WORKING PAPER

ITLS-WP-23-12

A data-driven conceptual framework for understanding the nature of hazards in railway accidents

By
Wei-Ting Hong[a], Geoffrey Clifton[a] and John D Nelson[a]

[a] Institute of Transport and Logistics Studies (ITLS), The University of Sydney, Australia

June 2023

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in
Transport and Logistics Management

The University of Sydney

Established under the Australian Research Council's Key Centre Program.

NUMBER:	Working Paper ITLS-WP-23-12
TITLE:	**A data-driven conceptual framework for understanding the nature of hazards in railway accidents**

ABSTRACT:

Hazards threaten railway safety by their potential to trigger railway accidents. Whilst there are a considerable number of prior works investigating railway hazards, few offer a holistic view of hazards across jurisdictions and time and demonstrate policy implementation due to the inability to analyse a large amount of safety-related textual data. The conceptual framework *HazardMap* is developed to overcome this gap, employing open-sourced Natural Language Processing topic model BERTopic for the automated analysis of textual data from Rail Accident Investigation Branch (RAIB), Australian Transport Safety Bureau (ATSB), National Transportation Safety Board (NTSB) and Transportation Safety Board of Canada (TSB) railway accident reports. The topic modelling depicts the relationships between hazards, railway accidents and investigator recommendations and is further extended and integrated with the existing risk theory and epidemiological accident models. Results show that each hazard in the railway system has different aspects and could trigger a railway accident when combined with other hazards. Each aspect can be partially or fully addressed by implementing hazard mitigation policies such as introducing new technologies or regulations. A case study of the application to the risk at level crossings is provided to illustrate how *HazardMap* works with real-world data. This demonstrates a high degree of coverage within the existing risk management system, indicating the capability of helping policymaking for managing risks with adequate accuracy. The primary contributions of the framework proposed are to enable a huge amount of knowledge accumulated for an intuitive policymaking process to be summarised, and to allow other railway investigators to leverage lessons learnt across jurisdictions and time with limited human intervention. Future research could incorporate data from road, aviation or maritime accidents.

KEY WORDS: ***Hazards analysis; Railway accident; Natural Language Processing (NLP); Implementation; data-driven framework***

AUTHORS: **Hong, Clifton, Nelson**

ACKNOWLEDGEMENTS: N/A

CONTACT:

INSTITUTE OF TRANSPORT AND LOGISTICS STUDIES (H04)

The Australian Key Centre in Transport and Logistics Management
The University of Sydney NSW 2006 Australia
Telephone: +612 9114 1813

	E-mail:	business.itlsinfo@sydney.edu.au
	Internet:	http://sydney.edu.au/business/itls

DATE:	June 2023

1 Introduction

The nature of railway accidents has drawn a considerable amount of attention in recent years. A variety of theories and frameworks are proposed in the literature to understand the mechanism of railway accidents from epidemiological (Peters et al., 2018), systemic (Read et al., 2021; Santos-Reyes &Beard, 2009), causation and sequencing (Wullems et al., 2013; Xia et al., 2012), and barrier of energy (Huang et al., 2020) perspectives. Many of them have been widely used in the railway industry and by railway accident investigation bodies. For instance, the Root Cause Analysis (RCA), the Accident Causation Model, and the Systems Theory are commonly used during the railway accident investigation to identify the causal relations between (underlying) factors during a railway accident (ATSB, 2009; Dai &Wang, 2010; Kinnersley &Roelen, 2007; RAIB, 2008).

In recent years, a growing focus has been on mitigating the limitation of the number of cases of railway accidents analysed before establishing a theory or framework. Whilst it is commonly agreed that a railway accident involves many perspectives given that the railway system is characterised by high interactive complexity (Li et al., 2019; Read et al., 2021), extending the number of cases analysed can be difficult due to the complicated nature of factors in each railway system and the limited capacity of analysis conducted by humans. Previous research has enabled an increase in the number of railway accidents analysed for the purpose of comparing the causal relations between accidents analysed by predefining a series of labels and manually reviewing railway accident reports (Kim &Yoon, 2013; Zhou &Lei, 2018). However, the findings cannot be further extended to railway accidents in other jurisdictions or be updated after new railway accident reports are published.

To overcome such obstacles, some prior works have explored leveraging the benefits of Natural Language Processing (NLP) and machine learning to consistently analyse a large body of textual data. NLP addresses the interface between human languages and computers by enabling the computer program to process a large amount of textual data through machine learning approaches. Several attempts to incorporate NLP into accident data analysis can be found in the context of maritime, aviation and road safety for the analysis of crowdsourced textual data (Kinra et al., 2020; Nelson et al., 2020; Syeda et al., 2019). Despite the extensive discussion of (semi-)automated textual data analysis in the literature, the focus is mainly on building the NLP model rather than interpreting the result. Additionally, most studies in this context utilise the supervised learning approach, requiring a significant amount of manual effort for training the model (Sizov &Öztürk, 2013; Wang et al., 2017). These limitations hinder researchers and practitioners from advancing the existing railway safety knowledge with the help of novel technologies.

The research objective of this study is to provide a holistic view of the nature of hazards in railway accidents across jurisdictions and across time by leveraging the power of NLP with little manual effort. Instead of establishing a customised model, this study only utilises open-sourced and off-the-shelf toolkits for building the NLP model so that the result and contribution of this study can be duplicated and reused. Therefore, the data-driven framework *HazardMap* is proposed to offer another view on railway accidents from the hazard-centred perspective. The result is considered beneficial for researchers and practitioners in advancing railway safety knowledge by enabling learning across jurisdictions and across time.

This paper is organised as follows. Firstly, a brief review of literature on existing frameworks and theories on railway accident analysis is conducted (Section 2). Subsequently, the NLP model design for semi-automated hazards analysis on railway accidents and Python API-based toolkits and models are introduced (Section 3). Next, the *HazardMap* is proposed based on the findings of the outcome (Section 4) and a case study of the analysis across jurisdictions and across time is offered (Section 5). Lastly, conclusions, suggestions for further works and limitations are elaborated (Section 6).

2 Literature context

Railway safety research can be divided into several categories based on the purpose of the analysis. For example, railway safety can be treated as a risk issue and addressed from the perspective of risk sources, likelihood and consequences (Liang et al., 2020; Parkinson et al., 2016; Yang &Li, 2020). On the other hand, the classical energy-barrier model argues that introducing barriers in practical safety management can prevent energies from impacting vulnerable targets (Braut et al., 2014; Syeda et al., 2019; Zhou &Ding, 2017). Regardless of various research purposes, identifying hazards that might potentially cause negative consequences is one of the most critical tasks prior to any other risk-relevant analysis (Rausand, 2013). Hazards in the railway system can be classified on the basis of the application of interests. For instance, classifying based on the main contributor such as technological hazards and natural hazards is popularly used in the literature (Ouyang et al., 2010; Rydstedt Nyman &Johansson, 2015; Whittingham, 2012). Some studies also define a set of hazards as endogenous hazards and exclude others as exogenous hazards for the research object of interest (Hulin et al., 2016; Li et al., 2021).

There has been a considerable number of methods for identifying hazards applied in the railway context, which can broadly be categorised in accordance with data-driven approaches, conceptual frameworks and systemic analysis. The data-driven approach concentrates on extracting hazards in the accident causation network (Bíl et al., 2017; Li et al., 2021; Zhang et al., 2021) or manually identifying hazards with proposed methods, such as the preliminary hazard analysis (PHA) (Guenab et al., 2008; Yan &Xu, 2019) and the failure modes, effects, and criticality analysis (FMECA) (Catelani et al., 2021; Ciani et al., 2019). On the other hand, conceptual frameworks attempt to capture critical insights from in-depth analysis of case studies (Bang et al., 2020; Li et al., 2021) or brainstorming (Berrado et al., 2010; Runyan &Yonas, 2008) to develop general methods and frameworks for further application. For instance, the Hazard and Operability (HAZOP) analysis has been adopted to identify potential deviations and undesired situations in railway operations (Bian &Wang, 2015; Li et al., 2015; Papen et al., 2011). Lastly, the systematic analysis considers the railway industry as a complex system and treats hazards as constructed issues that need to be addressed from the hierarchical control perspective (Ouyang et al., 2010; Zhang et al., 2021). The occurrence of railway accidents is the consequence of inadequate control which might be effected by individuals, organisations or systems (Rausand, 2013). Therefore, the emphasis of the systematic analysis is on inadequate controls resulting in hazards. Some methods such as the systems-theoretic accident model (STPA) (Gong &Li, 2018; Ouyang et al., 2010; Song et al., 2012) and the hierarchical socio-technical framework (Accou &Carpinelli, 2022; Akel et al., 2022; Ryan et al., 2021) are popularly used in the analysis of railway hazards.

Despite a wide discussion of applied hazard identification approaches in the literature, most of them suffer from the limited number of cases analysed given that most railway hazard-related documents are recorded in text form and manually reviewing textual data is extremely time-consuming (Rosadini et al., 2017). Therefore, growing attention has been drawn to the application of Natural Language Processing (NLP) to automating crowdsourced textual analysis. NLP is a technique for extracting knowledge of interest from unstructured text by enabling computers to process vast amounts of text (Ly et al., 2020; Marquez et al., 2000). The implementation of the NLP in the context of railway hazard analysis is limited. Most works focus on hazards classification (Dong et al., 2022; Hughes et al., 2018; Liu et al., 2022) and extracting hazards from textual data (An et al., 2013; Hadj-Mabrouk, 2019; Huang et al., 2022). The main contribution of these studies is offering an opportunity to enhance railway safety management by taking a holistic view of railway accidents across time (Hua et al., 2019; Syeda et al., 2019). However, several barriers to utilising the NLP hinder further application to the railway hazards analysis. For instance, most NLP models are customised for specific purposes and can only be used on certain datasets, making it extremely difficult to be extended in future works (Alawad et al., 2020; Dong et al., 2022; Lee et al., 2021). Additionally, the overreliance on the supervised learning approached is observed in the context of railway accident analysis which might restrict the insight into pre-determined dimensions and overlook other potential factors. Annotating data for the training process also requires intensive labour and human intervention, implying a potential obstacle to practical applications (Yang et al., 2022). Despite attempts to utilise unsupervised learning approaches, previous studies suffer from systematically interpretating the outcomes, resulting in a barrier for practitioners to implement to real-world operations (Bougacha et al., 2019; Lasisi &Attoh-Okine, 2020). This represents a gap in the literature which the present study seeks to address.

To sum up, railway hazard identification has been discussed in the literature for a long time. A potential shift from manual approaches to NLP-based approaches is observed. Enlarging the amount of data analysed for a comprehensive perspective of hazards and incorporating it with existing frameworks and theories are prioritised in recent studies (Dong et al., 2022; Hua et al., 2019; Syeda et al., 2019). Additionally, a growing number of studies has examined the potential of textual big data analysis for public policy decision-making (Bai et al., 2021; Kinra et al., 2020). However, the major challenge of the current implementation is the absence of a general model allowing a wide range of data sources and the need to train the NLP model with limited human intervention. Such an obstacle hinders practitioners and researchers from reusing models and frameworks proposed in the literature for further implementation. Therefore, the need for a generalised and semi-automated analysis framework applicable to most textual data is required to overcome such a research gap.

3 Semi-automated hazard analysis of railway accidents

To overcome the limitations mentioned above, this study develops a framework for describing the nature of hazards in railway accidents by leveraging open-sourced and publicly available Python API-based toolkits to build the NLP model. The topic modelling is applied to explore potential hazards from the thematic structure of textual data (Hristova &Netov, 2022). The extracted topics are further processed to re-construct the relations between hazards via proposed post-processing procedures.

3.1 Topic modelling

Topic modelling is a practical application in information retrieval and NLP to categorise text into domain topics and rank documents by topics (Bai et al., 2021; Dornick et al., 2021; Roque et al., 2019). A topic model reveals the relationship between topics and documents by exploring different features, such as the probability of occurrence of words and high dimensional word embeddings. The model assumes that a document contains a collection of underlying themes, and the distribution of words in the document over the whole corpus might derive topics representing these underlying themes. A set of keywords is identified to reflect underlying topics and their trend, which is useful for further methodological and practical applications (Blei and Mcauliffe, 2007).

