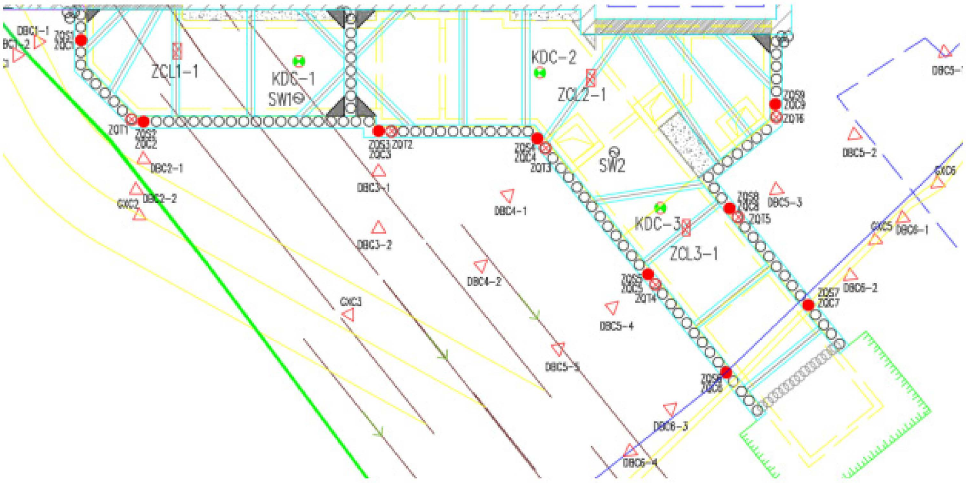
Monitoring item	Station number	Maximum cumulative deformation (mm)	Average daily rate of change (mm/d)	Cumulative deformation control value (mm)	Change rate control value (mm/d)	Early warning state	note
Pier settlement	Q1	-0.50	-0.05	±6	±2.00	normal	
Pier settlement	Q2	-0.53	-0.03	±6	±2.00	normal	
Pier settlement	Q3	-0.33	-0.05	±6	±2.00	normal	
Pier settlement	Q4	-1.28	-0.18	±6	±2.00	normal	

Figure 2. *Cont.*

No. 2 entrance and exit structure construction

No. 4 entrance support crown beam construction

（3） Monitoring data analysis this week: third-party monitoring data are selected for analysis, and specific analysis is as follows.



Monitoring item	Station number	Cumulative change (mm)	Cumulative quantity alarm value (mm)	Average speed of the week (mm/d)	Rate alarm value(mm/d)	Safe state
Surface settlement	2DBC3-2	-5.06	±30	-0.04	±2	Safe
Pile top settlement	ZQC3	-1.38	±20	-0.02	±2	Safe
Horizontal displacement of pile top	ZQS3(Y)	1.90	±40	0.07	±2	Safe
Horizontal displacement of pile	2ZQT3（8.5m)	8.21	±40	0.04	±2	Safe

Figure 2. Safety risk text of subway shield construction (part).

Table 1. Safety risk identification and correlation analysis of subway shield construction corpus.

Time	Construction Location	Risk Description
10 February 2022	Huchuan~Mafangshan Failure to export soil stockpiled	Failure to export soil stockpiled at the edge of the pit
10 February 2022	Chayesuo~Qingling Shield Structure Interval	Water dripping from exploratory holes of the left line start gate
13 January 2022	WuchangRailway~Central Garden Shield Structure Interval	Water and silt accumulation at the 410th ring of the left line
13 January 2022	Wuchang Railway Station	Longitudinal cracks in water retaining wall
13 January 2022	Wuchang Railway Station	Poor civilized construction on site, the earth was not transported out in time, and the accumulated mud was not cleaned up in time.
6 January 2022	Chayesuo~Qingling Shield Structure Interval	Water dripped continuously from the probe hole of the left line's starting hole, and water seepage was slight.
6 January 2022	Wuchang Railway Station	Slight seepage from the 322nd ring pipe sheet of the left line.
6 January 2022	Huchuan~Mafangshan Shield Structure Interval	Shield tunnel left line 401 tube sheet is damaged.

