IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 February 2023, accepted 18 March 2023, date of publication 5 April 2023, date of current version 10 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264763

RESEARCH ARTICLE

Unsupervised Machine Learning for Managing Safety Accidents in Railway Stations

HAMAD ALAWAD, (Member, IEEE), AND SAKDIRAT KAEWUNRUEN
Birmingham Centre for Railway Research and Education, University of Birmingham, B15 2TT Birmingham, U.K.

Corresponding author: Sakdirat Kaewunruen (s.kaewunruen@bham.ac.uk)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant under Agreement 691135; and in part by the European Commission's Sponsorship for the H2020-RISE "RISEN: Rail Infrastructure Systems Engineering Network," which enables a global research network that tackles the grand challenge of railway infrastructure resilience and advanced sensing in extreme environments (www.risen2rail.eu), under Project 691135. The work of Sakdirat Kaewunruen was supported in part by the Australian Academy of Science; and in part by the Japan Society for the Promotion of Sciences for the Invitation Research Fellowship (long-term) at the Railway Technical Research Institute and the University of Tokyo, Japan, under Grant JSPS-L15701. The authors are grateful to Rail Safety and Standards Board (RSSB) for the access to UK railway accident data.

ABSTRACT For both passenger and freight transportation, railroad operations must be dependable, accessible, maintained, and safe (RAMS). In many urban areas, railway stations risk and safety accidents represent an essential safety concern for daily operations. Moreover, the accidents lead to damage to market reputation, including injuries and anxiety among the people and costs. This stations under pressure caused by higher demand which consuming infrastructure and raised the safety administration consideration. To analysing these accidents and utilising the technology such AI methods to enhance safety, it is suggested to use unsupervised topic modelling for better understand the contributors to these extreme accidents. It is conducted to optimise Latent Dirichlet Allocation (LDA) for fatality accidents in the railway stations from textual data gathered RSSB including 1000 accidents in the UK railway station. This research describes using the machine learning topic method for systematic spot accident characteristics to enhance safety and risk management in the stations and provides advanced analysing. The study evaluates the efficacy of text by mining from accident history, gaining information, lesson learned and deeply coherent of the risk caused by assessing fatalities accidents for large and enduring scale. This Intelligent Text Analysis presents predictive accuracy for valuable accident information such as root causes and the hot spots in the railway stations. Further, the big data analytics ' improvement results in an understanding of the accidents' nature in ways not possible if a considerable amount of safety history and not through narrow domain analysis of the accident reports. This technology renders stand with high accuracy and a beneficial and extensive new era of AI applications in railway industry safety and other fields for safety applications.

INDEX TERMS Unsupervised machine learning, topic model, accidents analysis, railway station, safety.

I. INTRODUCTION

Trains as public transportation have been considered as safer than other means. However, passengers on trains stations sometimes face many risks because of many overlapping factors such as station operation, design, and passenger behaviours. Due to the gradually increasing demand and the heavily congested society and the state of some station's layout and complexity in design, there are potential risks during the operation of the stations. Furthermore, Passenger, people and public safety is the main concern of the railway industry and one of the critical parts of the system. European Union put into practice Reliability, Availability, Maintainability and Safety (RAMS)as a standard in 1999 known as EN 50126. Aiming to prevent railway accidents and ensure a high level of safety in railway operations. The RAMS analyses concepts lead to minimising the risks to acceptable levels and rise safety levels. However, that have been an urgent issue and still, the reports show several people are killed every year in the railway station, some accidents lead to injuries

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

or fatalities. For example, In Japan in 2016, 420 accidents occurred that included being struck by a train, which resulted in 202 deaths. This including of those 420 accidents, 179 (resulting in 24 fatalities) included falling from a platform and following injury or death as a consequence of hitting with a train [1]. In the UK, 2019/20, it has been reported that Most passenger injuries occur from accidents in stations. Greatest Major injuries are the outcome of slips, trips and falls, of which there were approximately 200 [2] play significant impact in reducing injuries on station platforms and provide quality, reliable and safe travel environment for all passengers, worker and public. Even if some accident does not result in deaths or injuries, such accidents cause delay, cost, fear and anxiety among the people, interruption in the operations and damage the industry reputation. Also, to provide or invest any control safety measurements the stations it is crucial to considering the risks associated with the railway incidents and risks in the station and identification of many factors related to the accident by a comprehensive knowledge of the root cause of accidents considering all the possible technology.

The objective of this research is to analysis a collection case of accidents between 01/01/2000 and 17/04/2020 data to introduce a smart method, which expected to develop the safety level future, the risk management process, and the way to collect data in the railway stations. This data been gathered by RSSBS and agreed to be used for the research purpose. Analysing an extensive amount of data recorded in a different form are a challenging job. Nowadays, it is hard to obtain for specific information in such mix digitization big data in including Web, video, images and other sources, it is research of a needle in a haystack. Thus, a powerful tool for assistance manage, search and understand these vast amounts of information is needed indeed [3], [4]. Many pre-processing techniques and algorithms are required to obtain valuable characteristics from an enormous amount of safety data in the stations including textual. The study covers the topic modelling to identify useful characteristics such the root cause of the accidents and also exploring the factors which are multiple groups of words or phrases that explain and summarize the content covered by an accident's reports reducing time with high accuracy of outcomes. Topic modelling techniques are robust smart methods that extensively applied in natural language processing to topic detection and semantic mining from unstructured documents. Consequently, It has been suggested in this work the LDA model which is one of the best-known probabilistic unsupervised learning methods that marks the topics implicit in collection of contexts [5]. Since increasing of applying new technologies and the revolution of data, the development of technology and utilising AI in many fields it suggested in this paper a smart analysis utilising the topic modelling techniques which can be very useful and effective to semantic mining and latent discovery context documents and datasets. The other source of data (Images-videos and numerical) been conducted utilising AI approaches which cover supervised learning [6], [7], so the unstructured textual data is targeted.

Hence, our motivation is to investigate the topic modelling approaches to risks and safety accident subjects in the stations. This work provides the method of topic modelling based on LDA with other models for advanced analytics, aiming to make contributions in the future of smart safety and risk management in the stations. Through applying the models, we investigate the safety accidents for fatality accident in the railway.

This paper establishes an innovative method in the area to studies how the textual source of data of railway station accident reports could be efficiently used to extract the root causes of accidents and establish an analysis between the textual and the possible cause. where the full automated process that has ability to get the input of text and provide outputs not yet ready [8]. Applying this method expected to come overcome issues such as aid the decision-maker in real time and extract the key information to be understandable from non-experts, better identify the details of the accident in-depth, design expert smart safety system and effective usage of the safety history records. A Such results could support in the analysis of safety and risk management to be systematic and smarter. Our approach uses state-of-the-art LDA algorithm to capture the critical texts information of accidents and their causes. The rest of this paper is arranged as follows: In Section II, related work in both accident analysis and text classification with deep learning have been presented. Section III describes in detail the approach that has been used along with evaluation criteria. Section IV provides details of our implementations and section V reports the results. Finally, Section VI presents the conclusion.

II. TOPIC MODEL FOR RAILWAY STATION SAFETY

Text data is essential nowadays more than before, which is valuable and can be easy to store in massive amounts to be processed and mining [9]. Using social media is expanding from the public, and the customer's reviews and reactions are necessary and powerful tool for quality services, sustainable tourism [10] and transport and other aspects such as maintenance. Many points can be raised from such technology of data mining see Figure 1. For instance, the call data which is valuable and raw for long-term history safety data contains many inputs such as risk indicators, the time and date of the week or the seasons. This big data can be classified by different methods, which contain information on safety hazard, can be used to reduce accidents, and form a proactive analysing approach [11].