A topic model can be trained in several ways, including supervised learning, semi-supervised learning and unsupervised learning. To ensure highly automated analysis and avoid human intervention during data analysis, unsupervised learning approaches are selected for building the topic model in this study. A considerable number of off-the-shelf programming packages (Python API-based toolkits) for advanced NLP applications are developed and publicly available. These include Spacy for deep learning workflows and pre-trained language models (Choi et al., 2015; Jugran et al., 2021), Stanford NLP for toolkits used in developing extensible pipeline and pre-trained models (Manning et al., 2014), and NLTK for a wide range of libraries to implement NLP tasks (Bird and Loper, 2004). Several package-oriented programming models have been developed based on these packages and the state-of-the-art technologies result in significant improvements of performance in the topic modelling contexts. Among existing topic modelling approaches, the BERTopic has demonstrated a better performance in the human language understanding and offering robust and interpretable results (Grootendorst, 2022; Hristova &Netov, 2022). The following sections elaborate on the details and applications of the BERTopic models.

3.2 BERTopic

The BERTopic is a topic model adopting the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) pre-trained language model (Devlin et al., 2018) to retrieve high-dimension vectors of texts for clustering, which has been proven to provide better performance on several NLP tasks (Devlin et al., 2018; Dornick et al., 2021). For implementation topics are generated through three steps: text vectorisation with a pre-trained language model, dimension reduction for optimising the modelling process, and topic representations with custom class-based TF-IDF (c-TF-IDF). The c-TF-IDF is an advanced method for converting original text into a series of representative numbers (which is also known as word embedding). In contrast to traditional approaches, the c-TF-IDF takes the semantic relationships between words into account, increasing the interpretability and accuracy of the outcomes.

For the text vectorisation with a pre-trained language model, documents in the corpus are embedded in vector space with high dimensions, allowing semantical comparisons. For instance, sentences such as "The train stops before the signal." and "The train fails to stop before the signal." will have a longer semantical distance in vector space than the representation created by the bag-of-words approach. The Sentence-BERT (SBERT) framework (Reimers and Gurevych, 2019) is used to convert texts into dense vector representations, which has been commonly applied to NLP tasks and high performance achieved (Ganesh et al., 2020; Labusch and Neudecker, 2020; Ly et al., 2020). The author of BERTopic also states that the language model used in the BERTopic is exchangeable so that the performance can be continuously improved through the development of NLP techniques (Devlin et al., 2018).

Once the dense vectors are generated, the spatial distance between data becomes less meaningful due to the multidimensions of local and global features. Therefore, the Uniform Manifold Approximation and Projection (UMAP) technique is introduced to reduce the dimensionality by projecting vectors to lower dimensional space (McInnes et al., 2018). Subsequently, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is used to cluster vectors in lower dimensional space (McInnes et al., 2017). The advantage of HDBSCAN is allowing noise to be modelled as outliers, avoiding unrelated documents being sorted to topics and influencing the representations of topics. The clustering approach can be replaced by other algorithms in the interest of accuracy and computational time.

Lastly, each identified cluster is assigned to one topic with a distribution of keywords. To highlight the difference between clusters, the custom class-based TF-IDF (c-TF-IDF) is used to rank keywords by the combination of Term Frequency (TF), and Inverse Document Frequency (IDF) (Devlin et al., 2018; Hakim et al., 2014). The weight of a term (t) over documents sorted to a topic (c) can be expressed as

$$W_{t,c} = tf_{t,c} \times \log\left(1 + \frac{A}{tf_t}\right) \qquad \text{Equation 1,}$$

where tf is the term frequency and A is the average number of keywords per topic. The output reflects the importance of a term in one topic rather than in one document, allowing us to understand the distributions of keywords on each topic. Furthermore, Equation 1 can be extended for dynamic topic modelling to reflect the evolution of topics over time. For instance, a topic relating to “over speeding” and “SPAD” can be found across the corpus, but the term “Eurotunnel” might not be found in documents before 1994. Such variance has been mixed, hindering researchers from understanding the temporal effect of “over speeding” and “SPAD” on the term “Eurotunnel”. To overcome such difficulties, Devlin et al. (2018) modifies the calculation of the weight $W_{t,c}$ by creating a local temporal representation at time i with the original equation.

$$W_{t,c,i} = tf_{t,c,i} \times \log\left(1 + \frac{A}{tf_t}\right) \qquad \text{Equation 2,}$$

Equation 2 adds an additional dimension to the weight and enables the representation of local variables without modifying the parameters of the trained BERTopic model and clustered documents. An overview of processes for establishing a BERTopic model is illustrated in Figure 1. Additional mathematical descriptions and details of Python API interfaces are given in Devlin et al. (2018).

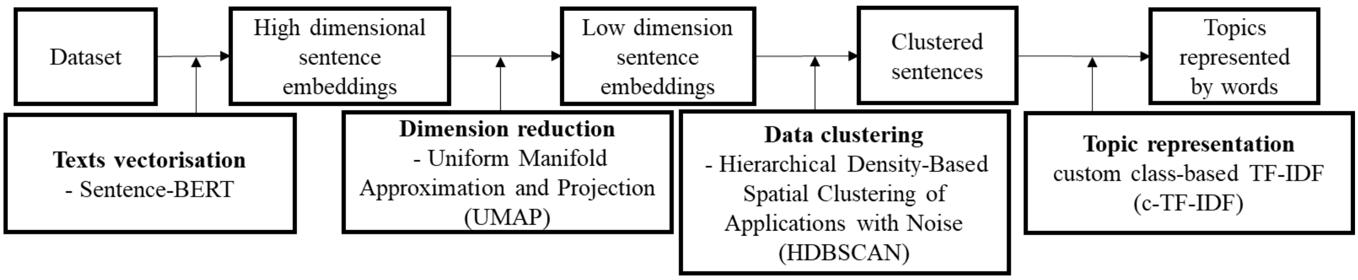


Figure 1, the overview of workflows for developing a BERTopic model

3.3 Dataset

Multiple data resources are applied in this study. However, the difference in language used in railway systems in recording accidents, culture, and regulations can make the analysis inconsistent and create significant bias. Hence, in this study only countries where the investigation bodies exhibit the following features are considered:

1) The investigator must have access to a comprehensive documentation system to reduce the complexity of processing. In other words, the framework of the accident report must be clear and consistent in the temporal aspect (for instance, the jurisdiction has a law or regulation on the format of generating accident reports).

2) The investigator must have been granted the independent authority to conduct the investigation. Furthermore, the investigation objective should aim to increase railway safety regardless of blame or liability.

3) The reports conducted by the investigator must contain recommendations which focus on issues relating to railway safety, such as the implementation of specific training or policy, introducing new technology, or revising existing standard operating procedures. Note that the recommendations must not contain inference or conclusion of apportioning liability.

4) This study only considers data from native English-speaking countries to secure the performance of the model, which implies that reports must be written in English. Additionally, the English language used in reports should be consistent regardless of time, the types of accident, or investigation engagement (for instance, the definition of derailment in each jurisdiction would not vary over time).

5) This study only considers investigating bodies which have published over 100 reports to ensure the performance of the model.

Based on these requirements, railway accident reports published by independent railway accident investigation bodies from the UK, the USA, Canada and Australia are used. Railway accident reports compiled by independent railway accident investigation organisations are regulated by a national framework and provide unbiased and blame-free details for promoting a railway safety culture. Despite the differences in writing styles and terminology used, all reports consist of the summary of the accident, the analysis, the investigation, key findings, conclusions and recommendations (if applicable). The database provided by investigators covers various periods of time. For the best understanding of railway accident knowledge, all retrievable railway accident reports in PDF format from the official websites of countries included are retrieved. Data from the ATSB and TSB is retrieved from websites directly because the full text is provided and crawlable via HTML. Scanned files are removed due to the technical difficulties of recognising the text converting it into an editable form.

Table 1 shows the overview of the processed railway accident dataset. The RAIB has published a series of review reports, such as the *"Investigation into the safety of automatic open level crossings on Network Rail's managed infrastructure"* (RAIB, 2011). These reviews are overlapped with published reports and are excluded from the dataset. The ATSB and TSB provide the full text of railway accident reports on the website which are retrieved directly. Despite the availability of early reports (pre-1996) published by the NTSB, only scanned files are retrievable and these are excluded from the dataset.

Table 1, the overview of the processed railway accident dataset

	No. of reports	No. of sentences	Period	Note
RAIB	339	124,990	2005-2019	Review reports are removed.
ATSB	250	84,679	1999-2021	Reports are retrieved from websites directly.
NTSB	274	92,406	1996-2021	Reports earlier than 1996 are scanned files.
TSB	415	104,720	1993-2021	Reports are retrieved from websites directly.

3.4 Descriptive results from applying the topic model

Table 2 presents the description of the top 5 topics with the highest occurrences in each dataset. An exhaustive description of the top 50 topics of each dataset is shown in Appendix A. The name of each topic is assigned in accordance with the representative terms identified. The result indicates potential similarities and differences between railway accidents analysed in each jurisdiction. It can be seen that the risk factor "fatigue" has been widely discussed and examined in the countries investigated. In contrast, the emphasis of each investigator in their reporting is slightly different. For instance, the NTSB concentrates on the statistical evidence collected from physical devices, such as the event recorder and toxicology test. On the other hand, the RAIB seems to devote much effort to discovering potential underlying factors, such as driver knowledge and communication. However, the interpretation would be extremely limited by only analysing topics with the highest occurrences. Without further analysis the relationship between each topic cannot be revealed, hindering users from understanding the mechanisms of railway accidents. Therefore, additional modifications are required to extend the result for a holistic view of the nature of railway accidents across countries.

Table 2, the description of the top 5 topics with the highest occurrences of the possibility of each dataset

	RAIB	ATSB	NTSB	TSB
Topic 0*	Trams	Fatigue	Emergency response	Fatigue
Topic 1	Fatigue	Sounded horn/ audibility	Drug, alcohol, and toxicology test	Emergency brake
Topic 2	Communication	Train speed	Event, audio and image recorder	Warning devices at grade crossings
Topic 3	RRV (Road Rail Vehicles)	Queensland Rail	Positive Train Control	Derailment of freight cars
Topic 4	Driver knowledge, training, instruction	Network Control Office (NCO)	Conditions of switches	The interface between controllers and crew

*The Python indexing system starts with 0.

4 The development of the general framework

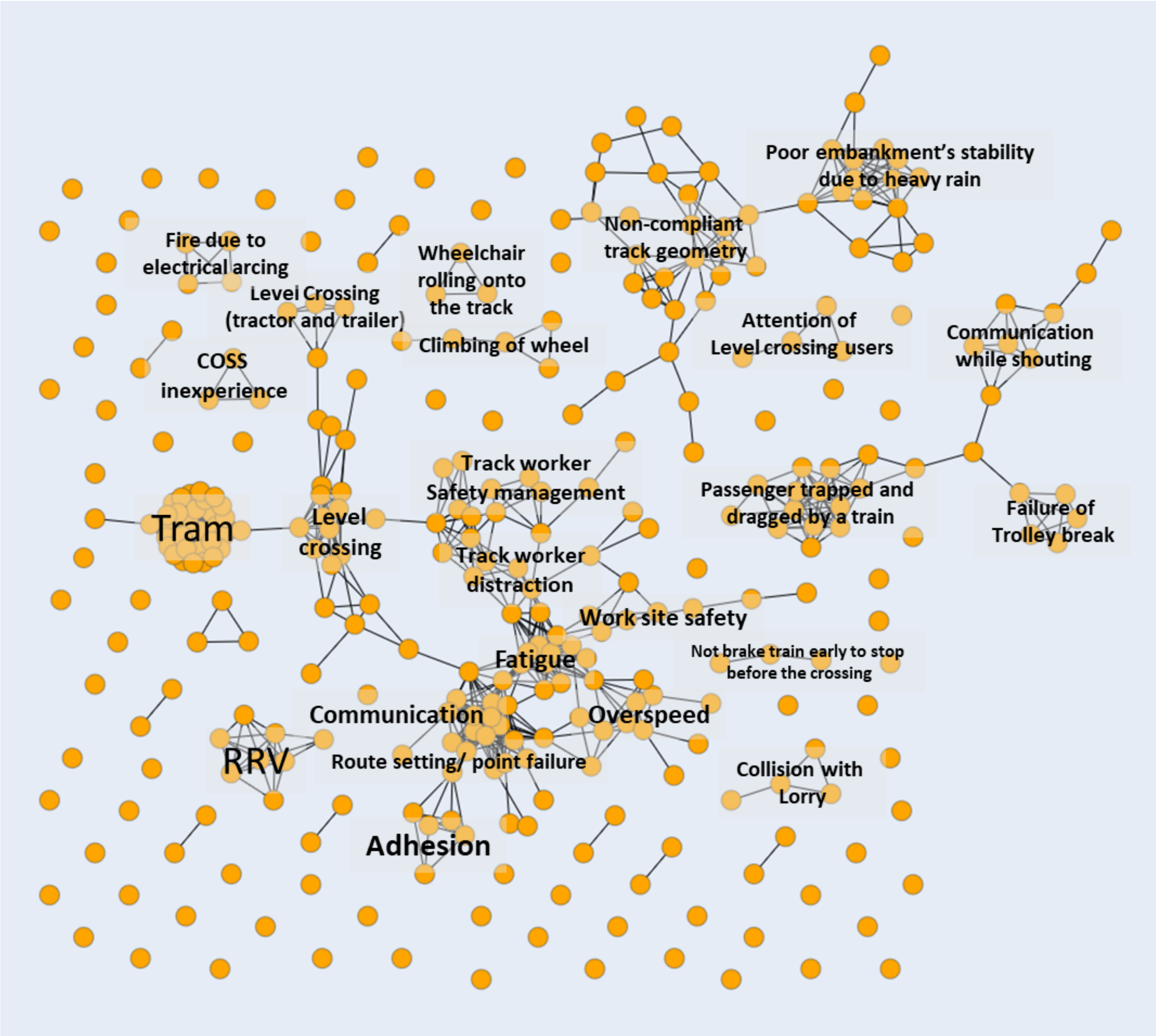
In traditional hazard identification approaches, manual review and analysis are the main methods to propose the framework depicting the nature of hazards identified in a particular case. However, there is no general model allowing a wide range of data sources which significantly limits the capability for reusing models and frameworks proposed in the literature for further implementation (Section 2). As demonstrated above (Section 3), the result retrieved from the NLP model is difficult to be interpreted individually without understanding the relationship between each topic. To address these issues, the *HazardMap* framework is proposed to extend the result of the topic model and existing theories in the literature. This section describes the development of the framework and a case study application is given in Section 5.