4.2. Text Preprocessing

The preprocessing process involves Chinese word segmentation of the corpus for identifying and analyzing safety risks in subway shield tunneling construction, dividing risk statements into independent words.

Taking the risk statement "In the shield tunneling section between Wuchang Railway Station and Central Garden Station, water seepage was found between piles and was not drained in a timely manner". as an example, a Chinese word segmentation was performed using code, and the segmentation result is as follows:

In the/shield/tunneling/section/between/Wuchang/Railway Station/and/Central Garden/Station, water/seepage/was/found/between/piles/and/was not/drained/in/a/timely manner/./

Using the built-in vocabulary of the Jieba word segmentation package to perform Chinese word segmentation on the corpus, we obtained segmentation feature items, some of which contain risk semantics, such as "foundation pit", "tunnel", "passage", "pipe segment", "shield", "enclosure", "construction", "soil", "structure", etc., as shown in Table 2. However, the effectiveness of using the built-in vocabulary of the Jieba toolkit for word segmentation is not satisfactory. In the above example, "Wuchang Railway Station" and "Central Garden Station" were split as interval names, which is not accurate in expressing the construction site. In addition, some professional vocabulary was also incorrectly split, affecting the accuracy of safety risk identification and association analysis. Therefore, it is necessary to develop a professional vocabulary to improve the effectiveness of word segmentation.

Table 2. Corpus Chinese word segmentation results (part).

No.	Characteristic Term	No.	Characteristic Term	No.	Characteristic Term
1	tunnel	5	tube sheet	9	management
2	channel	6	shield construction	10	monitor
3	construction	7	enclosure	11	framework
4	pit	8	officers	12	earthwork

4.3. Professional Vocabulary Development

In order to improve the accuracy of Chinese word segmentation, it is necessary to develop a professional thesaurus of subway shield construction safety risk and re-segment the corpus. The professional thesaurus constructed in this study includes a security risk thesaurus, custom thesaurus, and stop word thesaurus.

(1) Security risk thesaurus

In this study, the general dictionary of "Civil Construction" was downloaded from the Google input method, and "Engineering construction terms", "Safety Engineering", "Building construction lexicon", and "Building Structure" were downloaded from the Sogou lexicon to form a security risk lexicon. In addition, the text file userdict.txt is created to store the dictionary in the security risk database, and the load_userdic() method of Jieba is directly called to load the custom dictionary file when the Jieba is used.

(2) User-defined thesaurus

The safety risk thesaurus cannot fully meet the need for safety risk text mining for subway shield construction, so it is still necessary to further add professional vocabulary, mainly including line names and combination vocabulary. The name of the line mainly includes the names of subway lines, stations, and sections in the research area, avoiding the separation of the nouns representing the construction site. Combined words are for professional words, which is because the field of civil engineering is very wide, and it is difficult to fully cover professional terms. On the basis of Chinese word segmentation, some professional words are artificially combined (Tables 3 and 4).

Table 3. Custom thesaurus (part).

Category	
Line name	Line 3, Line 4, Wuchang Railway Station, Mafangshan Station, Fu'an Street Station, Jiangchu Avenue Station, Tea Leaf Institute Station,. . .
Combined vocabulary	Safety personnel, project progress, construction operations, construction personnel, construction program, shield interval, technical briefing, safety briefing, safety risk, construction organization design. . .

After citing the security risk thesaurus and the custom thesaurus, the effect of word segmentation is improved, and the results of word segmentation are as follows.

Table 4. Optimized segmentation results of security risk thesaurus and custom thesaurus.