Safety history is a rich source of knowledge discovery and risk management analysis. For instance, investigation reports after accidents by a responsible authority or expert person, are one of the most popular safety actions that it evaluates and analysis the accidents causes and the consequence of the risk which be very effective for analysing the behaviour, hidden risk cause and lessons can be learned. The text data has many source forms including social media, emails and call recording, such data exist in a raw and unstructured status which requiring transfer and cleaning as part from topic modelling to capture the needed information. A framework based on textual sources data using AI algorithms to

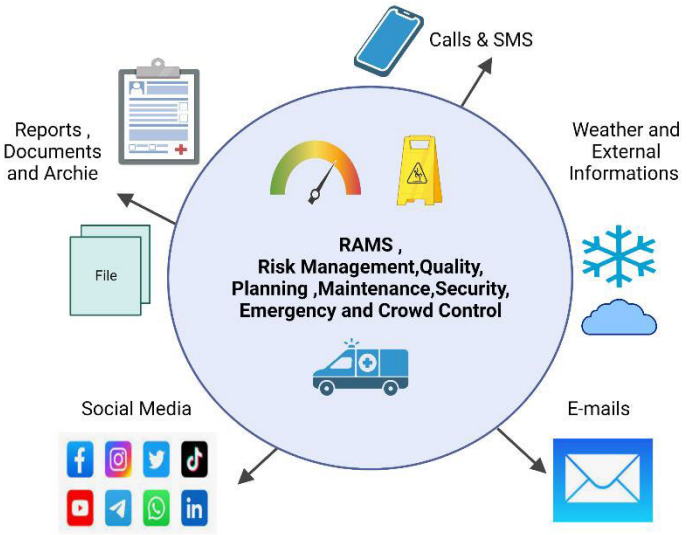

FIGURE 1. The sectors that can benefit from different textual data.

build a tag recordation system from safety documents been suggested see Figure 2. Such method has ability to explore and digest the complete history, it has powerful to tracking, navigate through time to reveal how specific events have changed and can be adapted to many kinds of data. Moreover, to enable automation and digitisation concepts, currently, more texts are available online and the human do not have ability to read, analysis, explore and study how connected to each other, such flow of textual. The topic model is fit to facilitate such issues and annotate large archives of records [3].

The lesson must be learned to prevent repeat accidents in the stations, and a massive effort happened in the field for controlling the issues, and recommendations from investigations have been yielded for high safety level. Usually, many reports and or document been recorded and were presented initiated from risk assessment until the accident investigations report from different organisations which is narratives are indispensable. Regardless of whether or not the text data is structured, many challenges have been expected, such as, massive data, time, cost, the shortage of experts and the context in the documents which may has nonstandard terms. These challenges and more can be decreased by the intelligent use of Deep Learning methods to automate and analysis as a part of the process [12].

III. RELATED WORK

Despite the scatter of applying such method and the differences in terms been using in the literature, there is a shortage of such applications in the railway industry. Moreover, the NLP has been implemented to detect defects in the requirements documents of a railway signalling manufacturer [13].Also, for translating terms of the contract into technical specifications in the railway sector [14]. Additionally, identifying the significant factors contributing

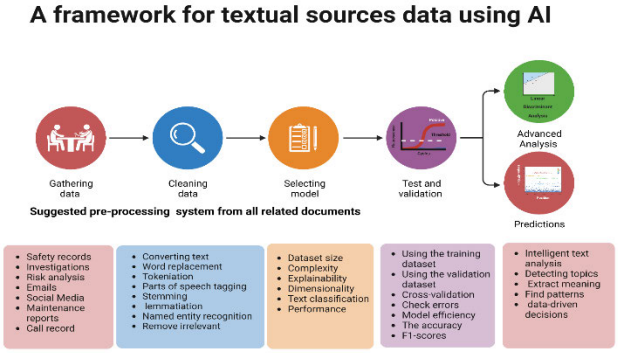

FIGURE 2. Framework of the topic suggested model and data mining cleaning procedure steps (Text Cleaning and Pre-processing).

to railway accidents, the taxonomy framework was proposed using (Self-Organizing Maps – SOM), to classify human, technology, and organization factors in railway accidents [15].Likewise, association rules mining has been used to identify potential causal relationships between factors in railway accidents [16]. In the field of the machine learning and risk, safety accident, and occupational safety, there are many ML algorithms been used such as SVM, ANN, extreme learning machine (ELM), and decision tree (DT) [7], [17].

Scholars have been conducted the topic modelling in, where such method has been proved as one of the most powerful methods in data mining [18] many fields and applied in various areas such as software engineering [19], [4], [20], medical and health [21], [22], [23], [24] and linguistic science [25], [26], etc., Furthermore, from the literature It has been utilised this technique in for predictions some areas such as occupational accident [17], construction [8], [27], [28] and aviation [29], [30], [31]. For Understand occupational construction incidents in the

construction and for construction injury prediction the method been conducted [32], [33], for analysing the factors associated with occupational falls [34], for steel factory occupational incidents [35] and Cybersecurity and Data Science [36]. Moreover, From 156 construction safety accidents reports in urban rail transport in china risks information, relationships and factors been extracting and identified for safety risk analysis [37]. From the literature it has been seen that,there is no perfect model for all text classifications issues and also the process of extracting information from text is an incremental [38], [11]. In the railway sector, a semi-automated method has been examined for classifying unstructured text-based close call reports which show high accuracy. Moreover, for future expectations, it has been reported that such technology could be compulsory for safety management in railway [11]. Applying text analysing methods in railway safety expected to solve issues such as time-consuming analysis and incomplete analysis. Additionally, some advantages have been proved, automated process, high productivity with quality and effective system for supervision safety in the railway system. Moreover, For the prevention of railway accidents, machine learning methods have been conducted. Many methods used for data mining including machine learning, information extraction (IE), natural language processing (NLP), and information retrieval (IR). For instance, to improve the identification of secondary crashes, a text mining approach (classification) based on machine learning been applied to distinguish secondary crashes based on crash narratives, which appear satisfactory performance and has great potential for identifying secondary crashes [39]. Such methods are powerful for railway safety, which aid decision-maker, investigate the causes of the accident, the relevant factors, and their correlations [40]. It has been proved that text mining has several areas of future work development and advances for safety engineering railway [41].

Text mining with probabilistic modelling and k-means clustering is helpful for the knowledge of causes factors to rail accidents. From that application analysis for reports about major railroad accidents in the United States and the Transportation Safety Board of Canada, the study has been designating out that the factors of lane defects, wheel defects, level crossing accidents and switching accidents can lead to the many of recurring accidents [42]. Text mining is used to understand the characteristics of rail accidents and enhance safety engineers, and more to provide a worth amount of information with more detail. An accident reports data for 11 years in the U.S. are analysed by the combination of text analysis with ensemble methods has been used to better understand the contributors and characteristics of these accidents, yet and more research is needed [41]. Also, from the U.S, railroad equipment accidents report are used to identify themes using a comparison text mining methods (Latent Semantic Analysis(LSA)and Latent Dirichlet Allocation(LDA)) [43]. Additionally, to identify the main factors associated with injury severity, data mining methods such as an ordered probit model, association rules, and classification and regression tree (CART) algorithms have been conducted. Using the U.S accidents highway railroad grade crossings database for the period 2007–2013, where Some factors have been discussed such the train speed, age, gender and the time [44]. In recent years, the revolution of big data is opportunities in the railway industry, and that is opening up for safety analysis depends on data [45], so, the approach to proactively identify high-risk scenarios been recommended such as applying the Natural Language Processing (NLP) analysis [46].

From Big Data Application Case A Supervision System has been introduced as a significant role tool in railway safety supervision system. Applying Text Mining Methods in Railway Safety from accident and fault analysis reports been conducted [47]. Also, As well as big data and natural language is an opportunity should be to use for processing for Analysing Railway Safety, NLP framework for analysing accident data been explained using investigation reports of railway accidents [48].Moreover, for Fault Diagnosis in Railway System, classification of maintenance text been proposed using (LDA) algorithm [49], and to improve the fault diagnosis performance [50]. In China railway, for prediction passenger capacity, the social network text data have been used with a combination of text mining and deep learning which show a good accuracy rate [51]. Also from the Chinese Railway, natural language processing has been applied for extraction and analysis of risk factors from accident reports [52]. In the context of deep learning, Data From 2001 to 2016 rail accidents reports in the U.S. examined to extract the relationships between railroad accidents' causes and their correspondent descriptions. Thus for automatic understanding of domain specific texts and analyze railway accident narratives, deep learning has been conducted, which bestowed an accurately classify accident causes, notice important differences in accident reporting and beneficial to safety engineers [53].Also text mining conducted to diagnose and predict failures of switches [54]. For high-speed railways, fault diagnosis of vehicle onboard equipment, the prior LDA model was introduced for fault feature extraction [55] and for fault feature extraction the Bayesian network (BN) is also used [56]. For automatic classification of passenger complaints text and eigenvalue extraction, the term frequency-inverse document frequency algorithm been used with Naive Bayesian classifier [57].