4.1 The topic model results

Firstly, the distribution of the number of sentences over each topic on documents is extracted and condensed to a topic-document matrix. Secondly, we assume that the distribution of each topic over documents is the projection of the extent to which this topic influences each railway accident. Multiple similar distributions indicate that these topics constitute a specific group of railway accidents with similar features. Therefore, the cosine similarity approach is applied to identify the similarity of distributions (Cheng et al., 2009; Qurashi et al., 2020). A topic-topic similarity matrix can be generated with each element between 0 and 1. The larger similarity score indicates that sentences under both topics are commonly used in the same group of documents.

Next, a distribution of topics including the relationship can be mapped by setting a threshold for the similarity score, linking each topic and forming a series of clusters representing various hazards. The threshold is determined by the researcher based on the nature of input data and analysis purposes. A higher threshold leads to scarce links between topics and forms a limited number of small clusters, whereas a lower threshold results in dense connections between topics and several large clusters containing almost all topics. Therefore, the threshold needs to be carefully determined by reviewing each outcome with a different similarity score.

This study uses the RAIB dataset as an example for demonstrating the application. The threshold for the RAIB dataset is set as 0.5 due to the appropriate balance between the number of clustered groups and the well-distributed hazards. Figure 2 shows the distribution of the relationship between hazards identified in the RAIB dataset. Each orange dot represents a topic identified by the BERTopic and the link refers to the similarity score of two topics that is larger than the threshold. The name assigned to each cluster is based on the inference of keywords of linked topics. According to this result, more potential hazards are identified compared with the interpretation of topics having high possibilities of occurrences. The connection between topics is also revealed to illustrate the underlying causal relations in the hazard group. In addition, the cross-country analysis becomes applicable by extracting the hazard of interests from different countries and comparing the mechanisms and causal relations.



Note: Each orange dot represents a topic identified by the BERTopic and the link refers to the similarity score of two topics that is larger than the threshold of 0.5 for the similarity score. The distance between dots is arbitrary.

Figure 2, the distribution of the relationship between hazards identified in the RAIB dataset

It is recognised that a similar accident may still occur even though recommendations made are adopted by the railway industry. Such a situation might not be directly related to the issue of how hazards are addressed but rather the way hazards are interpreted, indicating the need for a re-interpretation of hazards in the railway safety context. Additionally, the distribution of the relationship between hazards constitutes each cluster by aggregating connected hazards, indicating the nature of the complexity of a hazard. Therefore, the hazard should be interpreted by elements involved rather than the hazard itself given that it would trigger another accident in combination with other hazards or in other dimensions.

It is assumed that each hazard in the railway system is revealed in the form of accidents. From the analysis, it seems that a hazard has multiple aspects that result in different types of accidents. Therefore, a previously addressed hazard might appear again after combination with others, implying one hazard will never be fully addressed. However, reducing the impact caused by proposing appropriate recommendations toward accidents by revealing aspects of one hazard is still beneficial for improving railway safety.

Based on this description, we can conclude that each hazard has almost infinite aspects. For example, one aspect might result in an accident with the combination of other hazards and under specific conditions. Once the accident occurred, the impact would disrupt the railway system and recommendations are proposed to address the triggered aspect of this hazard by the investigator. After several occurrences of accidents and all controllable aspects have been addressed, this hazard is considered to be mitigated to the lowest possible level.

4.2 Theoretical basis

The concept of hazard is derived from the risk theory depicting potential threats that might cause harm. A hazard would not cause harm until it is out of control or triggered by other hazards or external factors (Rausand, 2013). The theory of hazards has been extended with other theories; for example, the domino effect uses the concept of hazards to illustrate accident scenarios and escalating situations (Gonzva et al., 2017). The modern analysis of hazards in the railway context focuses on understanding the mechanisms of hazards in the socio-technical system (Akel et al., 2022; Gong &Li, 2018; Ouyang et al., 2010). For the hazard identification processes, the sources can be classified as brainstorming, functional approaches and empirical analysis.

The brainstorming method aims to identify hazards in a system by retrieving knowledge or experience from experts via interviews, workshop or discussion (Berrado et al., 2010).The functional approach emphasises the structure of systems and understands the system as a hierarchy control system (Li &Liu, 2021; Li et al., 2015). The sources of hazards identified by functional approaches come from recognising unsafe interactions between factors even though the harm is not triggered. Similar to the brainstorming method, the process involves the engagement of experts in addition to the inferring process to identify underlying hazards at the organisational level (Li et al., 2019; Madigan et al., 2016). Lastly, the empirical analysis identifies hazards through in-depth case studies with existing frameworks or theories. A holistic view of hazards can be extracted by collecting the data relating to an investigation, evidence and analysis and the knowledge and experience by discussing with experts (Holmgren, 2006; Li et al., 2019; Zhan et al., 2017).

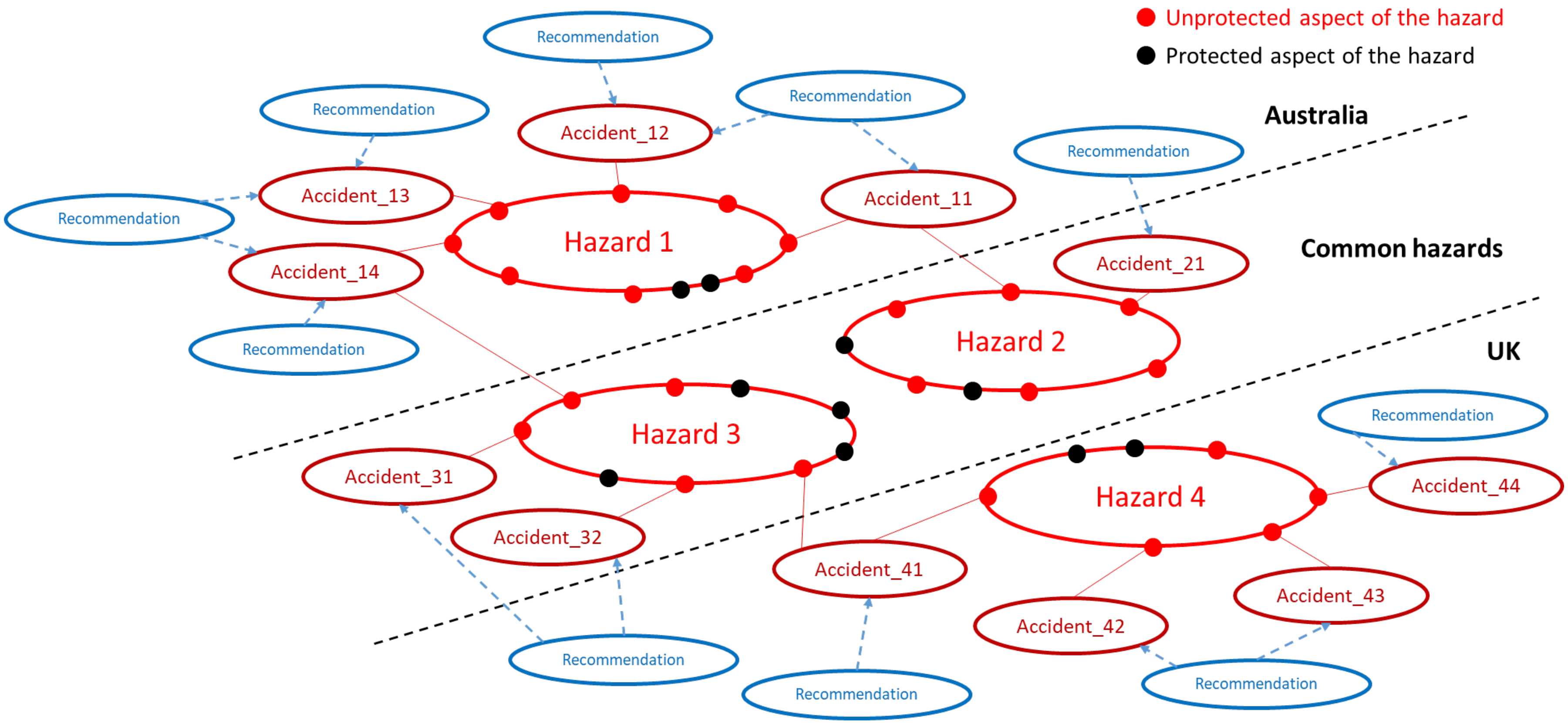
Therefore, the sources for identifying hazards are mainly professional knowledge, practice experience, analysis, and real-world recorded data, primarily in textual and statistical forms. By understanding the potential path and learning from accidents, the hazard can be managed to prevent similar accidents from occurring again. On the other hand, a hazard can derive a higher possibility of causing harm over time through the development of technology, the introduction of a new system or the change of legislation, all of which are difficult to be foreseen. Therefore, each hazard can trigger an accident in combination with other factors by different aspects. For example, the human factor “fatigue” might cause a railway accident due to transient, cumulative, and circadian sleep disorders (Fan &Smith, 2019). Each type can be considered as an aspect of the “fatigue” hazard and is discovered and managed in different ways. There might be another aspect belonging to the “fatigue” hazard existing in the railway system but not yet recognised. Despite the difficulty of observing unforeseen aspects of a hazard, recording and updating aspects found in historical railway accidents are critical to ensure known hazards are well managed and support decision-making when designing a new railway infrastructure project.

4.3 HazardMap

Figure 3 illustrates the conceptual framework *HazardMap*, inspired by the result of topic modelling (Figure 2). The *HazardMap* is a data-driven and epidemiological factor-based framework, looking at railway accidents from a hazard-centred perspective. Hazards illustrated in the *HazardMap* are derived from clusters of hazards, for example, the level crossing in Figure 2.

In Figure 3 each hazard has a series of aspects illustrated as the outline of the oval consisting of continuous dots. Two types of dots surround the hazard: the unprotected aspect (coloured in red) and the protected aspect (coloured in black). The unprotected aspect refers to the potential possibility that this hazard triggers a railway accident under specific conditions or in combination with other hazards. The unprotected aspect might not be identified until it triggers an accident or preventative implementation is placed in advance. The protected aspect represents the hazard that would no longer trigger an accident from this dimension because it has been identified and fully addressed by introducing permanent solutions, such as applying state-of-the-art technology or improving relevant processes. Note that any implementation of new policies or strategies might result in another hazard whilst fully addressing an aspect of a hazard.

Once a hazard triggers a railway accident (with the combination of other hazards or factors), the aspect would be highlighted in the *HazardMap* and connected to the triggered railway accident. Multiple aspects of hazards might trigger some railway accidents; for instance, accident 11 is triggered by aspects of hazard 1 and hazard 2. Subsequently, railway accident investigators would investigate and propose recommendations to address identified aspects of the specific hazard, aiming to prevent similar railway accidents from occurring again (converting red aspects into black). Some recommendations might also address hazards identified by previous railway accidents and reinforce the prevention of hazards, which is illustrated as multiple arrows toward different accidents in the *HazardMap*.