No.	Segmentation Result
1	The soil/piled up/at/the edge/of/the/foundation pit/was/not transported out/./
2	The on-site/civilized construction/is/poor/,/the/earthwork/is/not transported out/in/a timely/manner/, and/the accumulated/mud/is/not cleaned up/in/a timely/manner/./
3	A/longitudinal/crack/appeared/on/the/water retaining wall/at/Wuchang Railway Station/./
4	There/is/a lot of/accumulated water/in the/shield tunnel/section/from Huquan Station/to Mafangshan Station/that/has/not been drained/in/a timely/manner

(3) Stop word database

After the addition of security risk thesauruses and custom thesauruses, the segmentation effect is obviously improved, but it still contains a lot of meaningless information, such as punctuation marks, descriptive words, and numerical forms, which need to be disabled. There are three types of stopword lexicon: First, the Dictionary of Modern Chinese Function Words based on ICTCLAS3.0 software; Second, some descriptive words such as "reasons", "needs", "suggestions", and a series of words without mining value; Third, punctuation marks and words in number form (Table 5). At the same time, in order to reduce the influence of construction sites, the "interval name" and "station name" are extracted, and the line names in the custom thesaurus are included in the stop word thesaurus to improve the efficiency of text mining. The core code pattern used by the stop word library is as follows:

```
with open('Stop_database.txt') as f:
con = f.readlines()
stop_words = set()
for i in con:
i = i.replace("\n", "")
stop_words.add(i)
for word in seg_list_exact:
if word not in stop_words and len(word) > 1:
result_list.append(word)
return result_list
```

After adding the stop word library and running it, the results of the statement segmentation are as follows.

Table 5. Optimization result of word segmentation of stop word database.

No.	Segmentation Results
1	The soil/piled up/at the edge/of/the foundation pit/was/not transported out/./
2	The on-site/civilized construction/is/poor/,/the earthwork/is/not transported out/in/a timely manner/, and/the accumulated mud/is/not cleaned up/in/a timely manner/./
3	A longitudinal crack/appeared/on/the water retaining wall/at/Wuchang Railway Station/./
4	There/is/a lot of/accumulated water/in the/shield tunnel section/from Huquan Station/to Mafangshan Station/that/has/not been drained/in/a timely manner

The use of a subway shield construction safety risk professional thesaurus can effectively improve the effect of Chinese word segmentation and obtain normative risk characteristic vocabulary. The structured language obtained after Chinese word segmentation is still not recognizable by the computer, so it is also necessary to carry out text structural transformation to transform feature words into structured language that can be analyzed and calculated, which requires parametric assignment of feature items.

4.4. Text Structure Transformation

In this paper, a word frequency–inverse document frequency algorithm is used to assign the feature items. According to the ABC classification method, the influencing factors are classified as class A factors, and the feature items with cumulative TF-IDF value accounting for 80% are extracted as high-frequency words, limited by space, and high-frequency words with TF-IDF value greater than 60 are shown in Table 6.

Table 6. High-frequency words with TF-IDF values greater than 60 are shown.

No	Feature Item	TF-IDF	No	Feature Item	TF-IDF
1	Safety	199.3	8	monitor	87.6
2	Management	134.2	9	control	83.8
3	Safety Awareness	111.2	10	epidemic situation	81.1
4	Shield	99.6	11	protect	72.9
5	Tunnel	95.2	12	stabilization	72.4
6	Construction technique	94.1	13	quality	69.1
7	Inspection	91.9	14	device	65.7

4.5. Identification of Security Risk Factors

After the text structure transformation, 301 high-frequency words of subway shield construction safety risk are extracted, and their meanings in the text of construction safety risk are comprehensively analyzed. A total of 30 high-frequency words containing risk semantics are screened out, and an initial list of subway shield construction safety risk factors is established, as shown in Table 7.

Table 7. Initial list of safety risk factors for subway shield construction.