IV. THE LATENT DIRICHLET ALLOCATION

Stations as ML and natural language processing (NLP), topic method, Latent Dirichlet Allocation (LDA) are a kinds of statistical approach for defining the abstract ''topics'' that occur in a collection of context. The concept is to capture the text from multiple topics in the documents, the document is explained as a unique mixture of topics with different proportions see Figure 3, where different colure keywords from accident investigation report documents which exhibit multiple topics. Some terms are highlighted as examples such as the time, date and accident title or causes, and the topic is a distribution over a fixed vocabulary. This analysing

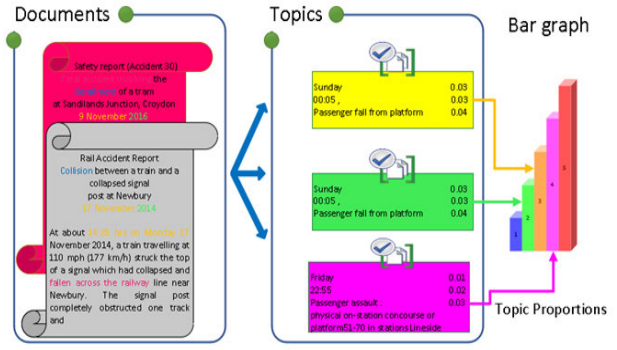

FIGURE 3. Illustrative graph for the steps of latent Dirichlet allocation (examples not from real data).

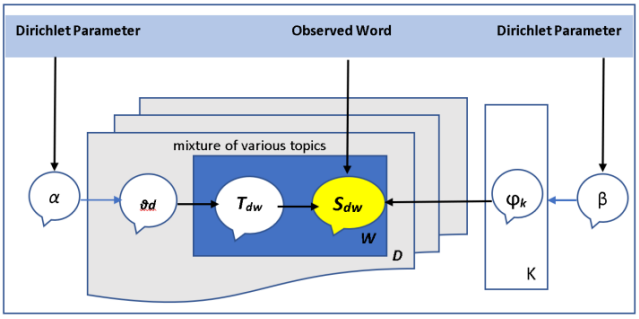

FIGURE 4. Plate diagram (Graphical model) for LDA with Dirichlet-distributed topic-word distributions.

present the ability to manage and summarize the textual data in automated real time manner [3].

The power of machine learning has the ability to learn, predict and describe qualitative and quantitative patterns lying in data such root cause of accidents, which leads to the study of hidden knowledge and the correlation factors in accidents in the railway or other fields. the methods such clustering and k means clustering been used for detect text from unstructured data [58], [5]. LDA as flexible generative probabilistic framework, assumes that each document can be demonstrated as a probabilistic distribution over latent topics, and that topic distribution in all documents take part in a common Dirichlet prior. Any latent topic in the LDA model is likewise demonstrated as a probabilistic distribution over words and the word distributions of topics participate in a common Dirichlet prior. As a generative system, the data from such process includes hidden variables which is a joint probability distribution over both the observed and hidden random variables. The process is executed via that the hidden variables (topic structure) given the observed variables (the words of the documents) as the conditional distribution (posterior distribution) [59]. The model can be described with the notation presented in Table 1 and in the Plate diagram or the graphical model shown in Figure 4 which are means of explaining the probabilistic theories behind LDA mode.

The topics are $\phi_{\mathbf{1:k}}$ and each ϕ_{k} is the distributions over words. The topic proportions document $\boldsymbol{\theta_d}$ for the dth document and $\boldsymbol{\theta_{dk}}$ is topic proportion for topic K in document d.

TABLE 1. Mathematical notation for LDA model in the field of text mining.

Notation	Description
D	Number of documents
Wd	Number of words in document d
Tdw	Topic of the (w-th) word in document of (d), and S is the specific word
Sdw	The specific word in document (d) for topic (t)
α	The parameter of the Dirichlet prior on the per-document topic distributions
θd	The topic distribution for document (d)
θdk	The topic proportion for topic k in document(d)
ϕk	The word distribution for topic k
β	The parameter of the Dirichlet prior on the per-topic word distribution
K	Number of topics and k

The observed words for document $\boldsymbol{d}$ are the $\boldsymbol{s_d}$ *and* $S_{\mathbf{dw}}$ is the *nth* word in document *d,* which is an component from the fixed vocabulary. Thus, the hidden and observed variables corresponds to generative process for LDA can be written in the particular mathematical form of the joint distribution as fallow:

$$p(\boldsymbol{\beta}_{\mathbf{1:K}}, \boldsymbol{\theta}_{\mathbf{1:D},}, \mathbf{Z}_{\mathbf{1:D}}, \mathbf{w}_{\mathbf{1:D},}) = \prod_{\mathbf{i=1}}^{\mathbf{K}} p\left(\boldsymbol{\beta}_{\mathbf{i}}\right) \prod_{\mathbf{d=1}}^{\mathbf{D}} p(\theta_{\mathbf{d},}) (\prod_{n=}^{N} p\left(\mathbf{Z}_{\mathbf{d,n}} \mid \boldsymbol{\theta}_{\mathbf{d},}\right) * p\left(\mathbf{w}_{\mathbf{d,n}} \mid \boldsymbol{\beta}_{\mathbf{1:K}}, \mathbf{Z}_{\mathbf{d,n}}\right))$$

The $\boldsymbol{S_{dw}}$ remarked in yellow as a variable which the only words are observed, $\boldsymbol{\alpha}$ is a matrix where each row is a document, and each column represents a topic and $\boldsymbol{\beta}$ is a matrix where each row represents a topic and each column represents a word. Both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters of the respective Dirichlet distributions. For computational the conditional distribution (posterior distribution) of the hidden variables which is the topic structure given the observed documents, the posterior can be formed as:

$$p\left(\boldsymbol{\beta}_{\mathbf{1:K}}, \boldsymbol{\theta}_{\mathbf{1:D},}, \mathbf{Z}_{\mathbf{1:D}}, \mathbf{w}_{\mathbf{1:D},} \mid \mathbf{w}_{\mathbf{1:D}}\right) = \frac{p(\boldsymbol{\beta}_{\mathbf{1:K}}, \boldsymbol{\theta}_{\mathbf{1:D},}, \mathbf{Z}_{\mathbf{1:D}}, \mathbf{w}_{\mathbf{1:D}})}{p(\mathbf{w}_{\mathbf{1:D}})}$$

in fact, different models of approximate inference algorithms can be analysed for LDA, for instance, Laplace approximation, variational approximation, and Markov chain Monte Carlo.in spite of the fact that the posterior distribution is intractable for exact inference, In this part, we present a simple model variational algorithm for inference in LDA [59], [5].

The LDA method has the strength to recognise sub-topics for risks range formed of many causes and represent each of the risks in an array of topic distributions. With LDA, the terms in the set of documents, produce a vocabulary that is then utilised to discover hidden topics.

V. PREPARING DATA

The textual data have some key information can be used such as the time, description of the accidents, location and the range age of the victim. The time of accidents occurred been

divided as the Parts of the Day for more mining to capture accurate times. The Morning from 5 am to 12 pm (Noon), Afternoon from 12 pm (midday) to 5pm (17:00). The Evening from 5 pm (17:00) to 9 pm (21:00) and the Night from 9 pm to 4 am.

The data set containing the fatalities occurring at rail stations between 01/01/2000 and 17/04/2020. The following information is almost available for each accident including hazardous description, and the below text data:

- The day of the week the fatality occurred on
- The time of day of the fatality occurred
- Information on the cause of the fatality
- The age range of the deceased
- The site type where the fatality occurred.
- The physical environment where the fatality occurred.

From the RSSB the raw data set has more than 2250 accidents that been registered in the railway stations in ten years. However, the data have been divided into datasets, for more clarifications and practices in precise views of future research scope; for example, the suicide dataset been excluded from this work data set. As apart from the data mining the raw data need to be processed to extract the knowledge for Text Cleaning and Pre-processing, it has been known that many documents have additional words like misspelling, punctuation, stop words, slang, and others which affect the algorithms and topic model results performance. Some of the techniques have been remarked to Pre-processing text data, convert the context to formal language and remove any Noise as follows:

- Tokenization which is breaks the context of text into meaningful elements called tokens, aiming for investigation of the words in a sentence by data gets split into parts [60], [61].
- Stop and noise Words which is the words that do not form a key in the classification algorithms, for example (a, after, about etc.,), so it needs to be removed [62].
- Capitalization where the words or Abbreviation written in capital letters, so converters to lower case can help account for such exceptions [63].

For Text Cleaning and Pre-processing, it has been known that many documents have additional words like misspelling, stop words, slang,son and others which affect the algorithms and topic model results performance [64]. Thus, for improving the quality of the data set and the model performance, filtering configuration is used, which allow selecting the fields that considered in the modelling or not, where some information's in the unprocessed data can lead to fuzzy overview, noise and or has missing values, favourably, in this data set no missing value. For topic model configuration, the non-language characters, Non-dictionary, and Numeric digits are excluded from the analyses. Also, uninformative words such as at, on, and or are removed.

From the view visualisation of the data set the times of nights and afternoon capture most accidents then morning and afternoons. However, the day of accident not easy to identify, also the adults seems to be involved in the accidents more than children. The lineside is the location which gathering more accidents and then the stairs /bridge and escalators see Figure 5.