Note: taking Accident_11, Hazard 1 and Hazard 2 as an example. Accident 11 is triggered by aspects (red dots) of Hazard 1 and Hazard 2 and recommendations are subsequently made to address aspects of Hazard 1 and Hazard 2 identified in Accident 11. The similar recommendation made after Accident 12 to address another aspect of Hazard 1.

Figure 3, the relations between hazards, accidents, and recommendations across countries (HazardMap)

Hazards are further categorised based on countries. Some hazards can only be found in specific areas, such as the fall of autumn leaves in the UK and the high temperatures hazards in Australia. Additionally, hazards that can be classified into more than one country are considered common hazards. Different aspects of these common hazards might reach the country-specific area and trigger a railway accident. Note that the locations of hazards on the *HazardMap* might move from one area to another to reflect the change in environment. For example, the high-temperature hazard might impact the UK railway system due to severe climate change. In this case, the high-temperature hazard might move from Australia to the common hazard area.

5 Policy implementation - a case study of the risk at level crossings

Level crossing accidents have been widely discussed in the literature and have influenced railway safety significantly for a long time (Adeolu et al., 2016; Blaho et al., 2020; Jonsson et al., 2019; Liang et al., 2018; Salmon et al., 2013). However, cross-country analysis based on policy implementation is seldom found in the literature. This case study provides an example of how the analysis process can be semi-automated and how the *HazardMap* is generated for a comprehensive view of the way that hazards relevant to level crossing accidents impact the railway system across the countries investigated.

Firstly, the *HazardMap* of each country for level crossing accidents is identified by developing the distribution of the relationship between hazards derived from the BERTopic model. Next, the threshold of covariance is determined based on manual review of each distribution with different thresholds of covariance, which is set to 0.5 for the RAIB, NTSB and TSB datasets and 0.7 for the ATSB dataset. Once the distribution is generated, clusters relevant to the level-crossing hazards are extracted manually. It is suggested that relevant topics are searched to identify the initial network by starting with the top-50 topics. The network is further extended by looking at each document's topic distribution in the initial network. Note that the network of interest might be connected to other clusters. Therefore, the boundary is required to be manually identified in case of including irrelevant topics.

At this step, the threshold of the mentioning rate of topics needs to be set to determine whether one document belongs to this network. A higher threshold results in a smaller number of selected documents with higher confidence of relativity and vice versa. To determine the best threshold of topic mentioning rate for each dataset, an initial rate can be set and documents with a mentioning rate close to the threshold should be manually reviewed. The threshold can be enlarged once most reviewed documents are irrelevant to the topic of interest and vice versa. Once relevant documents are retrieved, additional topics of interest can be further extracted for extending the network by reviewing dominant topics in documents.

An example of the extracted network for level-crossing hazards in the NTSB dataset is demonstrated (Figure 4). The topic selected for identifying the initial network is topic 10: private crossings (refer to Table A- 3 in the Appendix). Subsequently, the boundary is set after reviewing the relevance of topics on the edge as the initial network is connected to other clusters. Next, an initial threshold of the topic mentioning rate is set to 20%, and documents on the edge are reviewed. A final threshold is set to 10% and 36 documents are identified and labelled as level crossing (LC)-related incidents. Additionally, another relevant cluster containing two topics (topics 85 and 225) is recognised as well after reviewing dominant topics in documents retrieved from the initial network.

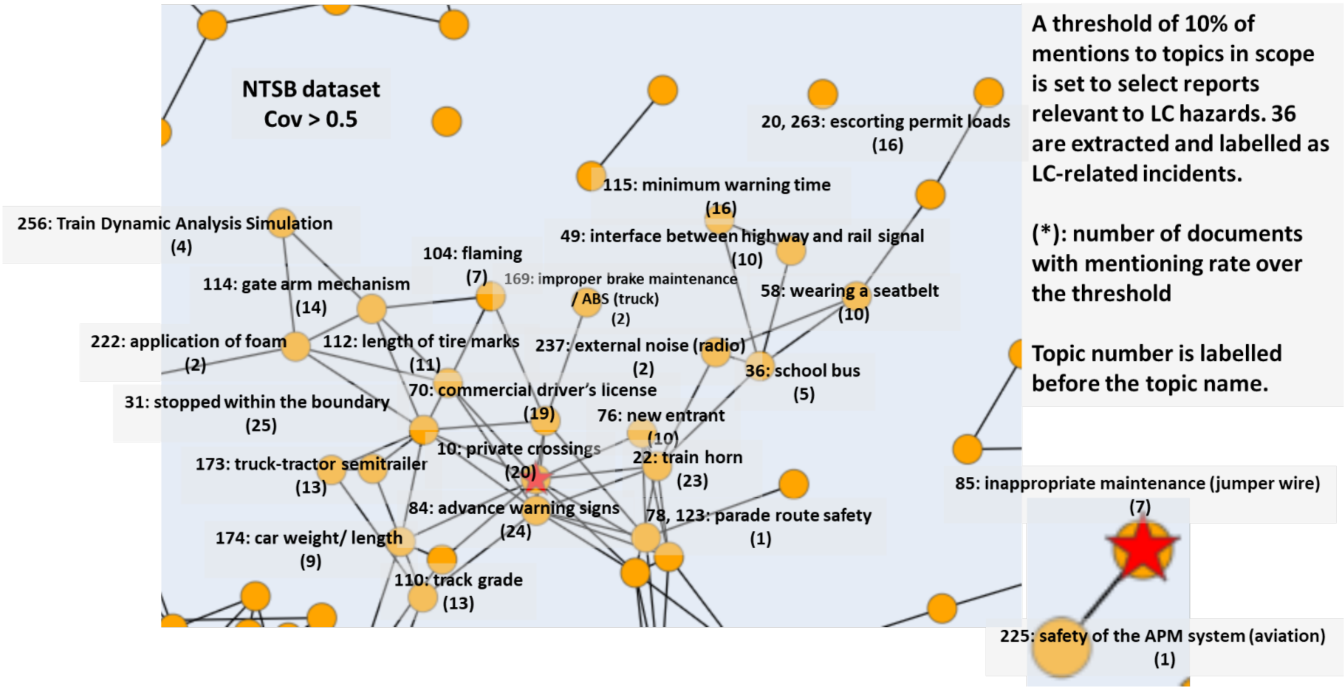


Figure 4, the distribution of topics and their relationship relevant to level crossing accidents in accident reports published by the NTSB (1996-2021)

After establishing the distribution of topics and their relationship relevant to level crossing accidents for each country, heterogenous terminology used in each country is identified and standardised by manual review. For instance, terms “level crossing” and “grade crossing” are linked to and presented as the same concept of “level crossing”. Topics with names standardised from the countries investigated are clustered again based on the characteristics of hazards. Lastly, the *HazardMap* can be created by plotting hazards from each country with different colours for representations.

There are ten main hazards identified in *HazardMap* relevant to level crossing accidents investigated by national railway accident investigators, namely level crossings design (Figure 5), human factor, types of users, types of level crossings, external hazards, maintenance, and others (Figure 6), policy/ management, employee training, and level crossing (LC) users education (Figure 7). Aspects of each hazard are coloured in accordance with identified countries. Overall, a significant number of aspects are observed in several hazards, including road signs, road users and policy/ management. The RAIB and ATSB cover almost all aspects of hazards, and much emphasis is placed on human factors by the NTSB. On the other hand, the TSB concentrates on types of users but overlooks the design of signs on the road and rail. Additionally, the NTSB has also investigated several potential behaviours of road users, such as stopping within the boundary and the regulation of users, whereas the ATSB focuses on the potential impact brought by the design of road signs and the condition of sighting distance. Thus, the difference in the approach that each country addresses level crossing hazards between countries can be explored.

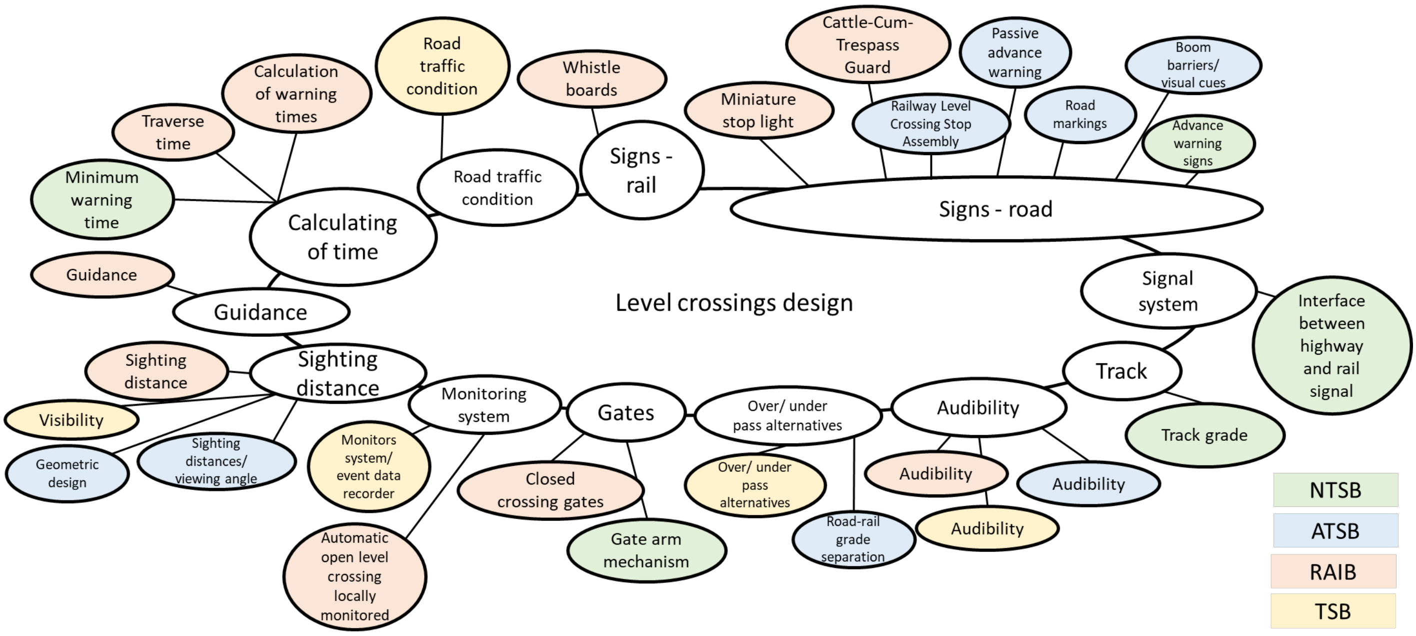


Figure 5, the applied HazardMap on level crossing accidents from all investigators – level crossing design

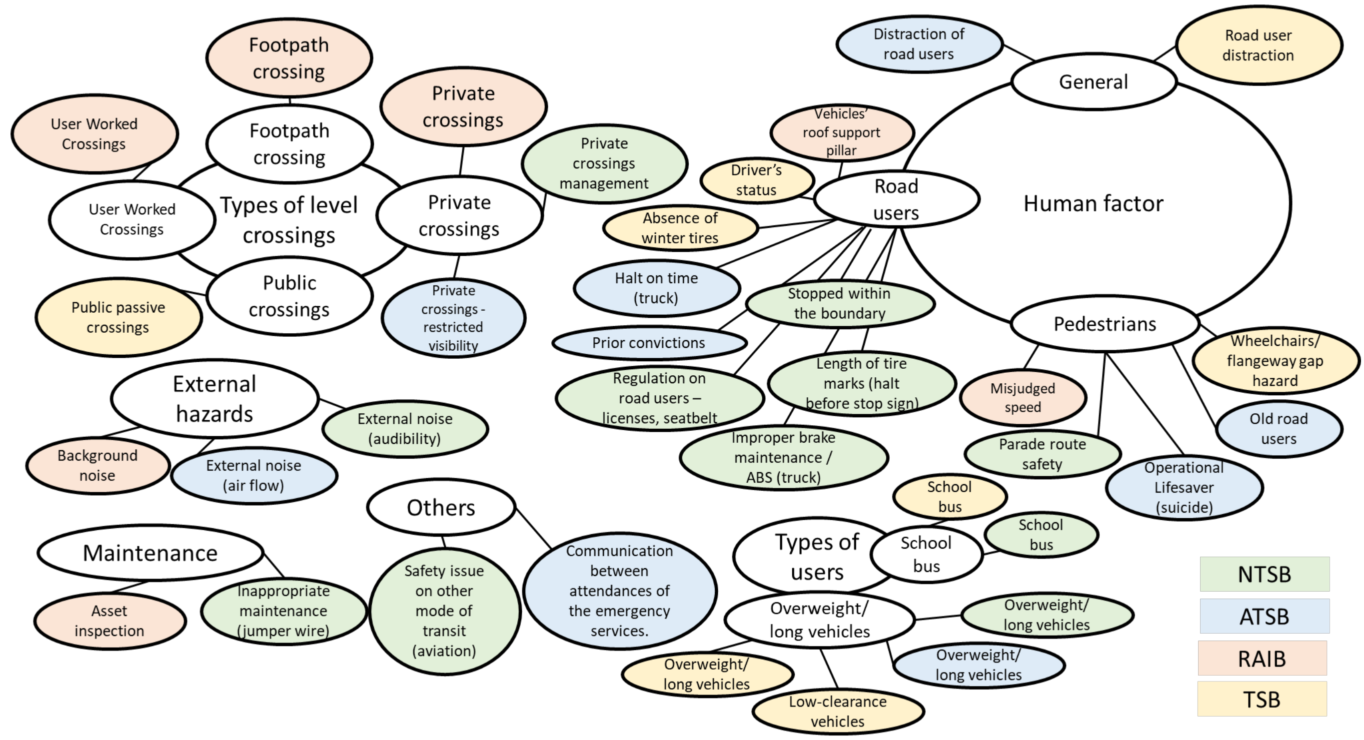


Figure 6, the applied HazardMap on level crossing accidents from all investigators – human factor, types of level crossings, external hazards, maintenance, types of users and others