No	High-Frequency Words	Risk Factor	TF-IDF
1	Administration	Chaotic on-site management	199.3
2	Safety consciousness	Lack of safety awareness	111.2
3	Shield tunneling	Shield tunneling angle deviation	99.6
4	Construction technique	Insufficient construction technology	94.1
5	Check	Insufficient on-site inspection	91.9
6	Monitor	Inadequate on-site monitoring	87.6
7	Epidemic situation	Inadequate epidemic prevention and control	81.1
8	Protect	Poor protection at the construction site	72.9
9	Quality	Quality defects	69.1
10	Device	Mechanical equipment malfunction	65.7
11	Patrol	Inadequate on-site inspection	61.0
12	Construction plan	Unreasonable construction plan	59.3
13	Management system	Incomplete safety management system	57.2
14	Supervise	Insufficient effective supervision	55.1
15	Survey	There are errors in the survey	54.9
16	Safety briefing	Insufficient safety briefing	53.6
17	Train	Insufficient safety training	53.5
18	Coordinate	Difficulty in on-site coordination	50.9
19	Command	Illegal command	49.9
20	Duration	Tight construction schedule	48.1
21	Pipeline	Pipeline relocation not marked	36.4
22	Structure	Structural defects	32.5
23	Emergency	Insufficient emergency management	29.2
24	Illegal construction	Illegal construction	25.8
25	Ventilation lighting	Insufficient ventilation lighting	19.5
26	Brace	Defects in tunnel support system	18.2
27	Material	Insufficient material usage	16.1
28	Model selection	Improper equipment selection	14.9
29	Support	Incomplete support system	13.7
30	Fire fighting	Insufficient fire-fighting equipment	10.9

At present, the current national standard that has compiled the list of safety risk factors and is most relevant to the safety risk management of subway shield construction is the "Subway Construction Safety Evaluation Standard (GB50715-2011)" issued by the Ministry of Housing and Urban–Rural Development, referred to as the "Evaluation Standard". The Evaluation Standard establishes an evaluation index system from four aspects: organization, technology, environment, monitoring, and early warning. The risk factors in the initial list of identified subway shield construction safety risk factors are compared and analyzed with the Evaluation Standard, as shown in Table 8.

Table 8. Comparative analysis of security risk factors.

No	Risk Factors	Corresponding Categories in the Evaluation Criteria
1	Chaotic on-site management	Organization–Construction Unit
2	Lack of safety awareness	/
3	Shield tunneling angle deviation	Technology–Shield tunneling construction
4	Insufficient construction technology	Technical–Construction Unit
5	Insufficient on-site inspection	Technology–Shield tunneling construction
6	Inadequate on-site monitoring	Organizational Monitoring Unit
7	Inadequate epidemic prevention and control	/
8	Poor protection at the construction site	Technical–Construction Management Measures
9	Quality defects	Organization–Construction Unit
10	Mechanical equipment malfunction	Technology–Shield tunneling construction
11	Inadequate on-site inspection	Organizational Supervision Unit
12	Unreasonable construction plan	Organization–Construction Unit
13	Incomplete safety management system	Organization–Construction Unit
14	Insufficient effective supervision	Organizational Supervision Unit
15	There are errors in the survey	Organization–Survey and Design Unit
16	Insufficient safety briefing	Technical–Construction Technical Measures
17	Insufficient safety training	Organization–Construction Unit
18	Difficulty in on-site coordination	/
19	Illegal command	Technology–Main Construction Processes
20	Tight construction schedule	/
21	Pipeline relocation not marked	Technology–Main Construction Processes
22	Structural defects	Technology–Shield tunneling construction
23	Insufficient emergency management	Technical–Construction Management Measures
24	Illegal construction	Technology–Main Construction Processes
25	Chaotic on-site management	Technical–Construction Management Measures
26	Lack of safety awareness	Technology–Shield tunneling construction
27	Shield tunneling angle deviation	Organization–Construction Unit
28	Insufficient construction technology	Technology–Shield tunneling method
29	Inadequate on-site monitoring	Technology–Shield tunneling construction
30	Inadequate epidemic prevention and control	Technical–Construction Management Measures