VI. MODEL ANALYSING

A DT is a determination support tool that applies a tree-like pattern of decisions and their likely outcomes [40], [53]. There are many possible (ML) approaches towards safety analysis. More exactly, we train a DT to classify the accidents and the patterns that occurred in these accidents in the stations [41], [52]. This model is applied to a wide variety of data, and it is preferable because its structured rules are simple to follow and understand. This technique is used to classify instances by classifying them based on feature values (Yuan and Shaw). The two general types of DTs are classification (where the class variable is discrete), and regression (where the class variable is continuous) [42], [43], [53], [54]. After, the data sets are uploaded and then a DTs model is designed and visualised. The DT for the predictive model provides a visualisation of the prediction case. The DTs have useful information; branches are used to make a branching decision. It shows the decisions that led to a given prediction. The tool presents the model prediction path on the side of the tree which gives this tool an advantage.

A. TEXT ANALYSIS

The dataset that been used in this work has key text attributes, and information's as the day of week and time of day that accident had occurred, also including the hazardous event description and Precursor description. Moreover, the age of the victim and the Site type with data of physical environment information been remarked. Form a quick overview noticed to the cloud, which is a key visualisation since it has summaries a bulk of the text. Generally, variety of accidents has been occurred such as fall off from train platforms or in the gab, trapped in the door, struck by trains, and electrical shuck or suicide and so on. These accidents can occur when alighting from or boarding to the train or also when there is no train stopped at the station.

The Saturday been remarked linked to the accidents in the stations which is a day usually not crowded such the workdays, but may that public be going out more as it is off the workday. The details and reasons behind that need more investigations such as the factor of deficiency of assistance staff and intoxication impact. The night and evening times obtain accident more than morning and afternoons. This also raised many factors to need to be redefined, such as the light condition, the seasonality and weather effect. More information can be gained such as the trespass accidents, which one of the brightness words in the cloud which linked to the station and the contact with the vehicles of the train which present the importance of isolate passenger from the train at safe distance. In another view, this raises the query of overcrowding risk effect which may force the people to be close to the trains and track. Moreover, infrastructure been appeared which reflect the stations age and the impact of intensive usage. The lineside in the platforms is the hot spot place that interacts the human with the machine and forms an accidents trap. The consequence like such crushed is appear which raise the engineering solution to indicate the objects and stop trains in accidents situations [6], [65].

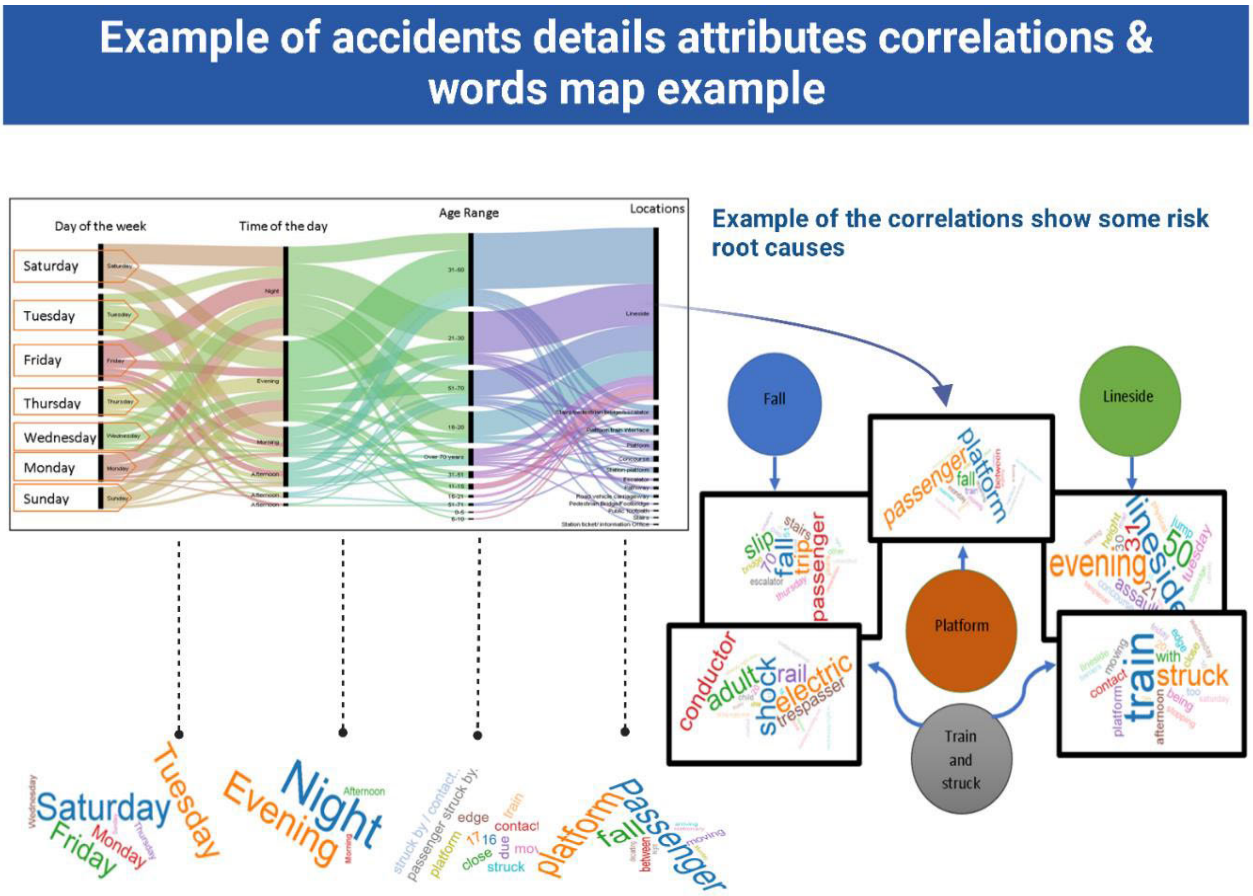

FIGURE 5. Example of an alluvial diagram of raw accident data: accident details, locations, day, time, and passenger age and the examples from dataset inputs word cloud visualization.

In addition, to review and visualise graphic statistics of the dataset, that shows the distribution of the accidents among factors. The time of the day (AM &PM), the range of the time Afternoon, Morning and Night produce more solid information, the children and elderly passenger been involved in accidents in the morning. Nevertheless, the Night-time captured most accidents (See Figure 6-b). Moreover, these factors overlapping with the days of the week, some accidents been occurred on Saturday morning and seems that Afternoon is safe over all the week (See Figure 6-c). Their many valuable illustrations can be found from the details of data, yet the details of the detailed need to be gathered in the future which proved the importance of the data in such case of safety analytics. Clearly, if this approach been considered in the future safety and reporting systems, the system of reporting and gathering data will be improved to be more valuable for advanced analytics such as AI methods.

Moreover, it is expected that passenger behaviour is primary towards inside accidents, as passengers on the platform tend to walk or stand near the platform edges to avoid crowded areas, and there are others who run to catch the trains or stand too close in order to get on the train before others, and this can be coupled with slow responses of moving trains, or little time to react. From frequency view of the words, the risks related to fall/slip and trip, struck/crashed, train and platform with the passenger, are noticed as risk joint words from the details of the accident occurrence.

Even though the available data does not provide a deep understanding of the causes, the information was analysed independently and correlated with all the input attributes, and the factors related to the outcomes show the importance of the details of data for each unwanted event for future data gathering. For example, in a fall in accident cases, the position (status) that passengers had at the time of falling and the position (forward position) that land in is key factor impact also impact the rail track as sharp and solid on the seriousness of the accident's results, and the details of the lights, flooring and the platform slope and the infrastructure status. To detect the relevant topics within the text and learn from the topics underlying a collection of documents, the Latent Dirichlet Allocation (LDA) algorithm is implemented, which has some configurations (see Table 2).