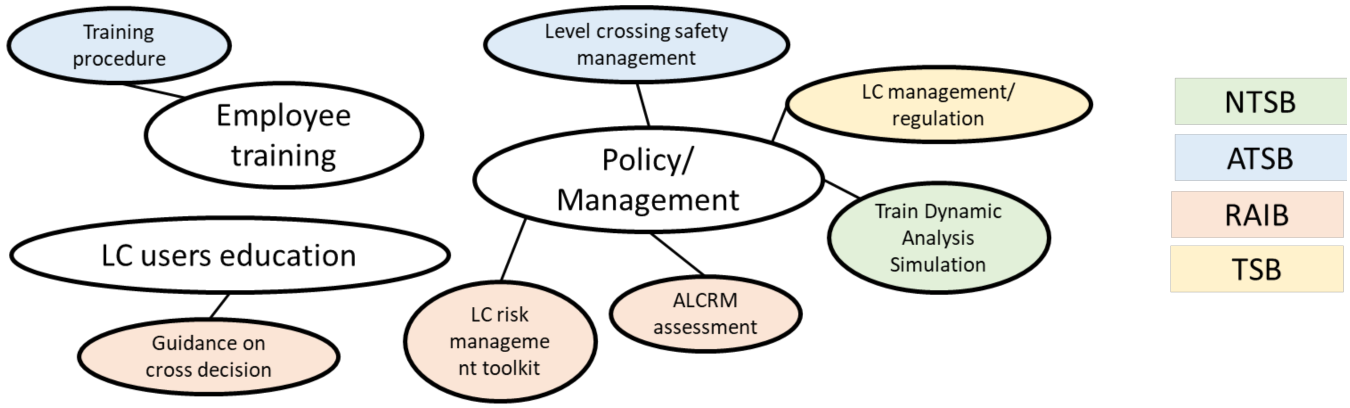


Figure 7, the applied HazardMap on level crossing accidents from all investigators – policy/ management, employee training, and LC users education

From the perspective of policy implementation, the conceptual framework based on *HazardMap* could be transformed to a policy plan map outlining strategic directions for managing risks at level crossings. Critical aspects of level crossing hazards extracted from lessons learnt across jurisdictions and time can efficiently mitigate the risk which has not (yet) triggered an accident. *HazardMap* also allows the framework for adding knowledge to be updated with very limited human intervention. The accumulated knowledge can also be disseminated to jurisdictions developing a new railway project but without sufficient experience and help them to build the policy framework for mitigating specific risks.

Finally, a cross-validation of the level crossing case study is conducted by using the Australian Level Crossing Assessment Model (ALCAM) (Lees, 2006; SPICER, 2007). The ALCAM is an assessment system for identifying potential risks related to level crossing systems in Australia and prioritising the upgrade of dangerous level crossings by evaluating each level crossing with risky factors. Factors used in the ALCAM are extracted to conduct the comparison with aspects and hazards in the *HazardMap*.

Overall, the ALCAM elaborates on distractions of road users, road and signs designs, types of road users and types of level crossing in detail. Results from our model cover almost all topics in the ALCAM except for the following: proximity to sites or public facilities, the likelihood of vandalism to controls, seasonal/infrequent train patterns, train speed, train schedule and possible sun glare sighting. On the other hand, less emphasis is placed by the ALCAM on the conditions of road users' vehicles, such as regulation on road users, improper vehicle brake maintenance and the absence of the usage of winter tires. Also, suicide/ trespass prevention and communication with emergency services are not included in the ALCAM. It should be noted that the lack of a characteristic in our model means the relation between this characteristic and level crossing is not significant from the analysis but may be substantial with other accidents. For instance, the connection between fatigue (road users and train drivers) and level-crossing accidents is not found but the link with speeding is found.

6 Conclusions and recommendations

This study proposes the *HazardMap* framework to depict the nature of hazards in the railway system and their mechanisms illustrated by a case study comparing four different countries. Over 1,200 railway accident reports, containing 400,000 sentences, published by national railway accident investigation bodies of four countries are analysed. The *HazardMap* is developed based on the output of the BERTopic, enabling the consideration of topics with a low probability of occurrences and the visualisation of the relations between hazards and accidents across countries. The *HazardMap* can also describe how each hazard triggers a railway accident by revealing the unaddressed aspects. Therefore, railway accident investigators in different countries can understand the potential path that one hazard impacts the railway system by reviewing the mechanism found in other countries to come up with corresponding solutions before it triggers another railway accident.

Additionally, the gap between policy implementation and real-world data is filled through the semi-automated process development described in this paper. Policymakers are able to draft the policy plan map covering a wide range of cases around the world without the need to manually review a large number of railway accident reports. *HazardMap* can also assist in revealing critical aspects of the hazard of interest and directing investigators to factors required to be considered. A case study of the level crossing risk is provided, and the result complement an existing risk assessment framework used in practice (ALCAM). The railway hazard knowledge can be updated automatically by inputting the new data and extending the *HazardMap*. The value of the conceptual model proposed is to summarise a huge amount of knowledge accumulated and make it easy to be applied to practical policymaking processes. Nevertheless, it also significantly supports the knowledge dissemination and allows inexperienced jurisdictions or railway industries to leverage lessons learnt across jurisdictions and time with limited human intervention to mitigate risks and enhance railway safety. All toolkits used are open-sourced and off-the-shelf, offering high flexibility to further improve the analysis process and enabling railway industry practitioners to apply *HazardMap* to existing data without the need for fine-tuning procedures.

Despite the advantages of using *HazardMap*, it is recognised that several limitations might influence the result. Firstly, the outcome heavily relies on the characteristics of the input data, implying that critical features missing in the original data would result in the absence of features in the constructed *HazardMap*. Nevertheless, human interpretation might still be required whilst processing systematic factors or underlying causes. The name of each extracted topic also needs to be determined manually by reviewing keywords of each topic. Future work might concentrate on potential solutions for reducing human intervention required while interpreting results from the topic model. A shared decision-making platform based on the *HazardMap* might also be worthwhile investigating.

Acknowledgments

The authors gratefully acknowledge the funding provided in the form of a Taiwan Ministry of Education - University of Sydney Scholarship (funding code is SC3261).

Appendix A. the result of topics extracted and descriptions of each dataset

Table A- 1, topic descriptions of the RAIB dataset

<table>
<tr><th>Topic</th><th>Topic - local</th><th>Topic – interval</th><th>Topic - global</th></tr>
<tr><td>6</td><td>Speed</td><td rowspan="2">Emergency brake at high speed</td><td rowspan="6">Trains apply emergency brake at high speed due to 1.) AWC isolation 2.) work site hazards</td></tr>
<tr><td>42</td><td>Emergency Brake</td></tr>
<tr><td>27</td><td>AWS isolation/ active</td><td rowspan="2">AWS isolation due to error warning/ failure of signal system</td></tr>
<tr><td>11-1</td><td>Sounded horn</td></tr>
<tr><td>11-2</td><td>Train horn</td><td rowspan="2">Work site safety and hazards – site workers</td></tr>
<tr><td>30</td><td>Site lookout</td></tr>
<tr><td>12</td><td colspan="2">CCTV, monitor/ recording</td><td rowspan="3">Unawareness of Platform-tram interface or pedestrians on track/ level crossing due to fatigue or incomplete monitoring system</td></tr>
<tr><td>1</td><td colspan="2">Fatigue</td></tr>
<tr><td>0-1</td><td>Trams / pedestrian</td><td>Platform-tram interface/ striking pedestrian</td></tr>
<tr><td>0-2</td><td>Trams/ Sandilands*</td><td colspan="2">Tram-specified accident (i.e., Overturning)</td></tr>
<tr><td>2</td><td>Communication – Signaller (radio, GMR-S, etc.…)</td><td colspan="2" rowspan="2">1. Signaller–driver interface
2. Stuff training/ knowledge</td></tr>
<tr><td>4</td><td>Driver knowledge, training, instruction</td></tr>
<tr><td>13</td><td>Time</td><td colspan="2" rowspan="2">Background information</td></tr>
<tr><td>25</td><td>Location</td></tr>
<tr><td>17</td><td>Sanding/ adhesion</td><td>Track-wheel interface</td><td rowspan="2">Relation between set of units and track-wheel interface</td></tr>
<tr><td>20</td><td colspan="2">Set/ type of train (single, multiple, diesel, electric unit)</td></tr>
<tr><td>15</td><td>Suspension system (Bogie, wheel…)</td><td colspan="2" rowspan="2">Cause and result of flange climbing</td></tr>
<tr><td>36</td><td>Contact between flange and gauge</td></tr>
<tr><td>28</td><td>Deaths and injuries</td><td colspan="2" rowspan="2">Consequence of accidents</td></tr>
<tr><td>40</td><td>Property loss</td></tr>
<tr><td>48</td><td colspan="2">Grinding repairs</td><td rowspan="3">Track inspection/ recording/ maintenance</td></tr>
<tr><td>14</td><td>Track maintenance/ inspection</td><td rowspan="2">Track defects inspection</td></tr>
<tr><td>49</td><td>Track geometry faults</td></tr>
<tr><td>39</td><td>Bolts failure</td><td rowspan="2">Design failure of the switch</td><td rowspan="5">1. Failure of signalling system
2. Failure of on-board equipment
3. Failure of infrastructure</td></tr>
<tr><td>47</td><td>(Nonadjustable) Stretcher bar</td></tr>
<tr><td>32</td><td>Wire-pantograph interface</td><td rowspan="2">Faults of wire-pantograph interface and inactive of power system protection</td></tr>
<tr><td>37</td><td>Power system protection (circuit breaker)</td></tr>
<tr><td>29</td><td>Failure mode of the axle</td><td>Axle</td></tr>
</table>

<table>
<tr><td>33</td><td>Holdfast panel-sleeper interface</td><td>Level-crossing infrastructure</td><td rowspan="3"></td></tr>
<tr><td>18</td><td>Switch interlocking system</td><td>Signalling system</td></tr>
<tr><td>46</td><td>Obstacle detection of doors</td><td>Door system</td></tr>
<tr><td>9</td><td>Weather conditions</td><td colspan="2" rowspan="2">Natural disasters</td></tr>
<tr><td>10</td><td>Natural hazards (landslip, flood…)</td></tr>
<tr><td>22</td><td>PICOP (Person in Charge of Possession)</td><td rowspan="2">Work site safety and hazards – on-track possession</td><td rowspan="5">Work site safety and hazards with engineering units</td></tr>
<tr><td>43</td><td>SSOW (Safe System of Work)</td></tr>
<tr><td>44</td><td>Engineering units - RGU (Rail Grinding Unit)</td><td rowspan="3">1. Conditions of engineering units
2. Incidents and recommendations relating to engineering units</td></tr>
<tr><td>3</td><td>Engineering units – RRV (Road Rail Vehicles)</td></tr>
<tr><td>21</td><td>Engineering units - track trolley</td></tr>
<tr><td>45</td><td>Drugs and alcohol test**</td><td colspan="2">Drug and alcohol conditions of the staffs</td></tr>
<tr><td>35</td><td>Shunters/ shunting activity</td><td colspan="2">Hazards and regulations relating to shunters</td></tr>
<tr><td>16</td><td>Fire hazards</td><td colspan="2" rowspan="2">Fire incidents and response</td></tr>
<tr><td>31</td><td>Emergency service systems</td></tr>
<tr><td>19</td><td>COSS (Controller of Site Safety), driver</td><td>COSS-driver interface</td><td rowspan="2">Work site safety – COSS-driver interface and planning</td></tr>
<tr><td>41</td><td>COSS and site safety planning</td><td>Work site safety and hazards – site workers</td></tr>
<tr><td>7</td><td>Declarative – risk assessment</td><td colspan="2" rowspan="2">Recommendations made by the RAIB on hazards identification and risk assessment</td></tr>
<tr><td>34</td><td>Hazards identification and risk assessment</td></tr>
<tr><td>38</td><td>Earthworks***</td><td rowspan="2">Infrastructure maintenance strategy and further improvement</td><td rowspan="3">Incidents and recommendation relating to Network Rail</td></tr>
<tr><td>5</td><td>Recommendations for Network Rail</td></tr>
<tr><td>24</td><td colspan="2">Network Rail’s safety issues</td></tr>
</table>