After comparative analysis, it is found that the safety risk factors in the initial list of safety risk factors of subway shield construction based on text mining are basically attributed to the Evaluation standards, which covered the safety evaluation standards of shield construction and can reflect the main safety risks of subway shield construction. In addition, the lack of safety awareness, the difficulty of epidemic prevention and control, the difficulty of organization and coordination, the tight construction period, and the fatigue of workers are not part of the evaluation criteria, and the above four types of factors are analyzed:

(1) Lack of safety awareness: Management personnel, construction technicians, and other ideological slack have insufficient vigilance toward dangerous sources. The lack of safety awareness directly affects the behavior habits of personnel and is easy to produce coupling effects of other risk factors to further cause safety accidents. That factor is therefore retained.

(2) Difficulties in epidemic prevention and control: Difficulties in epidemic prevention and control refer to the shortage of construction personnel caused by the parties involved in the subway shield construction due to the spread of the epidemic and the suspension of the project caused by epidemic prevention and control. Since the implementation of the Notice on Further Optimizing the Implementation of Prevention and Control Measures and a series of policies, the epidemic has leveled off. Therefore, this factor is not retained.

(3) It is difficult to organize and coordinate: The coordination of various units and professions in the process of organizing shield construction is not in place, resulting in interface conflicts. There are many units involved in the shield construction process,

and there are cross operations, and the organization and coordination are not in place, which may lead to security risks. That factor is therefore retained.

(4) Tight construction period: The short construction period target caused by the builder, or the construction period is compressed due to the construction party and other reasons. The tight construction period will cause the construction side to rush work, and the structural quality of shield construction may decline, which brings the risk of rush work to all units. That factor is therefore retained.

The initial list of risk factors is partially adjusted, the 30 risk factors are modified to 29, the list of subway shield construction safety risk factors is established, and the risk factors are described by combining safety risk texts and standard specifications, as shown in Table 9.

Table 9. List of safety risk factors of subway shield construction.

No	Risk Factor	Risk Interpretation
S_1	Lack of safety awareness	The construction general contracting unit's inadequate safety management of shield tunneling construction includes insufficient management personnel, inadequate control over subcontractors, and failure to establish a reasonable safety feedback mechanism.
S_2	Shield tunneling angle deviation	Management personnel, construction technicians, and others are prone to ideological laxity and lack of vigilance toward hazardous sources.
S_3	Lack of proficiency in construction techniques	During the process of shield tunneling and pipe assembly, there is a deviation in the posture of the shield tunneling process.
S_4	Improper equipment operation	Refers to the lack of technical proficiency in subway shield tunneling and inadequate pre-construction training.
S_5	Inadequate on-site monitoring	Refers to the non-standard operation of mechanical equipment in various processes of subway shield tunneling construction.
S_6	Poor protection at the construction site	The monitoring plan is unreasonable, or the monitoring timeliness, frequency, and accuracy are insufficient.
S_7	Quality defects	The construction site protection was not in place, and the site enclosure was not carried out as required.
S_8	Mechanical equipment malfunction	The built project does not meet the regulatory requirements and has quality defects.
S_9	Inadequate on-site inspection	Refers to the functional defects, damages, and other situations of mechanical equipment itself.
S_{10}	Unreasonable construction plan	Refers to the supervisory unit's failure to conduct inspections as required or the on-site supervision being superficial, etc.
S_{11}	Incomplete safety management system	The content of the construction plan is not scientific, and the feasibility and safety of the construction are not fully considered. Lack of professional knowledge and skills among construction personnel.
S_{12}	Insufficient effective supervision	It refers to the lack of complete safety management regulations and unclear division of rights and responsibilities.
S_{13}	There are errors in the survey	Refers to the supervision unit not being present for supervision or the supervision being superficial, etc.
S_{14}	Insufficient safety briefing	Refers to the failure of the survey and design unit to conduct on-site surveys as required or the surveys being superficial,