The nodes have been created in the topic map and they illustrate each topic via word probability with different sizes and colours (see Figure 6-a). This collection of models is powerful for visualisation, and each circle presents a topic, the size notes how common the topic is in the data, and the

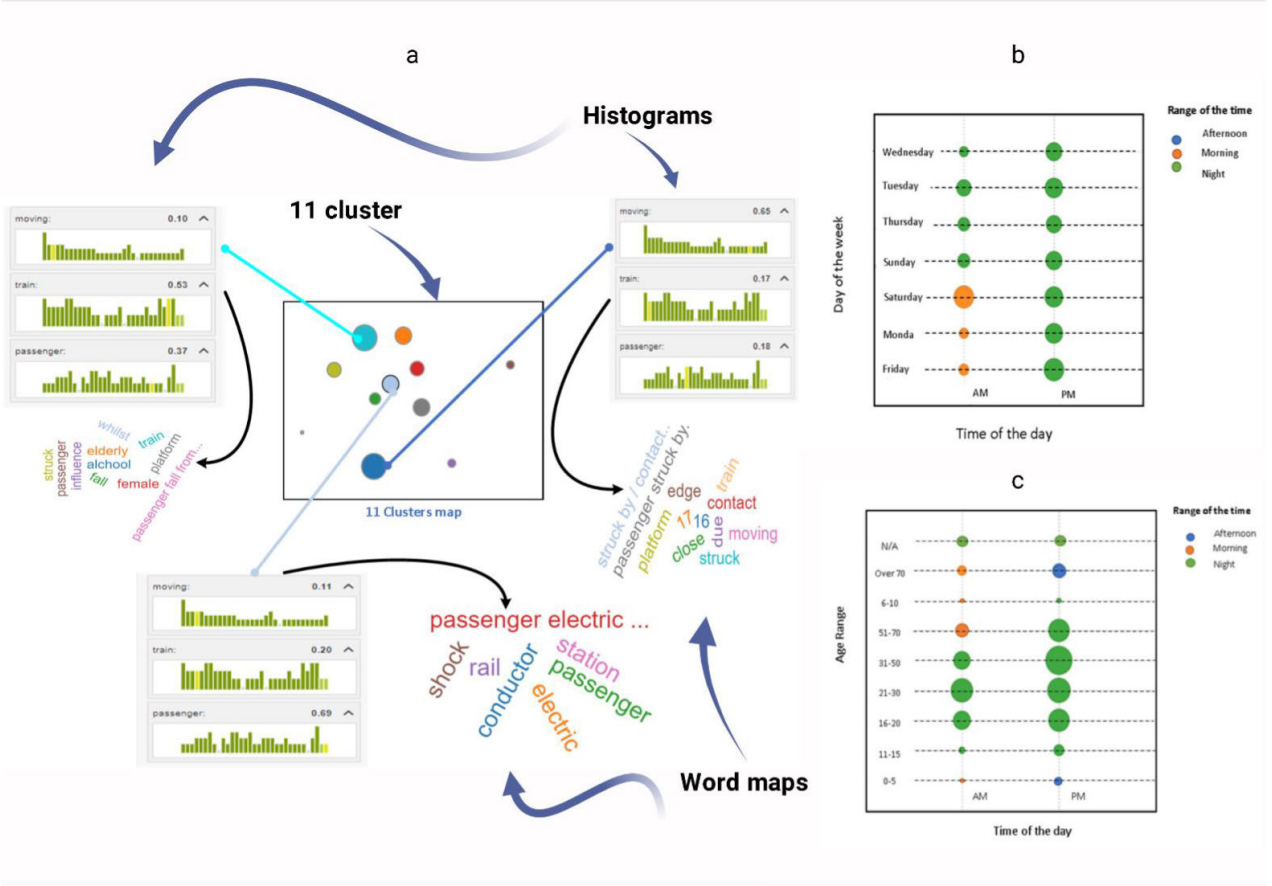

FIGURE 6. (a) 11 cluster maps with examples of word maps and histograms, (b) Comparisons of input subjects for different factors (time of the day and the day of the week with the age range), (c)Comparisons of input subjects for different factors (time of the day and the range of the time with the age range).

TABLE 2. Set up configurations for text analytics.

The model configuration	Details
Number of topics	10
Number of top terms	10
Language	English
Tokenize	Active for all text
Stemming	Active for all text
Filter terms	Active (Non-dictionary +non-language words)
Sampling	100% (163) accidents-deterministic)

topics that are close together are semantically linked in the data.

In selecting the passenger as a topic in specific, the word maps and the distribution show that this topic linked to falling accidents in the platform and the departing, more than the arriving at night. Additionally, for instance, the relation with the fall and slip/trip as a group, and in many locations such as stairs, escalators and bridges, and the afternoon time connect with the accidents caused by contact with trains. This presents some query not been answered, such as if this occurred because the passenger is running to catch up the train or their other hidden reasons such as narrow lineside, slippery floor and congestion. These analytics illustrates the importance of the details of the accidents reporting and the key factors that can be gain from such a method to provide safety measures.

Each topic node has many probabilities of the words. More details can get from the topic distribution, for example for the selection of the words, the most associated with the specific topic will be captured, as long as the words changing, the topic distribution updated. Falling occurred in many locations such as escalators and bridges and was linked with elderly passengers, which opens up the importance of accessibility in the stations to most vulnerable people like the disabled, aged and families, also the age of mature passengers (31-50) held more dangers in the platforms in the stations. Electrical shock is linked with conductors and small age groups, which raise the importance of standards review for safety related to electrical equipment in the stations. The contact of the passengers and the train in the platform while the train moving is shows an associated cause between topics of train and moving. Additionally, there is a correlation between the fall and slip/trip as a group, and many locations such as stairs, escalators and bridges, and the edge in the lineside, as well as the contact caused with trains. Such analytics presents some mysteries, which required detailed reporting to be answered in the future by the safety community, such as if this occurred because the passenger is running to catch the train or other hidden reasons such as narrow linesides, slippery floor, floor damage, or congestion. These analytics illustrate the importance of the details of accident reporting and the

TABLE 3. Topic batch with topic distribution examples.

The text input information examples						
Day of week ▶ Time of Day ▶ Hazardous Event Description ▶ Age Band ▶ Physical environment						
Sunday Evening Passenger slip, trip or fall in a station, his age 31-50, stairs/pedestrian						
Monday Night, MOP (adult trespasser) electric shock (conductor rail) ,21-30 Pathway						
Wednesday Night, MOP (trespasser) struck/crushed by train while on railway infrastructure at a station ,21-30 Lineside						
Thursday Evening, Passenger slip, trip or fall in a station ,31-50 Stairs/pedestrian bridge/escalator						
The predictions probabilities						
struck	railway	fall	platform	trespasser	train, struck	lineside
0.0045	0.0053	0.6862	0.0427	0.0053	0.03588	0.1458
0.0060	0.0037	0.0060	0.0324	0.0317	0.00378	0.0392
0.1358	0.1758	0.0037	0.0037	0.2309	0.01962	0.0347
0.0241	0.0037	0.7537	0.0453	0.0060	0.00453	0.1117

TABLE 4. The probabilistic description example of the words among the topics.

Physical environment	Stairs/pedestrian bridge/ Escalator	Lineside	Lineside	Lineside
Age band	31-50	51-70	16-20	21-30
Precursor description	Passenger slip, trip or fall on stairs	Passengers fall between moving train	MOP (adult trespasser) electric shock (conductor	MOP (adult trespasser) electric shock (conductor
Sub Hazardous Event description	Passenger slip, trip or fall (stairs)	Passengers fall between moving train	MOP (adult trespasser) electric shock (conductor	MOP (adult trespasser) electric shock
Hazardous Event description	Passenger slip, trip or fall in a station	Passengers fall between train and platform	MOP (adult trespasser) electric shock (conductor rail)	MOP (adult trespasser) electric shock (conductor rail)
Time of Day	Evening	Afternoon	Night	Morning
Day of week	Thursday	Tuesday	Monday	Saturday

TABLE 5. The probabilistic description example of the words among the topics.

fall	platform	trespasser	train, struck	lineside	struck	railway	train, struck	station	shock
1	0.04532	0.00604	0.00453	0.11178	0.02417	0.00378	0.00453	0.04	0.01
0	0.55363	0.00378	0.19109	0.0574	0.00529	0.0068	0.03852	0.039	0
0	0.02946	0.05891	0.00604	0.0136	0.00831	0.00604	0.04834	0.088	0.74
0	0.00378	0.05136	0.00529	0.09215	0.00755	0.03021	0.01284	0.144	0.65
0	0.63435	0.00382	0.13359	0.00382	0.00611	0.01145	0.13206	0.056	0.01

key factors that can be gained from such a method to provide safety measures. Each topic node has many probabilities of the words. More details can be gained from the topic distribution; for example, by selecting the words, the words most associated with the specific topic will be captured; as long as the words are changing, the topic distribution is updated. The development of data analytics can be with advanced techniques such as linking the outcomes with the dictionaries with the meaning and providing suggestions for the designer and analyst as an expert system or providing a recommendation for the safety employees as a tool of an expert system. Also, the topics modification and each topic have the distribution of the common words (see Table 4).

Applying a batch topic distribution with some sentences that contain information leads to predicting the properties related to the inputs (see Table 5 & 6 below).

TABLE 6. Cluster outcomes of the textual data.