Table A- 2, topic descriptions of the ATSB dataset

topic	Topic - local	Topic – interval	Topic - global
33	Description of gross mass and containers on wagons	Background information of incidents	Freight trains derailment incidents
46	Description of train information (length, number of crew…)		
14	Consequence of wagons after derailment	Consequent of derailment	
48	Details of bogies' condition during derailment		
45	Track infrastructure details		
17	Organisations receiving the draft of the accident report		
6	Conditions of ballast crib and shoulder	Buckling hazards	Derailment due to buckling hazards and flange climbing
11	Conditions of sleeper		
8	Flange climb accident	Flange climb hazards	
27	The gauge whilst accidents		
44	Infrastructure maintenance regime and inspection	Monitoring asset condition via fault monitoring and maintenance regimes	
47	Track patrols/ inspection		
23	Bearing failure		Failures of axle
10	Conditions of axle bearing	Ineffective axle inspection system	
25	Defects inspection (continuous ultrasonic testing)		
40	Asset Standards Authority (ASA)/ buffer stop	Asset owner- leaser interface	
28	Chicago Freight Car Leasing Australia (CFCLA)/ draft key		
36	Falling jumbo coils		
35	Rail creep/ monuments	High temperature hazards to tracks	Derailment due to rail creep
43	Track temperature		
34	Determined environmental conditions		
18	Conditions of battery cells	Wire-pantograph interface	
21	Conditions of Overhead Line Equipment (OHLE)/ circuit breaker		
2	Speed of the train	Conditions of the train	Investigation into level-crossing accidents
5	Data logger/ Hasler data	On-board recorders	
1	Sounded horn/ audibility	Events during level crossing incidents	
22	Driver behaviour during Level crossing		
24	Sighting distance/ viewing angle	Design of level crossing	
13	Conditions of signal/ turnout indication/ colour light	Signal condition during accidents	
29	Signal displaying during accident		

<table>
<tr><td>19</td><td colspan="3">Train conditions</td></tr>
<tr><td>31</td><td>Consequence of the accident</td><td colspan="2">Fatal/ severe/ mirror injuries</td></tr>
<tr><td>26</td><td colspan="2">Collision accidents</td><td rowspan="3">Collision between trains on track</td></tr>
<tr><td>41</td><td>Description of train’s movement</td><td rowspan="2">Driver-train controller interface</td></tr>
<tr><td>42</td><td>Special Proceed Authority (SPA)</td></tr>
<tr><td>9</td><td>Track Occupancy Authority (TOA)</td><td>Worksite safety – worker-train interface</td><td rowspan="2">Worksite safety planning – stuffs, signalling systems, and trains</td></tr>
<tr><td>16</td><td>Protection Officer (PO) arrangements</td><td>Worksite safety – worksite safety planning</td></tr>
<tr><td>4</td><td>Network Control Office (NCO) and crew</td><td colspan="2" rowspan="2">NCO-crew/driver interface</td></tr>
<tr><td>7</td><td>Shunt operations</td></tr>
<tr><td>37</td><td>Australian Level Crossing Assessment Model</td><td colspan="2" rowspan="2">Level crossing hazard mitigation strategy</td></tr>
<tr><td>39</td><td>Level crossing safety</td></tr>
<tr><td>12</td><td>Alcohol and drugs tests</td><td>Conditions of stuffs during the rail safety work</td><td rowspan="4">Human factors examination</td></tr>
<tr><td>0</td><td>Fatigue investigation</td><td>Distraction due to fatigue</td></tr>
<tr><td>15</td><td>Medical examinations and fitness of standards</td><td>Medical qualification reviewing</td></tr>
<tr><td>30</td><td colspan="2">Maintenance of competency (MOS) assessment (training, knowledge gaining for stuffs)</td></tr>
<tr><td>32</td><td>V/Line Pty Ltd</td><td colspan="2" rowspan="2">Specific organisations mentioned in reports with high frequency</td></tr>
<tr><td>3</td><td>Queensland Rail (QR)</td></tr>
<tr><td>20</td><td colspan="3">SPAD events due to violation of rules or procedures</td></tr>
</table>

Table A- 3, topic descriptions of the NTSB dataset

<table>
<tr><th>topic</th><th>Topic – local</th><th>Topic – interval</th><th>Topic - global</th></tr>
<tr><td>2</td><td colspan="3">Event, audio and image recorder</td></tr>
<tr><td>29</td><td>Occurrence of emergency brake</td><td rowspan="2">Speed at the occurrence of emergency brake</td><td rowspan="8">Grade crossings hazards</td></tr>
<tr><td>37</td><td>Speed of the train recorded</td></tr>
<tr><td>15</td><td>Condition of the signal aspect</td><td rowspan="2">Issue of grade crossings design</td></tr>
<tr><td>49</td><td>Pre-emption/ “all-red-flash” design of grade crossings</td></tr>
<tr><td>10</td><td>High-risk private highway–railroad grade crossings</td><td rowspan="2">Hazard of private highway–railroad grade crossings</td></tr>
<tr><td>36</td><td>Bus driver training about grade crossing in school district</td></tr>
<tr><td>22</td><td>Sounded horn/ audibility</td><td rowspan="2">General grade crossings hazards</td></tr>
<tr><td>31</td><td>Hazard of stopping within the boundary of the crossing</td></tr>
<tr><td>28</td><td>Consequence of derailment</td><td colspan="2" rowspan="2">Condition of the train</td></tr>
<tr><td>44</td><td>Components/ units of the train</td></tr>
<tr><td>16</td><td>Conditions of the tunnel ventilation system</td><td rowspan="2">Subway environment control system</td><td rowspan="2">Subway environmental hazards</td></tr>
<tr><td>47</td><td>Electrical arcing due to water intrusion</td></tr>
<tr><td>24</td><td>Pressure of brake/ relief valve</td><td rowspan="2">Failure of brake system</td><td rowspan="2">On-board equipment hazards</td></tr>
<tr><td>38</td><td>Air leakage from the brake pipe</td></tr>
<tr><td>6</td><td colspan="3">Conditions of staffs’ duty</td></tr>
<tr><td>18</td><td colspan="3">Weather conditions</td></tr>
<tr><td>48</td><td colspan="2">Unsafe offloading practices of solvent blend wastes</td><td rowspan="3">Tank cars hazards</td></tr>
<tr><td>14</td><td>Cracked or broken joint bars/ bolts</td><td rowspan="2">Tank cars failure and certifications</td></tr>
<tr><td>40</td><td>Specifications for tank cars</td></tr>
<tr><td>30</td><td colspan="3">Damages to assets</td></tr>
<tr><td>19</td><td>Bridges’ capacity to carry floods</td><td rowspan="2">Hazards of bridges</td><td rowspan="2">Infrastructure hazards</td></tr>
<tr><td>20</td><td>Escorting permit loads</td></tr>
<tr><td>8</td><td>Parasitic oscillation of track circuit modules</td><td colspan="2" rowspan="2">Failures of trains’ circuits</td></tr>
<tr><td>25</td><td>Failures of emergency windows/ doors</td></tr>
<tr><td>11</td><td>Organizational culture of safety oversight</td><td rowspan="2">Rail safety oversight framework</td><td rowspan="5">Worksites hazards</td></tr>
<tr><td>42</td><td>Regulation of State oversight agency</td></tr>
<tr><td>3</td><td colspan="2">Installation of Positive Train Control</td></tr>
<tr><td>17</td><td>Unsafe work practices culture of Amtrak’s management</td><td rowspan="2">Safety culture</td></tr>
<tr><td>33</td><td>Safety Management Manual and safety culture</td></tr>
<tr><td>46</td><td>Operating rules for employees</td><td rowspan="2">Requirements of employees’ conditions</td><td rowspan="3">Hazards of employees’</td></tr>
<tr><td>1</td><td>Drug, alcohol, and toxicology test</td></tr>
<tr><td>32</td><td colspan="2">Efficiency of tests</td></tr>
</table>

23	Colour vision test	Medical conditions	medical conditions
39	Obstructive sleep apnea		
13	Conditions of track inspections		Switches and tracks hazards
4	Conditions of switches	Failure of switches	
35	Subdivision of tracks		
12	Fatalities and injuries		
26	Interface between conductors and railroad cars	Conductors' failure	
41	Conditions of conductors		
0	Emergency response after accidents	Emergency response of train operators	
21	Operation of CSX Transportation and MARC Train		
43	Track warrant authority (interface between train crews and the dispatchers)	Interface between train crews and the dispatchers	
9	Usage of cell phones and text messages		
27	Radio communications between crew members and dispatchers		

Table A- 4, topic descriptions of the TSB dataset

<table>
<tr><th>topic</th><th>Topic - local</th><th>Topic – interval</th><th>Topic - global</th></tr>
<tr><td>46</td><td>Fracture surface due to fatigue</td><td rowspan="2">Rail fracture surface hazards</td><td rowspan="4">Rail fracture hazard</td></tr>
<tr><td>48</td><td>Rail fracture</td></tr>
<tr><td>9</td><td>Condition of tie plates and secured spikes</td><td rowspan="2">Overview of track information</td></tr>
<tr><td>38</td><td>Track information</td></tr>
<tr><td>1</td><td>Emergency brake application</td><td rowspan="2" colspan="2">Occurrence of emergency brake</td></tr>
<tr><td>11</td><td>Brake pipe pressure</td></tr>
<tr><td>21</td><td>Conditions of air brake tests</td><td rowspan="2">Brake tests before departing</td><td rowspan="4">CROR on special instructions</td></tr>
<tr><td>28</td><td>Certified car inspector</td></tr>
<tr><td>10</td><td>Application of hand brakes</td><td rowspan="2">CROR on brake and movement</td></tr>
<tr><td>26</td><td>Canadian Rail Operating Rules (CROR)</td></tr>
<tr><td>37</td><td>Yard assignment description</td><td rowspan="2" colspan="2">Interface between workers in the yard</td></tr>
<tr><td>43</td><td>Failure of transfer between yardmasters</td></tr>
<tr><td>3</td><td>Derailment of freight cars</td><td rowspan="2" colspan="2">Derailment of freight trains</td></tr>
<tr><td>42</td><td>Location where locomotive came rest</td></tr>
<tr><td>6</td><td>Malfunction of switches</td><td rowspan="2" colspan="2">Defects on rail tracks</td></tr>
<tr><td>34</td><td>Damage on tracks</td></tr>
<tr><td>14</td><td>L/V ratios of single-wheel</td><td rowspan="2" colspan="2">Measurement on rail wheel</td></tr>
<tr><td>36</td><td>Observation of wheel flange marks</td></tr>
<tr><td>22</td><td>Wheel Impact Load Detector (WILD)</td><td rowspan="3">Wheel overloading hazards</td><td rowspan="4">Wheel- bearing system interface</td></tr>
<tr><td>32</td><td>Risks associated with Transcona wheel shop loose wheels</td></tr>
<tr><td>27</td><td>Excessive truck hunting/ Constant Contact Side Bearings</td></tr>
<tr><td>35</td><td colspan="2">Alert for roller bearing temperature</td></tr>
<tr><td>8</td><td colspan="2">Sounded horn/ audibility</td><td rowspan="3">Grade crossings hazards</td></tr>
<tr><td>2</td><td>Warning devices/ rules of grade crossing</td><td rowspan="2">Drivers' interface at grade crossings</td></tr>
<tr><td>13</td><td>Behaviour of grade crossing users (driver)</td></tr>
<tr><td>4</td><td>Interface between Rail Traffic Controllers and crew members</td><td rowspan="2">Interface between RTC and others</td><td rowspan="4">Communications hazards</td></tr>
<tr><td>17</td><td>Interface between foreman and RTC</td></tr>
<tr><td>19</td><td colspan="2">Display of indication signals</td></tr>
<tr><td>25</td><td colspan="2">Radio communications</td></tr>
<tr><td>40</td><td colspan="3">Hazards related to train marshalling</td></tr>
<tr><td>24</td><td colspan="3">Absence of on-board voice recorders</td></tr>
<tr><td>7</td><td>Emergency response</td><td rowspan="2" colspan="2">Hazards of poor design of emergency exits and response</td></tr>
<tr><td>16</td><td>Failure of emergency exit (on board)</td></tr>
<tr><td>5</td><td colspan="3">Flood/ drainage system</td></tr>
<tr><td>39</td><td colspan="3">Failure of the thermite weld</td></tr>
</table>