Table 9. *Cont.*

No	Risk Factor	Risk Interpretation
S_{15}	Insufficient safety training	Failure to organize safety briefing before construction or insufficient depth of briefing.
S_{16}	Difficulty in on-site coordination	It refers to failure to conduct regular safety training as required or insufficient professionalism in training.
S_{17}	Illegal command	Poor coordination among various units and specialties during the organization of shield tunneling construction, resulting in interface conflicts, etc.
S_{18}	Tight construction schedule	Refers to management personnel who fail to give incorrect instructions or issue inappropriate instructions in accordance with relevant rules and regulations.
S_{19}	Pipeline relocation not marked	It refers to the short construction period target caused by the construction party or the compression of the construction period due to reasons such as the construction party.
S_{20}	Structural defects	Failure to keep detailed records of pipeline location and place clear markings on-site during pipeline relocation
S_{21}	Insufficient emergency management	There are defects in the geological structure of the area where the shield tunnel is located, which may cause settlement, deformation, and other problems.
S_{22}	Illegal construction	Lack of effective emergency plans and untimely handling of some engineering hazards.
S_{23}	Insufficient ventilation and lighting	Improper construction behavior by construction personnel who fail to comply with relevant regulations.
S_{24}	Worker fatigue work	The aging and damage of ventilation and lighting equipment in the shield tunnel section, which leads to equipment failure, cannot guarantee the normal progress of construction activities.
S_{25}	Defects in tunnel support system	The instability of the tunnel support system is caused by untimely support or construction defects.
S_{26}	Insufficient material usage	The rules, performance, quality, etc., of construction materials do not meet actual needs.
S_{27}	Improper equipment selection	The selection of construction equipment models does not meet the actual engineering needs, or the selected equipment itself has quality defects.
S_{28}	Incomplete support system	There are problems with the installation of shield tunnel segment support, and the support construction is incomplete.
S_{29}	Insufficient fire-fighting equipment	The on-site fire-fighting equipment is not fully equipped and does not meet the fire-fighting conditions.

Finally, the word cloud map is drawn using the visualization method, as shown in Figure 3.

Figure 3. Risk feature item word cloud map.

5. Conclusions

This article mainly studies the safety risk identification of subway shield construction. Firstly, a large number of safety risk texts are collected using the "Subway engineering safety risk early warning system", and 13,129 risk statements are extracted from the texts. Secondly, the risk statements are segmented into Chinese words via Jieba segmentation and the development of a professional thesaurus (security risk thesaurus, custom thesaurus, stop word thesaurus). Then, the TF-IDF algorithm is used to realize the structural transformation of the text, 301 high-frequency words of subway shield construction safety risk are extracted, 30 high-frequency words containing risk semantics are screened out, and the initial list of subway shield construction safety risk factors is established. Finally, a comparative analysis is made with the current national standards and norms to obtain the optimized safety risk factor set of subway shield construction, which contains a total of 29 risk factors. The case text is used to visualize the word cloud map drawn from the Chinese word segmentation results.

Via the analysis of the above research results, this paper believes that the risk of high importance needs to be dealt with first. Therefore, this paper puts forward the following suggestions:

(1) For the general contractor of shield construction, it is very important to formulate the safety management system of shield construction, including the establishment of safety awareness of site managers, the construction safety control of subcontractors, and the establishment of a reasonable site safety feedback mechanism. Via the mining of a large number of risk texts, we find that most of the accident sources are due to the lack of vigilance of the site management personnel to the source of danger, the thought of slack, not realizing the importance of construction safety.

(2) Technical defects of construction personnel also account for a large proportion of shield construction accidents, including unreasonable monitoring schemes, non-standard operation of shield construction machinery and equipment, and improper use of protective equipment by construction personnel. Therefore, the technical training and assessment of construction personnel before construction is very important, and the implementation of regular technical content review can prevent shield construction accidents.

(3) The failure of construction equipment is also one of the reasons for accidents. The safety inspection of subway shield construction equipment is not carried out regularly, or the inspection is not implemented, which will directly lead to the accident. The

inspection of functional damage and defects of the equipment itself before construction can reduce construction site accidents caused by equipment problems.