Centroid name	Age band	Physical environment	Day of week	Time of Day
Global	31-50	['; Platform', 'Platform']	['Saturday']	['Morning']
Cluster 0	31-50	['Lineside', 'Station platform']	['Friday', 'Thursday', 'Wednesday']	['Evening', 'Morning', 'Night']
Cluster 1	31-50	['Lineside']	['Monday', 'Saturday', 'Tuesday']	['Evening']
Cluster 2	31-50	['Road', 'Road vehicle carriageway', 'carriageway', 'vehicle']	['Tuesday']	['Evening']
Cluster 3	21-30	['Lineside', 'Other', 'Other; Platform']	['Monday', 'Saturday', 'Sunday', 'Tuesday']	['Evening', 'Night']
Cluster 4	31-50	['; Platform', 'Lineside', 'Platform', 'Platform train interface', 'Station', 'Station platform', 'train']	Tuesday']	['Afternoon']
Cluster 5	31-50	['; Platform', 'Other', 'Other; Platform', 'Platform', 'Platform train interface', 'Station', 'Station platform', 'train']	['Friday', 'Monday', 'Saturday', 'Wednesday']	['Afternoon', 'Night']
Cluster 6	31-50	['Concourse', 'Footbridge', 'Platform train interface', 'Stairs', 'Stairs/pedestrian bridge/escalator', 'bridge', 'escalator', 'pedestrian', 'train']	['Saturday', 'Sunday', 'Tuesday']	['Evening', 'Morning']
Cluster 7	Over 70 years	['Concourse', 'Stairs', 'Stairs/pedestrian bridge/escalator', 'Station', 'bridge', 'escalator', 'pedestrian']	['Friday', 'Thursday']	['Afternoon']

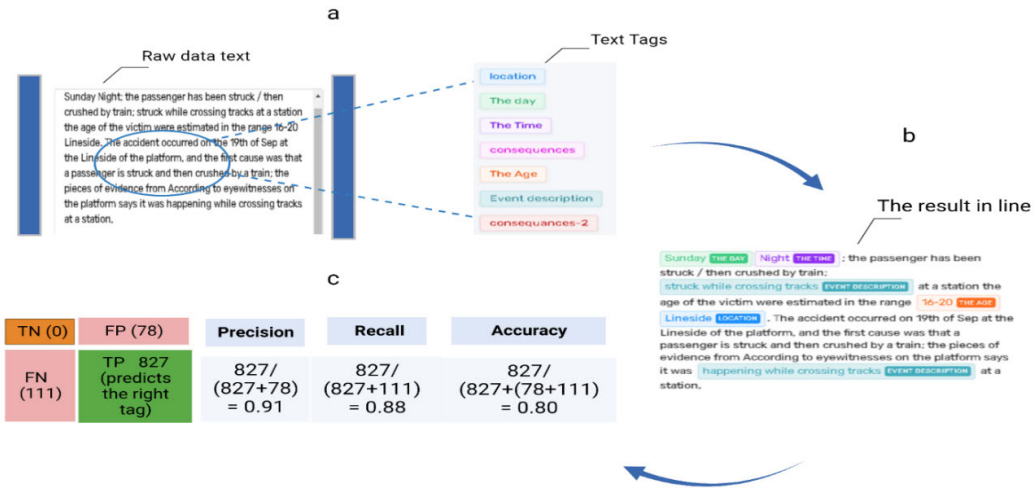

FIGURE 7. (a) Text example for extracting information from trained topic model, (b) Classifier of overall statistics (c) Confusion matrix.

B. CLUSTERING AND VALIDATION

Evaluating for further analysis, the cluster has been conducted in which the data set can be used. In this case, the G-Means are considered to find the best number of cluster group with a critical value (5); contrarily, K-Means can be used for a specific number of clusters (K). 11 clusters are used in the cluster algorithm to show the largest problem, and the correlations in the topic. However, 8 clusters appear as the most common correlations in the topic (the clusters with the maximum total of cases). The second cluster reveals the elderly passengers were involved in some accidents at the time when the train is moving. The next cluster shows the electricity risk more specific to the conductor's parts in the platforms. These accidents required details to obtain more safety measures. Analyses can provide causes roots, correlations, and any hidden patterns, for

example, electric shocks, including contact with Overhead Line Equipment (OHLE) or Overhead Contact Line (OCL), which may occur accidentally by carrying long objects (selfie stick/conductive materials) or vandalism and trespassing. Such details will improve the standards and obligation to add more protections for the public, passengers, and workforce at the PTI. All probability of the instances fields is shown in Table 6 below, which can be used as workflow in future projects and generalised. A guide system of the context can be converted to be numerical and cover a huge range of the stations.

For testing, topic evaluation by a short text, which is a text describing an accident, has been used to extract the information, which shows the ability to capture information from the textual content. This depends on the tags (Labelled Itemset) being used for the training model, Labels like the day, the time, event description and location (Figure 7-a). For the overall trained model, statistics measures present excellent outcomes for all the tags, as shown in (Figure 7-c). The analysis presents the power of the approach to be an expert system and guide the reporting process, also noting any hidden root causes of the accidents.

Some data are accurately detected such as the day, location, ages and time of the occurrence of the accident. Others, such as details of the accidents, show less accuracy to detect, but these types of data can be captured by providing more training data, that will in turn provide more balance for all tags and reflect on accuracy.

VII. DISCUSSION

Applying such a method shows the ability and the power of the new technology in the safety of the railway industry which has not been used widely to enhance safety and risk management. The systematic reducing of accidents in the station is beneficial for the public and stockholder, sustainability, and the safety community. Topic moulding is a proactive approach, where the input can be analysed in real-time, and actions can be taken. Moreover, multi-input can be used in parallel with supporting decision-makers. This approach reduces the dependence on experts, where they are costly and not available all the time and reduce the impact of manpower knowledge retirement. The development is key for many aspects which reflect on the quality, reliability, and satisfaction for both workers and passengers. The increase of technology such as AI and IoT with the growth of data requires more investment and research of such methods. Even with the limitation of data and delayed application in the field, text from data is analysed via text modelling which opens the novel approach of applying technology in the safety system. Such types of data are essential as they contain the history of safety such as risk evaluation, accident reports and periodic safety analyse documents, and can also use live text from media or calls. Moreover, social media has a textual source of information that can be harvested and analysed as it is related to safety, security, and quality. The method provides support for the safety authority of the decision-maker from many sides, including improving the service, quick response rates and advanced analysis. Also, the system can build an expert system in a specific area such as the railway stations and learning from the new documents and can cover an entire country's stations as well as pr training the system from many sources. In the digitalisation concept, the integration between the source of data is possible, which forms a smart safety system in the railway industry. This novel analytics opens a new window for applying AI technologies in the field. Also, railway organizations can use a such method to cover all RAMS parameters while safety overlaps with maintenance, reliability and many other factors. Analysing the safety in the stations is part of RAMS analysis of railway networks which can help managers find the key components of failure in the safety of the network [66], [67], [68].

The LDA provides a statistical model with the ability to learn, and this method does not only deal with the huge data in real-time automatically. It can also provide an expert system and decision support for the safety authority for the researcher. This concept corresponds with the future rail digitalization and the BD revolution. The fixability, effectiveness, and accuracy raise the importance of gathering more data in the station with all the possible details. The accessibility, privacy, skills and the IT infrastructure are some hurdles for the short term, but which it is expected the industry will overcome in the medium and long term. Finally, there are concerns with the vast acceleration of AI exploring texts and language, like the generative pre-trained transformer model (ChatGPT). However, using these models responsibly and cautiously and considering appropriate measures to mitigate potential risks is essential. Moreover, human health and safety in work or other areas, such as railway safety that can save lives, must be highly prioritised.

VIII. CONCLUSION

Topic models have an important role in many fields and in such case of safety and risk management in the railway stations for texts mining. In Topic modelling, a topic is a list of words that occur in statistically significant methods. A text can be voice records investigation reports, or reviews risk documents and so on.

This research displays various cases for the power of unsupervised machine learning topic modelling in promoting risk management, safety accidents investigation and restructuring accidents recording and documentation on the industry-based level. The description of the root causes accident, the suggested model, it has been showing that the platforms are the hot point in the stations. The outcomes reveal the station's accidents to be occurring owing to four main causes: falls, struck by trains, electric shock. Moreover, the night time and days of the week seems to contact to the risks are significant.

With increased safety text mining, knowledge is gained on a wide scale and different periods resulting in greater efficiency RAMS and providing the creation of a holistic perspective for all stakeholders.