0	Fatigue	Hazards of fatigue
31	Risk of memory lapse	
18	Risk of flammable materials	
20	Safety management system	

Reference

Accou, B., &Carpinelli, F. (2022). Systematically investigating human and organisational factors in complex socio-technical systems by using the “SAfety FRactal ANalysis” method. *Applied Ergonomics*, *100*, 103662. https://doi.org/10.1016/J.APERGO.2021.103662

Adeolu, O. D., Cornelius, O. A., &Bamidele, A. B. (2016). Evaluation of railway level crossing attributes on accident causation in Lagos, Nigeria. *Indonesian Journal of Geography*, *48*(2), 108–117. https://doi.org/10.22146/ijg.17520

Akel, A. J. N., Gravio, G.Di, Fedele, L., &Patriarca, R. (2022). Learning from Incidents in Socio-Technical Systems: A Systems-Theoretic Analysis in the Railway Sector. *Infrastructures 2022, Vol. 7, Page 90*, *7*(7), 90. https://doi.org/10.3390/INFRASTRUCTURES7070090

Alawad, H., Kaewunruen, S., &An, M. (2020). Learning from Accidents: Machine Learning for Safety at Railway Stations. *IEEE Access*, *8*, 633–648. https://doi.org/10.1109/ACCESS.2019.2962072

An, M., Lin, W., &Huang, S. (2013). An Intelligent Railway Safety Risk Assessment Support System for Railway Operation and Maintenance Analysis. *The Open Transportation Journal*, *7*(1), 27–42. https://doi.org/10.2174/1874447801307010027

ATSB. (2009). *Railway Accident Investigation Guidelines for Railway Network Owners, Railway Operators and Emergency Services Personnel* (3rd ed.). Australian Government.

Bai, X., Zhang, X., Li, K. X., Zhou, Y., &Yuen, K. F. (2021). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, *102*, 11–24. https://doi.org/10.1016/J.TRANPOL.2020.12.013

Bang, H. N., Miles, L. S., &Gordon, R. D. (2020). Challenges in managing technological disasters in Cameroon: Case study of Cameroon’s worst train Crash—the Eséka train disaster. *International Journal of Disaster Risk Reduction*, *44*, 101410. https://doi.org/10.1016/J.IJDRR.2019.101410

Berrado, A., El-Koursi, E., Cherkaoui, A., Khaddour, M., Khaddour, M. A., Berrado, A., Mohammed, U., &Khaddour, M. (2010). A Framework for Risk Management in Railway Sector: Application to Road-Rail Level Crossings. *Open Transportation Journal*, 19p. https://hal.science/hal-00542424

Bian, K., &Wang, H. (2015). Hazard and Operability Analysis on Risk Factors of Railway Dangerous Goods Transport. *ICTE 2015 - Proceedings of the 5th International Conference on Transportation Engineering*, 2746–2753. https://doi.org/10.1061/9780784479384.351

Bíl, M., Andrášik, R., Nezval, V., &Bílová, M. (2017). Identifying locations along railway networks with the highest tree fall hazard. *Applied Geography*, *87*, 45–53. https://doi.org/10.1016/j.apgeog.2017.07.012

Bird, S., &Loper, E. (2004). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions* (pp. 214–217). https://aclanthology.org/P04-3031

Blaho, P., Peceny, L., &Gasparik, J. (2020). Causality of accidents at railway-crossings in Slovakia and its prevention. *2020 12th International Science-Technical Conference AUTOMOTIVE SAFETY, AUTOMOTIVE SAFETY 2020*. https://doi.org/10.1109/AUTOMOTIVESAFETY47494.2020.9293528

Blei, D. M., &Mcauliffe, J. D. (2007). Supervised Topic Models. *Advances in Neural Information Processing Systems*, *20*. www.digg.com

Bougacha, R., Wakrime, A. A., Kallel, S., Ayed, R. B., Collart-Dutilleul, S., BenAyed, R., &Collart-Dutilleul, S. (2019). A Model-based Approach for the Modeling and the Verification of Railway Signaling System. In E.Damiani, G.Spanoudakis, &L.Maciaszek (Eds.), *PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON EVALUATION OF NOVEL APPROACHES TO SOFTWARE ENGINEERING (ENASE)* (Issue 14th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE), pp. 367–376). https://doi.org/10.5220/0007728403670376

Braut, G. S., Solberg, Ø., &Njå, O. (2014). Organizational effects of experience from accidents. Learning in the aftermath of the Tretten and Åsta train accidents. *Transportation Research Part A: Policy and Practice*, *69*, 354–366. https://doi.org/10.1016/J.TRA.2014.08.013

Catelani, M., Ciani, L., Galar, D., Guidi, G., Matucci, S., &Patrizi, G. (2021). FMECA Assessment for Railway Safety-Critical Systems Investigating a New Risk Threshold Method. *IEEE Access*, *9*, 86243–86253. https://doi.org/10.1109/ACCESS.2021.3088948

Cheng, M. Y., Tsai, H. C., &Chiu, Y. H. (2009). Fuzzy case-based reasoning for coping with construction disputes. *Expert Systems with Applications*, *36*(2 PART 2), 4106–4113. https://doi.org/10.1016/j.eswa.2008.03.025

Choi, J. D., Tetreault, J., &Stent, A. (2015). It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, *1*, 387–396. https://doi.org/10.3115/V1/P15-1038

Ciani, L., Guidi, G., &Patrizi, G. (2019). A Critical Comparison of Alternative Risk Priority Numbers in Failure Modes, Effects, and Criticality Analysis. *IEEE Access*, *7*, 92398–92409. https://doi.org/10.1109/ACCESS.2019.2928120

Dai, B., &Wang, T. (2010). Risk assessment based on accident theory in urban railway transportation. *Proceedings - 2010 International Conference on Intelligent System Design and Engineering Application, ISDEA 2010*, *2*, 318–320. https://doi.org/10.1109/ISDEA.2010.374

Devlin, J., Chang, M.-W., Lee, K., &Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. http://arxiv.org/abs/1810.04805

Dong, K., Romanov, I., McLellan, C., &Esen, A. F. (2022). Recent text-based research and applications in railways: A critical review and future trends. *Engineering Applications of Artificial Intelligence*, *116*, 105435. https://doi.org/10.1016/J.ENGAPPAI.2022.105435

Dornick, C., Kumar, A., Seidenberger, S., Seidle, E., &Mukherjee, P. (2021). Analysis of Patterns and Trends in COVID-19 Research. *Big Data, Iot, and Ai for a Smarter Future*, *185*, 302–310. https://doi.org/10.1016/J.PROCS.2021.05.032

Fan, J., &Smith, A. P. (2019). Mental Workload and Other Causes of Different Types of Fatigue in Rail Staff. *Communications in Computer and Information Science*, *1012*, 147–159. https://doi.org/10.1007/978-3-030-14273-5_9/TABLES/7

Ganesh, P., Chen, Y., Lou, X., Khan, M. A., Yang, Y., Chen, D., Winslett, M., Sajjad, H., &Nakov, P. (2020). Compressing Large-Scale Transformer-Based Models: A Case Study on BERT. *ArXiv*. http://arxiv.org/abs/2002.11985

Gong, Y., &Li, Y. (2018). STAMP-based causal analysis of China-Donghuang oil transportation pipeline leakage and explosion accident. *Journal of Loss Prevention in the Process Industries*, *56*, 402–413. https://doi.org/10.1016/j.jlp.2018.10.001

Gonzva, M., Barroca, B., Gautier, P. É., &Diab, Y. (2017). Modeling disruptions causing domino effects in urban guided transport systems faced by flood hazards. *Natural Hazards*, *86*(1), 183–201. https://doi.org/10.1007/s11069-016-2680-7

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://doi.org/10.48550/arxiv.2203.05794

Guenab, F., Boulanger, J. L., &Schön, W. (2008). Safety of railway control systems: A new Preliminary risk analysis approach. *2008 IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2008*, 1309–1313. https://doi.org/10.1109/IEEM.2008.4738082

Hadj-Mabrouk, H. (2019). Contribution of Artificial Intelligence to Risk Assessment of Railway Accidents. *Urban Rail Transit*, *5*(2), 104–122. https://doi.org/10.1007/s40864-019-0102-3

Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., &Muliady, W. (2014, January 12). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. *Proceedings - 2014 6th International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology Through University-Industry Collaboration, ICITEE 2014*. https://doi.org/10.1109/ICITEED.2014.7007894

Holmgren, M. (2006). Maintenance-related incidents and accidents: aspects of hazard identification. In *NA: Vol. NA* (Issue NA, p. NA-NA). https://doi.org/NA

Hristova, G., &Netov, N. (2022). Media Coverage and Public Perception of Distance Learning During the COVID-19 Pandemic: A Topic Modeling Approach Based on BERTopic. *2022 IEEE International Conference on Big Data (Big Data)*, 2259–2264.

https://doi.org/10.1109/BIGDATA55660.2022.10020466

Hua, L., Zheng, W., &Gao, S. (2019). Extraction and Analysis of Risk Factors from Chinese Railway Accident Reports. *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019*, 869–874. https://doi.org/10.1109/ITSC.2019.8917094

Huang, W., Shuai, B., Zhang, R., Xu, M., Xu, Y., Yu, Y., &Antwi, E. (2020). A New System Risk Definition and System Risk Analysis Approach Based on Improved Risk Field. *IEEE Transactions on Reliability*, *69*(4), 1437–1452. https://doi.org/10.1109/TR.2019.2942373

Huang, Y., Zhang, Z., Tao, Y., &Hu, H. (2022). Quantitative risk assessment of railway intrusions with text mining and fuzzy Rule-Based Bow-Tie model. *Advanced Engineering Informatics*, *54*, 101726. https://doi.org/10.1016/J.AEI.2022.101726

Hughes, P., Shipp, D., Figueres-Esteban, M., &vanGulijk, C. (2018). From free-text to structured safety management: Introduction of a semi-automated classification method of railway hazard reports to elements on a bow-tie diagram. *Safety Science*, *110*, 11–19. https://doi.org/10.1016/j.ssci.2018.03.011

Hulin, B., Kaindl, H., Rathfux, T., Popp, R., Arnautovic, E., &Beckert, R. (2016). Towards a Common Safety Ontology for Automobiles and Railway Vehicles. *Proceedings - 2016 12th European Dependable Computing Conference, EDCC 2016*, 189–192. https://doi.org/10.1109/EDCC.2016.15

Jonsson, L., Björklund, G., &Isacsson, G. (2019). Marginal costs for railway level crossing accidents in Sweden. *Transport Policy*, *83*, 68–79. https://doi.org/10.1016/J.TRANPOL.2019.09.004

Jugran, S., Kumar, A., Tyagi, B. S., &Anand, V. (2021). Extractive Automatic Text Summarization using SpaCy in Python NLP. *2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2021*, 582–585. https://doi.org/10.1109/ICACITE51222.2021.9404712

Kim, D. S., &Yoon, W. C. (2013). An accident causation model for the railway industry: Application of the model to 80 rail accident investigation reports from the UK. *Safety Science*, *60*, 57–68. https://doi.org/10.1016/j.ssci.2013.06.010

Kinnersley, S., &Roelen, A. (2007). The contribution of design to accidents. *Safety Science*, *45*(1–2), 31–60. https://doi.org/10.1016/j.ssci.2006.08.010

Kinra, A., Beheshti-Kashi, S., Buch, R., Nielsen, T. A. S., &Pereira, F. (2020). Examining the potential of textual big data analytics for public policy decision-making: A case study with driverless cars in Denmark. *Transport Policy*, *98*, 68–78. https://doi.org/10.1016/J.TRANPOL.2020.05.026

Labusch, K., &Neudecker, C. (2020). Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT. *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*. https://qurator.ai

Lasisi, A., &Attoh-Okine, N. (2020). An Unsupervised Learning Framework for Track Quality Index and

Safety. *Transportation Infrastructure Geotechnology*, *7*(1), 1–12. https://doi.org/10.1007/S40515-019-00087-6/TABLES/2

Lee, J. S., Min Kim, H., IlKim, S., &Min Lee, H. (2021). Evaluation of structural integrity of railway bridge using acceleration data and semi-supervised learning approach. *Engineering Structures*, *239*, 112330. https://doi.org/10.1016/J.ENGSTRUCT.2021.112330

Lees, C. (2006, September). The Australian level crossing assessment model. *International Level Crossing Safety and Trespass Symposium, 9th, 2006, Montreal, Quebec, Canada*.