At present, the safety risk management of the subway shield mainly relies on structured data such as objective monitoring data and subjective expert scores for risk analysis, and insufficient use is made of a large number of risk records retained in the construction process. Therefore, compared with the traditional expert experience, the research results of this paper are more objective and reliable and also have more research significance.

Author Contributions: Conceptualization, C.T.; methodology, C.T. and C.S.; validation, C.T. and C.S.; formal analysis, C.T.; investigation, C.T.; resources, C.T.; data curation, J.Z.; writing—original draft preparation, C.T.; writing—review and editing, C.S.; visualization, J.Z.; supervision, Z.G.; project administration, Z.G.; funding acquisition, Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: Sanya Yazhou Bay Science and Technology City, Grant No: SKJC-2022-PTDX-021.

Data Availability Statement: All data generated or analyzed during the study are available from the corresponding author upon request. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Q.; Wang, P.; Deng, X.; Wu, W. Analysis of safety ability of construction personnel in subway shield collapse accident. *Disaster Sci.* **2010**, *25*, 5. (In Chinese)
2. Chen, H.; Wang, N.; Wang, J. Construction risk assessment of shield tunneling underpass based on triangular fuzzy Cloud Bayesian Network. *Def. Transp. Eng. Technol.* **2020**, *18*, 6. (In Chinese)
3. Yin, Y. Research on Safety Risk Identification of Subway Tunnel Construction with Shield Method Based on SNA. Master's Thesis, Wuhan University of Technology, Wuhan, China, 2020.
4. Liu, Z.; Qin, X. Safety risk assessment of subway shield construction based on Analytic Hierarchy Process. *Constr. Technol.* **2018**, *47*, 1269–1272. (In Chinese)
5. Chung, H.; Lee, I.M.; Jung, J.H.; Park, J. Bayesian Networks-based Shield TBM Risk Management System: Methodology Development and Application. *KSCE J. Civ. Eng.* **2018**, *23*, 452–465. [CrossRef]
6. Nezarat, H.; Sereshki, F.; Ataei, M. Ranking of geological risks in mechanized tunneling by using Fuzzy Analytical Hierarchy Process (FAHP). *Tunn. Undergr. Space Technol.* **2015**, *50*, 358–364. [CrossRef]
7. Yazdani-Chamzini, A. Proposing a new methodology based on fuzzy logic for tunnelling risk assessment. *Statyba* **2014**, *20*, 82–94. [CrossRef]
8. Zhou, C.; Ding, L.Y.; Skibniewski, M.J.; Luo, H.; Zhang, H.T. Data based complex network modeling and analysis of shield tunneling performance in metro construction. *Adv. Eng. Inform.* **2018**, *38*, 168–186. [CrossRef]
9. Xue, Y.; Li, X.; Qiu, D.; Ma, X.; Kong, F.; Qu, C.; Zhao, Y. Stability evaluation for the excavation face of shield tunnel across the Yangtze River by multi-factor analysis. *Geomech. Eng.* **2019**, *19*, 283–293.
10. Ren, H. Risk lifecycle and risk relationships on construction projects. *Int. J. Proj. Manag.* **1994**, *12*, 68–74. [CrossRef]
11. Chen, H.; Li, H.; Hu, X. Influencing Factors and accident causative model of subway shield tunneling safety. *J. Railw. Eng.* **2020**, *37*, 6.
12. Wang, J.; Guo, L.J.; Li, Y.Q. Risk Analysis of a Subway Tunnel Shield Based on a Fuzzy Synthetic Approach. In Proceedings of the International Conference on Electric Information and Control Engineering, Wuhan, China, 15–17 April 2011.
13. Pan, H.Z.; Gou, J.; Wan, Z.; Ren, C.; Chen, M.; Gou, T.; Luo, Z. Research on Coupling Degree Model of Safety Risk System for Tunnel Construction in Subway Shield Zone. *Math. Probl. Eng.* **2019**, *2019*, 5783938. [CrossRef]
14. Cao, L.; Kalinski, M. Risk Analysis of Subway Shield Tunneling. In Proceedings of the Geo-Risk Conference, Denver, CO, USA, 4–7 June 2017.
15. Huang, Z.; Zhang, W.; Sun, H.; Pan, Q.; Zhang, J.; Li, Y. Risk uncertainty analysis in shield tunnel projects. *Tunn. Undergr. Space Technol.* **2023**, *132*, 104899. [CrossRef]
16. Liu, S. Dynamic Safety Risk Assessment Method and Engineering Application of Tunnel Construction Process Based on Data Mining. Master's Thesis, Shandong University, Jinan, China, 2018.
17. Liu, Z. Research on the Causative Mechanism and Prevention Strategies of Subway Construction Safety Accidents. Master's Thesis, China University of Mining and Technology, Xuzhou, China, 2022.
18. Xu, N. Research on Safety Risk Transmission of Urban Rail Transit Construction Projects Based on Data Mining. Master's Thesis, China University of Mining and Technology, Xuzhou, China, 2018.
19. Ji, C. Subway Construction Safety Risk Probability and Cost Control Analysis. Master's Thesis, Zhejiang University, Hangzhou, China, 2022.