Application of the unsupervised machine learning technique is useful for safety since, which is solving, exploring hidden patterns and deal with many challenges such as:

- Text data from many perspectives and in unstructured forms

- Power for discovery, dealing with missing values, and spot safety and risk kyes from data
- Smart labelling, clustering, centroids, sampling, and associated coordinates
- Capture the relationships, causations, more for ranking risks and related information
- Prioritisation risks and measures implementations
- Aid the process of safety review and learning from the long and massive experience.
- Can be used the scale and weighted as configuration options which can be used for assessing risks.

Although this paper highlights the innovative of unsupervised machine learning in accidents classification of railway accidents and root cause analyses, it is a necessity to focus on expanded research on the huge data topics concerning the diversity of the station's locations, size and safety cultures and other factors with further techniques of unsupervised machine learning algorithms in the future. Finally, this research enhances safety, but it raises the importance of data in text form and suggests redesigning the way of gathering data to be more comprehensive.

REFERENCES

[1] S. Terabe, T. Kato, H. Yaginuma, N. Kang, and K. Tanaka, "Risk assessment model for railway passengers on a crowded platform," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 1, pp. 524–531, Jan. 2019, doi: 10.1177/0361198118821925.

[2] *Annual Health and Safety Report 19/2020*, RSSB, London, U.K., 2020.

[3] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.

[4] M. Gethers and D. Poshyvanyk, "Using relational topic models to capture coupling among classes in object-oriented software systems," in *Proc. IEEE Int. Conf. Softw. Maintenance*, Sep. 2010, pp. 1–10, doi: 10.1109/ICSM.2010.5609687.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, nos. 4–5, pp. 993–1022, Mar. 2003, doi: 10.1016/B978-0-12-411519-4.00006-9.

[6] H. Alawad, S. Kaewunruen, and M. An, "A deep learning approach towards railway safety risk assessment," *IEEE Access*, vol. 8, pp. 102811–102832, 2020, doi: 10.1109/ACCESS.2020.2997946.

[7] H. Alawad, S. Kaewunruen, and M. An, "Learning from accidents: Machine learning for safety at railway stations," *IEEE Access*, vol. 8, pp. 633–648, 2020, doi: 10.1109/ACCESS.2019.2962072.

[8] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports," *Autom. Construct.*, vol. 62, pp. 45–56, Feb. 2016, doi: 10.1016/j.autcon.2015.11.001.

[9] J. Sido and M. Konopik, "Deep learning for text data on mobile devices," in *Proc. Int. Conf. Appl. Electron.*, Sep. 2019, pp. 1–4, doi: 10.23919/AE.2019.8867025.

[10] A. Serna and S. Gasparovic, "Transport analysis approach based on big data and text mining analysis from social media," *Transp. Res. Proc.*, vol. 33, pp. 291–298, Jan. 2018, doi: 10.1016/j.trpro.2018.10.105.

[11] P. Hughes, D. Shipp, M. Figueres-Esteban, and C. van Gulijk, "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," *Saf. Sci.*, vol. 110, pp. 11–19, Dec. 2018, doi: 10.1016/j.ssci.2018.03.011.

[12] A. Chanen, "Deep learning for extracting word-level meaning from safety report narratives," in *Proc. Integr. Commun. Navigat. Surveill. (ICNS)*, Apr. 2016, pp. 5D2-1–5D2-15, doi: 10.1109/ICNSURV.2016.7486358.

[13] A. Ferrari, G. Gori, B. Rosadini, I. Trotta, S. Bacherini, A. Fantechi, and S. Gnesi, "Detecting requirements defects with NLP patterns: An industrial experience in the railway domain," *Empirical Softw. Eng.*, vol. 23, no. 6, pp. 3684–3733, Dec. 2018, doi: 10.1007/s10664-018-9596-7.

[14] G. Fantoni, E. Coli, F. Chiarello, R. Apreda, F. Dell'Orletta, and G. Pratelli, "Text mining tool for translating terms of contract into technical specifications: Development and application in the railway sector," *Comput. Ind.*, vol. 124, Jan. 2021, Art. no. 103357, doi: 10.1016/j.compind.2020.103357.

[15] G. Yu, W. Zheng, L. Wang, and Z. Zhang, "Identification of significant factors contributing to multi-attribute railway accidents dataset (MARA-D) using SOM data mining," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 170–175, doi: 10.1109/ITSC.2018.8569336.

[16] Y. Wang, W. Zheng, H. Dong, and P. Gao, "Factors correlation mining on railway accidents using association rule learning algorithm," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, vol. 22, Sep. 2020, pp. 1–6, doi: 10.1109/ITSC45102.2020.9294317.

[17] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents," *Comput. Oper. Res.*, vol. 106, pp. 210–224, Jun. 2019, doi: 10.1016/j.cor.2018.02.021.

[18] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.

[19] S. W. Thomas, "Mining software repositories using topic models," in *Proc. 33rd Int. Conf. Softw. Eng. (ICSE)*, May 2011, pp. 1138–1139, doi: 10.1145/1985793.1986020.

[20] H. U. Asuncion, A. U. Asuncion, and R. N. Taylor, "Software traceability with topic modeling," in *Proc. ACM/IEEE 32nd Int. Conf. Softw. Eng.*, vol. 1, May 2010, pp. 95–104, doi: 10.1145/1806799.1806817.

[21] Z. Jiang, X. Zhou, X. Zhang, and S. Chen, "Using link topic model to analyze traditional Chinese medicine clinical symptom-herb regularities," in *Proc. IEEE 14th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Oct. 2012, pp. 15–18, doi: 10.1109/HealthCom.2012.6380057.

[22] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing Twitter for public health," in *Proc. Int. AAAI Conf. Weblogs Social Media (ICWSM)*, 2011, pp. 265–272.

[23] W. Zhao, W. Zou, and J. J. Chen, "Topic modeling for cluster analysis of large biological and medical datasets," *BMC Bioinf.*, vol. 15, no. 11, pp. 1–11, Oct. 2014, doi: 10.1186/1471-2105-15-S11-S11.

[24] H.-M. Lu, C.-P. Wei, and F.-Y. Hsiao, "Modeling healthcare data using multiple-channel latent Dirichlet allocation," *J. Biomed. Informat.*, vol. 60, pp. 210–223, Apr. 2016, doi: 10.1016/j.jbi.2016.02.003.

[25] S. Bauer, A. Noulas, D. Ó. Séaghdha, S. Clark, and C. Mascolo, "Talking places: Modelling and analysing linguistic content in foursquare," in *Proc. ASE/IEEE Int. Conf. Privacy, Secur., Risk Trust, ASE/IEEE Int. Conf. Social Comput. (SocialCom/PASSAT)*, Sep. 2012, pp. 348–357, doi: 10.1109/SocialCom-PASSAT.2012.107.

[26] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, "Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, vol. 1, 2015, pp. 99–107.

[27] B. Zhong, X. Pan, P. E. D. Love, J. Sun, and C. Tao, "Hazard analysis: A deep learning and text mining framework for accident prevention," *Adv. Eng. Informat.*, vol. 46, Oct. 2020, Art. no. 101152, doi: 10.1016/j.aei.2020.101152.

[28] H. Baker, M. R. Hallowell, and A. J.-P. Tixier, "Automatically learning construction injury precursors from text," *Autom. Construct.*, vol. 118, Oct. 2020, Art. no. 103145, doi: 10.1016/j.autcon.2020.103145.

[29] B. Lyall-Wilson, N. Kim, and E. Hohman, "Modeling human factors topics in aviation reports," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, Nov. 2019, vol. 63, no. 1, pp. 126–130, doi: 10.1177/1071181319631095.

[30] Y. Luo and H. Shi, "Using lda2vec topic modeling to identify latent topics in aviation safety reports," in *Proc. IEEE/ACIS 18th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2019, pp. 518–523, doi: 10.1109/ICIS46139.2019.8940271.

[31] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, "Natural language processing for aviation safety reports: From classification to interactive analysis," *Comput. Ind.*, vol. 78, pp. 80–95, May 2016, doi: 10.1016/j.compind.2015.09.005.

[32] C.-W. Cheng, C.-C. Lin, and S.-S. Leu, "Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry," *Saf. Sci.*, vol. 48, no. 4, pp. 436–444, Apr. 2010, doi: 10.1016/j.ssci.2009.12.005.

[33] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Application of machine learning to construction injury prediction," *Autom. Construct.*, vol. 69, pp. 102–114, Sep. 2016, doi: 10.1016/j.autcon.2016.05.016.

[34] N. Nenonen, "Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database," *Appl. Ergonom.*, vol. 44, no. 2, pp. 215–224, Mar. 2013, doi: 10.1016/j.apergo.2012.07.001.