Li, C., Tang, T., Chatzimichailidou, M. M., Jun, G. T., &Waterson, P. (2019). A hybrid human and organisational analysis method for railway accidents based on STAMP-HFACS and human information processing. *Applied Ergonomics*, *79*, 122–142. https://doi.org/10.1016/j.apergo.2018.12.011

Li, J., &Liu, J. (2021). Quantitative Evaluation on Hazard Causation of High-speed Railway Signal System. *2021 CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes, SAFEPROCESS 2021*. https://doi.org/10.1109/SAFEPROCESS52771.2021.9693615

Li, K., Yao, X., Chen, D., Yuan, L., &Zhou, D. (2015). HAZOP study on the CTCS-3 onboard system. *IEEE Transactions on Intelligent Transportation Systems*, *16*(1), 162–171. https://doi.org/10.1109/TITS.2014.2329692

Li, Q., Zhang, Z., &Peng, F. (2021). Causality-network-based critical hazard identification for railway accident prevention: Complex network-based model development and comparison. *Entropy (Basel, Switzerland)*, *23*(7), 864-NA. https://doi.org/10.3390/e23070864

Li, X. Q., Shi, T. Y., Li, P., Zhou, W., &IEEE. (2019). Application of Bagging Ensemble Classifier based on Genetic Algorithm in the Text Classification of Railway Fault Hazards. *2019 2ND INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND BIG DATA (ICAIBD 2019)*, *2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 286–290. https://doi.org/10.1109/ICAIBD.2019.8836988

Li, Y., Gong, S., Zhang, Z., Liu, M., Sun, C., &Zhao, Y. (2021). Vulnerability evaluation of rainstorm disaster based on ESA conceptual framework: A case study of Liaoning province, China. *Sustainable Cities and Society*, *64*, 102540. https://doi.org/10.1016/J.SCS.2020.102540

Liang, C., Ghazel, M., &Cazier, O. (2018). *Using Bayesian Networks for the Purpose of Risk Analysis at Railway Level Crossings*. *51*(9), 142–149. https://doi.org/10.1016/j.ifacol.2018.07.024

Liang, C., Ghazel, M., Cazier, O., &Bouillaut, L. (2020). Advanced model-based risk reasoning on automatic railway level crossings. *Safety Science*, *124*. https://doi.org/10.1016/j.ssci.2019.104592

Liu, J., Nie, T., Li, X., &Dong, X. (2022). Research on data processing method of railway safety information based on NLP technology. *Proceedings of SPIE - The International Society for Optical Engineering*, *12081*. https://doi.org/10.1117/12.2624272

Ly, A., Uthayasooriyar, B., &Wang, T. (2020). A survey on natural language processing (nlp) and applications in insurance. *ArXiv*. http://arxiv.org/abs/2010.00462

Madigan, R., Golightly, D., Madders, R., &Madigan David; Madders, Richard, R. G. (2016). Application of Human Factors Analysis and Classification System (HFACS) to UK rail safety of the line incidents. *Accident Analysis and Prevention*, *97*(NA), 122–131. https://doi.org/10.1016/j.aap.2016.08.023

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., &McClosky, D. (2014). *The Stanford CoreNLP Natural Language Processing Toolkit*. 55–60. https://doi.org/10.3115/V1/P14-5010

Marquez, L., Marquez, L., &Salgado, J. G. (2000). *Machine Learning and Natural Language Processing*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.3498

McInnes, L., Healy, J., &Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/JOSS.00205

McInnes, L., Healy, J., &Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/arxiv.1802.03426

Nelson, T., Ferster, C., Laberee, K., Fuller, D., &Winters, M. (2020). Crowdsourced data for bicycling research and practice. *Transport Reviews*, *41*(1), 97–114. https://doi.org/10.1080/01441647.2020.1806943

Ouyang, M., Hong, L., Yu, M. H., &Fei, Q. (2010). STAMP-based analysis on the railway accident and accident spreading: Taking the China-Jiaoji railway accident for example. *Safety Science*, *48*(5), 544–555. https://doi.org/10.1016/j.ssci.2010.01.002

Papen, V., Harvey, R., Qaasim, H., &Spielholz, P. (2011). Implementing Hazard & Operability (HAZOP) Studies Throughout the Project Life Cycle. *TRID*.

Parkinson, H. J., Bamford, G., &Kandola, B. (2016). The development of an enhanced bowtie railway safety assessment tool using a big data analytics approach. *IET Conference Publications*, *2016*(CP703). https://doi.org/10.1049/cp.2016.0510

Peters, J. L., Zevitas, C. D., Redline, S., Hastings, A., Sizov, N., Hart, J. E., Levy, J. I., Roof, C. J., &Wellenius, G. A. (2018). Aviation Noise and Cardiovascular Health in the United States: a Review of the Evidence and Recommendations for Research Direction. *Current Epidemiology Reports*, *5*(2), 140–152. https://doi.org/10.1007/S40471-018-0151-2

Qurashi, A. W., Holmes, V., &Johnson, A. P. (2020). Document Processing: Methods for Semantic Text Similarity Analysis. *INISTA 2020 - 2020 International Conference on INnovations in Intelligent SysTems and Applications, Proceedings*. https://doi.org/10.1109/INISTA49547.2020.9194665

RAIB. (2008). *Passenger trapped in a closed train door, Tooting Broadway, Northern Line, London Underground*. www.raib.gov.uk.

RAIB. (2011). *Investigation into the safety of automatic open level crossings on Network Rail's managed* *infrastructure*. https://assets.publishing.service.gov.uk/media/547c8fda40f0b60241000167/R122011_110728_AOCLs_Class_Inv.pdf

Rausand, M. (2013). *Risk assessment: theory, methods, and applications* (John Wiley & Sons. (ed.); Vol. 115).

Read, G. J. M., Cox, J. A., Hulme, A., Naweed, A., &Salmon, P. M. (2021). What factors influence risk at rail level crossings? A systematic review and synthesis of findings using systems thinking. *Safety Science*, *138*, 105207. https://doi.org/10.1016/J.SSCI.2021.105207

Reimers, N., &Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.48550/arxiv.1908.10084

Roque, C., Lourenço Cardoso, J., Connell, T., Schermers, G., &Weber, R. (2019). Topic analysis of Road safety inspections using latent dirichlet allocation: A case study of roadside safety in Irish main roads. *Accident Analysis & Prevention*, *131*, 336–349. https://doi.org/10.1016/J.AAP.2019.07.021

Rosadini, B., Ferrari, A., Gori, G., Fantechi, A., Gnesi, S., Trotta, I., &Bacherini, S. (2017). Using NLP to detect requirements defects: An industrial experience in the railway domain. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10153 LNCS* (pp. 344–360). https://doi.org/10.1007/978-3-319-54045-0_24

Runyan, C. W., &Yonas, M. (2008). Conceptual Frameworks for Developing and Comparing Approaches to Improve Adolescent Motor-Vehicle Safety. *American Journal of Preventive Medicine*, *35*(3), S336–S342. https://doi.org/10.1016/J.AMEPRE.2008.06.019

Ryan, B., Golightly, D., Pickup, L., Reinartz, S., Atkinson, S., &Dadashi, N. (2021). Human functions in safety - developing a framework of goals, human functions and safety relevant activities for railway socio-technical systems. *Safety Science*, *140*, 105279. https://doi.org/10.1016/J.SSCI.2021.105279

Rydstedt Nyman, M., &Johansson, M. (2015). Merits of using a socio-technical system perspective and different industrial accident investigation methods on accidents following natural hazards - A case study on pluvial flooding of a Swedish railway tunnel 2013. *International Journal of Disaster Risk Reduction*, *13*, 189–199. https://doi.org/10.1016/j.ijdrr.2015.06.004

Salmon, P. M., Read, G. J. M., Stanton, N. A., &Lenné, M. G. (2013). The crash at Kerang: Investigating systemic and psychological factors leading to unintentional non-compliance at rail level crossings. *Accident Analysis and Prevention*, *50*, 1278–1288. https://doi.org/10.1016/j.aap.2012.09.029

Santos-Reyes, J., &Beard, A. N. (2009). A systemic analysis of the Edge Hill railway accident. *Accident Analysis and Prevention*, *41*(6), 1133–1144. https://doi.org/10.1016/j.aap.2008.05.004

Sizov, G., &Öztürk, P. (2013). *Automatic Extraction of Reasoning Chains from Textual Reports*. 61–69.

https://aclanthology.org/W13-5009

Song, T., Zhong, D., &Zhong, H. (2012). A STAMP analysis on the China-Yongwen railway accident. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7612 LNCS*, 376–387. https://doi.org/10.1007/978-3-642-33678-2_32

SPICER, T. (2007). Implementing the Australian Level Crossing Assessment Model (ALCAM) in Victoria. *AUSTRALASIAN TRANSPORT RESEARCH FORUM (ATRF), 30TH, 2007, MELBOURNE, VICTORIA, AUSTRALIA, VOL 30*.

Syeda, K. N., Shirazi, S. N., Naqvi, S. A. A., Parkinson, H. J., &Bamford, G. (2019). Big Data and natural language processing for analysing railway safety: Analysis of railway incident reports. *Human Performance Technology: Concepts, Methodologies, Tools, and Applications.*, 781–809. http://search.ebscohost.com.proxy-ub.rug.nl/login.aspx?direct=true&db=psyh&AN=2019-25313-040&site=ehost-live&scope=site

Wang, S. S., Dong, H. H., Zhou, Y., Jia, L. M., Qin, Y., &IEEE. (2017). Exploring Traffic Accident Locations from Natural Language Based on Spatial Information Retrieval. In *2017 29TH CHINESE CONTROL AND DECISION CONFERENCE (CCDC)* (Issue 29th Chinese Control And Decision Conference (CCDC), pp. 3490–3495).

Whittingham, R. B. (2012). Preventing corporate accidents: An ethical approach. In *Preventing Corporate Accidents: An Ethical Approach*. https://doi.org/10.4324/9780080570297

Wullems, C., Toft, Y., &Dell, G. (2013). Improving level crossing safety through enhanced data recording and reporting: The CRC for rail innovation's baseline rail level crossing video project. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, *227*(5), 554–559. https://doi.org/10.1177/0954409713502014

Xia, M., Li, X., Jiang, F., &Wang, S. (2012). Cause analysis and countermeasures of locomotive runaway accident based on fault tree analysis method. *Procedia Engineering*, *45*, 38–42. https://doi.org/10.1016/j.proeng.2012.08.117

Yan, F., &Xu, K. (2019). Methodology and case study of quantitative preliminary hazard analysis based on cloud model. *Journal of Loss Prevention in the Process Industries*, *60*, 116–124. https://doi.org/10.1016/J.JLP.2019.04.013

Yang, J., Han, S. C., &Poon, J. (2022). A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, *64*(5), 1161–1186. https://doi.org/10.1007/S10115-022-01665-W/TABLES/6

Yang, L., &Li, K. (2020). Safety risk analysis of railway accident with text-based bow-tie model. *Proceedings of 2020 IEEE 3rd International Conference of Safe Production and Informatization, IICSPI 2020*, 200–204. https://doi.org/10.1109/IICSPI51290.2020.9332329

Zhan, Q., Zheng, W., &Zhao, B. (2017). A hybrid human and organizational analysis method for railway accidents based on HFACS-Railway Accidents (HFACS-RAs). *Safety Science*, *91*, 232–250. https://doi.org/10.1016/j.ssci.2016.08.017

Zhang, S., Tang, T., &Liu, J. (2021). A Hazard Analysis Approach for the SOTIF in Intelligent Railway Driving Assistance Systems Using STPA and Complex Network. *Applied Sciences 2021, Vol. 11, Page 7714*, *11*(16), 7714. https://doi.org/10.3390/APP11167714

Zhou, C., &Ding, L. Y. (2017). Safety barrier warning system for underground construction sites using Internet-of-Things technologies. *Automation in Construction*, *83*, 372–389. https://doi.org/10.1016/J.AUTCON.2017.07.005

Zhou, J. L., &Lei, Y. (2018). Paths between latent and active errors: Analysis of 407 railway accidents/incidents' causes in China. *Safety Science*, *110*(November 2017), 47–58. https://doi.org/10.1016/j.ssci.2017.12.027