20. Son, B.Y.; Lee, E.B. Using Text Mining to Estimate Schedule Delay Risk of 13 Offshore Oil and Gas EPC Case Studies During the Bidding Process. *Energies* **2019**, *12*, 1956. [CrossRef]
21. Li, J. Research on Construction Risk Monitoring of Subway Foundation Pit Engineering Based on Association Rules. Master's Thesis, Tianjin University of Technology, Tianjin, China, 2011.
22. Zhang, S.; Chen, X.W.; Wu, F. Analysis and Application of Automated Data Collection System in Subway. In Proceedings of the 6th Advanced Forum on Transportation of China (AFTC 2010), Beijing, China, 16 October 2010.
23. Hsu, Y.C.; Su, J.S. Responsive passengers' letter system—A case of Taipei subway. In Proceedings of the 12th ISSAT International Conference on Reliability and Quality in Design, Chicago, IL, USA, 3–5 August 2006.
24. Mo, Z.G.; Luo, H.B. Study on the Maintenance of Subway Equipment Based on Data Mining Techniques. In Proceedings of the 2nd International Conference on Smart City and Systems Engineering (ICSCSE), Changsha, China, 11–12 November 2017.
25. Juan, L.; Fang, Y.S.; Jian, H. Analysis of Subway Delay Based on Data Mining. In Proceedings of the 4th IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 12–15 April 2019.
26. Chang, H. Research on Data-Driven Subway Construction Safety Risk Assessment and Response. Master's Thesis, China University of Mining and Technology, Xuzhou, China, 2023. [CrossRef]
27. Mou, Q.; Ding, X.; Liu, Z.; Wu, X. Hazard source identification Algorithm based on Metro operation log text mining. *China Work. Saf. Sci. Technol.* **2022**, *18*, 204–210.
28. Ye, C. Research on Classification and Retrieval of Underground Construction Hidden Danger Text Based on Deep Learning. Master's Thesis, Huazhong University of Science and Technology, Wuhan, China, 2020. (In Chinese).
29. Pan, X.; Zhong, B.; Black, Y. Research on Key points mining and visualization of Subway Construction Safety hazard detection based on LDA. *Inf. Technol. Civ. Eng. Constr.* **2021**, *13*, 7–14.
30. Hei, Y. Analysis and Visualization of Subway Construction Hidden Dangers Based on Text Mining. Master's Thesis, Huazhong University of Science and Technology, Wuhan, China, 2019. (In Chinese).
31. Li, X.; Wang, J.; Xu, N.; Zhou, Z. Analysis of risk factors of subway construction safety accidents based on text mining. *Tunn. Constr.* **2017**, *37*, 160–166. (In Chinese)