[35] A. Verma, S. D. Khan, J. Maiti, and O. B. Krishna, "Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports," *Saf. Sci.*, vol. 70, pp. 89–98, Dec. 2014, doi: 10.1016/j.ssci.2014.05.007.

[36] T. Bechor and B. Jung, "Current state and modeling of research topics in cybersecurity and data science," *Syst. Cybern. Inf.*, vol. 17, no. 1, pp. 129–156, 2019.

[37] J. Li, J. Wang, N. Xu, Y. Hu, and C. Cui, "Importance degree research of safety risk management processes of urban rail transit based on text mining method," *Information*, vol. 9, no. 2, p. 26, Jan. 2018, doi: 10.3390/info9020026.

[38] C. van Gulijk, P. Hughes, and M. Figueres-Esteban, "The potential of ontologies for safety and risk analysis," in *Proc. 26th Eur. Saf. Rel. Conf. (ESREL)*, 2017, p. 210, doi: 10.1201/9781315374987-197.

[39] X. Zhang, E. Green, M. Chen, and R. R. Souleyrette, "Identifying secondary crashes using text mining techniques," *J. Transp. Saf. Secur.*, vol. 12, no. 10, pp. 1338–1358, Nov. 2020, doi: 10.1080/19439962.2019.1597795.

[40] H. Hadj-Mabrouk, "Analysis and prediction of railway accident risks using machine learning," *AIMS Electron. Electr. Eng.*, vol. 4, no. 1, pp. 19–46, 2020, doi: 10.3934/ElectrEng.2020.1.19.

[41] D. E. Brown, "Text mining the contributors to rail accidents," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 346–355, Feb. 2016, doi: 10.1109/TITS.2015.2472580.

[42] T. Williams, J. Betak, and B. Findley, "Text mining analysis of railroad accident investigation reports," in *Proc. Joint Rail Conf.*, Apr. 2016, doi: 10.1115/JRC2016-5757.

[43] T. Williams and J. Betak, "A comparison of LSA and LDA for the analysis of railroad accident text," *Proc. Comput. Sci.*, vol. 130, pp. 98–102, Jan. 2018, doi: 10.1016/j.procs.2018.04.017.

[44] H. Ghomi, M. Bagheri, L. Fu, and L. F. Miranda-Moreno, "Analyzing injury severity factors at highway railway grade crossing accidents involving vulnerable road users: A comparative study," *Traffic Injury Prevention*, vol. 17, no. 8, pp. 833–841, Nov. 2016, doi: 10.1080/15389588.2016.1151011.

[45] C. van Gulijk, P. Hughes, M. Figueres-Esteban, R. El-Rashidy, and G. Bearfield, "The case for IT transformation and big data for safety risk management on the GB railways," *Proc. Inst. Mech. Eng. O, J. Risk Rel.*, vol. 232, no. 2, pp. 151–163, Apr. 2018, doi: 10.1177/1748006X17728210.

[46] J. Sresakoolchai and S. Kaewunruen, "Integration of building information modeling and machine learning for railway defect localization," *IEEE Access*, vol. 9, pp. 166039–166047, 2021, doi: 10.1109/ACCESS.2021.3135451.

[47] X. Q. Li, T. Y. Shi, P. Li, F. Gao, and W. G. Xiang, "Application of text mining techniques in railway safety supervision system," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 189, no. 6, Nov. 2018, Art. no. 062009, doi: 10.1088/1755-1315/189/6/062009.

[48] K. N. Syeda, S. N. Shirazi, S. A. A. Naqvi, H. J. Parkinson, and G. Bamford, "Big data and natural language processing for analysing railway safety," in *Human Performance Technology*. Hershey, PA, USA: Hershey, 2019, pp. 781–809.

[49] S. Wu, "Short text mining for fault diagnosis of railway system based on multi-granularity topic model," in *Proc. 8th Int. Conf. Logistics, Informat. Service Sci. (LISS)*, Aug. 2018, pp. 1–6, doi: 10.1109/LISS.2018.8593228.

[50] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, "Bilevel feature extraction-based text mining for fault diagnosis of railway systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 49–58, Jan. 2017, doi: 10.1109/TITS.2016.2521866.

[51] C. Wang, X. Pan, and Y. Wang, "Social networks and railway passenger capacity: An empirical study based on text mining and deep learning," in *Proc. 4th ACM SIGSPATIAL Int. Workshop Saf. Resilience (EM-GIS)*, Nov. 2018, pp. 1–6, doi: 10.1145/3284103.3284125.

[52] L. Hua, W. Zheng, and S. Gao, "Extraction and analysis of risk factors from Chinese railway accident reports," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 869–874, doi: 10.1109/ITSC.2019.8917094.

[53] M. Heidarysafa, K. Kowsari, L. Barnes, and D. Brown, "Analysis of railway accidents' narratives using deep learning," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 1446–1453, doi: 10.1109/ICMLA.2018.00235.

[54] C. Lin and G. Wang, "Failure cause extraction of railway switches based on text mining," in *Proc. Int. Conf. Comput. Sci. Artif. Intell.*, Dec. 2017, pp. 237–241, doi: 10.1145/3168390.3168402.

[55] F. Wang, T.-H. Xu, Y. Zhao, and Y.-R. Huang, "Prior LDA and SVM based fault diagnosis of vehicle on-board equipment for high speed railway," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 818–823, doi: 10.1109/ITSC.2015.138.

[56] Y. Zhao, T.-H. Xu, and W. Hai-Feng, "Text mining based fault diagnosis of vehicle on-board equipment for high speed railway," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 900–905, doi: 10.1109/ITSC.2014.6957803.

[57] L. Li, W. Li, and D. Gong, "Naive Bayesian automatic classification of railway service complaint text based on eigenvalue extraction," *Tehnički Vjesnik*, vol. 26, no. 3, pp. 778–785, Jun. 2019, doi: 10.17559/TV-20190420161815.

[58] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007, doi: 10.1109/TKDE.2007.1048.

[59] C. LaBonne, B. Burke, and M. Whitman, "Role of MAP kinase in mesoderm induction and axial patterning during Xenopus development," *Development*, vol. 121, no. 5, pp. 1475–1486, 1995, doi: 10.1016/0168-9525(96)81381-3.

[60] R. Polig, K. Atasu, and C. Hagleitner, "Token-based dictionary pattern matching for text analytics," in *Proc. 23rd Int. Conf. Field Program. Log. Appl.*, Sep. 2013, pp. 1–6, doi: 10.1109/FPL.2013.6645535.

[61] T. Verma, R. Renu, and D. Gaur, "Tokenization and filtering process in RapidMiner," *Int. J. Appl. Inf. Syst.*, vol. 7, no. 2, pp. 16–18, Apr. 2014, doi: 10.5120/ijais14-451139.

[62] H. Saif, M. Fernandez, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 810–817.

[63] M. K. Dalal and M. A. Zaveri, "Automatic text classification: A technical review," *Int. J. Comput. Appl.*, vol. 28, no. 2, pp. 37–40, Aug. 2011, doi: 10.5120/3358-4633.

[64] K. Kowsari, J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.

[65] H. Alawad, S. Kaewunruen, and A. Min, "Utilizing big data for enhancing passenger safety in railway stations," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 603, no. 5, Sep. 2019, Art. no. 052031, doi: 10.1088/1757-899X/603/5/052031.

[66] S. Hidirov and H. Guler, "Reliability, availability and maintainability analyses for railway infrastructure management," *Struct. Infrastruct. Eng.*, vol. 15, no. 9, pp. 1221–1233, Sep. 2019.

[67] M. Catelani, L. Ciani, D. Galar, G. Guidi, S. Matucci, and G. Patrizi, "FMECA assessment for railway safety-critical systems investigating a new risk threshold method," *IEEE Access*, vol. 9, pp. 86243–86253, 2021.

[68] Z. Zhang, L. Jia, and Y. Qin, "RAMS analysis of railway network: Model development and a case study in China," *Smart Resilient Transp.*, vol. 3, no. 1, pp. 2–11, May 2021.

HAMAD ALAWAD (Member, IEEE) received the bachelor's degree in industrial engineering from King Saud University, the master's degree in fire and safety engineering from U.K., and the Ph.D. degree from the Birmingham Centre for Railway Research and Education, University of Birmingham, U.K. His research interests include safety, machine learning, and artificial intelligence.

SAKDIRAT KAEWUNRUEN received the Ph.D. degree in structural engineering from the University of Wollongong, Australia. He has expertise in transport infrastructure engineering and management, successfully dealing with all stages of infrastructure life cycle and assuring safety, reliability, resilience and sustainability of rail infrastructure systems. He is a chartered engineer and has over 500 technical publications. He is a member of BSI, CEN and ISO standard committees for railway applications.