International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

UNSUPERVISED MACHINE LEARNING FOR MANAGING SAFETY ACCIDENTS IN RAILWAY STATIONS

[1] Mr. Vijay Kumar L, [2] S. Sravani,[3] V. Spurthi,[4] V. Snehitha
[1]Assistant Professor,[234]Students
Department Of CSE
Malla Reddy Engineering College for Women

ABSTRACT

Railroad operations must be reliable, easily accessible, well-maintained, and safe (RAMS) in order to move both passengers and freight. An important everyday safety issue in many metropolitan areas is the danger of accidents at train stations. In addition to casualties, public worry, and financial losses, accidents wreak havoc on market reputation. Stations like this are feeling the heat from increased demand, which is putting a strain on infrastructure and making safety a top administrative priority. The use of unsupervised topic modeling to better understand the contributors to these severe incidents is advised for the purpose of analyzing them and using technology, such as AI approaches, to increase safety. This study aims to optimize Latent Dirichlet Allocation (LDA) using textual data collected from RSSB, which includes 1000 accidents that occurred at UK railway stations, with the purpose of reducing mortality accidents at these locations. To improve station safety and risk management, this study details the use of a machine learning topic approach to systematic detect accident characteristics, and it offers advanced analysis. The research assesses the effectiveness of text mining from accident records in obtaining information, lessons learned, and a comprehensive understanding of the risks associated with evaluating accidents that result in deaths on a large and long-term scale. Predictive accuracy for important accident data, including hotspots at railway stations and underlying causes, is shown by this Intelligent Text Analysis. In addition, the advancement of big data analytics allows for a better understanding of the nature of accidents compared to restricted domain analysis of accident reports, which was previously impossible without a large quantity of safety history. A new age of useful and widespread artificial intelligence applications in railway sector safety and other domains for safety applications has dawned, made possible by this technology's high level of accuracy.

1. INTRODUCTION

Many people believe that trains are the safest form of public transportation. Station operations, design, and customer habits are just a few of the numerous overlapping elements that put people aboard trains in danger. Operating the stations is not without its hazards, what with the ever-increasing demand, the already-congested society, and

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

the current architecture and complexity of certain stations. One of the most important aspects of the railway system is the safety of passengers, as well as other individuals and the general public. In 1999, the European Union implemented EN 50126, a standard that stands for Reliability, Availability, Maintainability and Safety (RAMS). The goal is to make railway operations very safe and to stop accidents from happening. Reducing risks to manageable levels and increasing safety are the outcomes of the RAMS analysis ideas. Nevertheless, that has been a pressing concern, and tragically, there are still reports of several deaths and injuries occurring year at the train station. Just to illustrate, In 2016, 202 people lost their lives in 420 incidents in Japan that included being hit by a train. Among the 420 incidents, 179 (or 24 deaths) included people falling from platforms or being injured or killed after being struck by a train [1]. Most passenger injuries in the UK in 2019 and 2020 were reportedly caused by incidents that occurred at stations. Most Excellent The important effect on minimizing injuries on station platforms and providing a quality, dependable, and safe travel environment for all passengers, workers, and the public is borne out by the fact that over 200 major injuries were the result of slips, trips, and falls [2]. Delays, costs, public fear and worry, interruptions to operations, and harm to the industry's image may occur as a consequence of accidents even if no one is hurt or killed. In addition, before implementing or funding any station safety measures, it is essential to assess the potential dangers posed by railway incidents and the station itself, as well as to identify the various factors that contribute to accidents through an exhaustive understanding of their causes and taking into account all available technologies.

As a result, we're interested in learning more about the topic modeling methodologies used to security and accident topics in the stations. With the hope of making a future contribution to smart station risk management and safety, this study presents an LDA-based topic modeling technique in conjunction with other models for advanced analytics. We study railway safety incidents including fatalities by using the models.

2. LITERATURE SURVEY

To efficiently evaluate accident data, Chen et al. provide a reliable method that uses clustering algorithms. By exploring the use of several unsupervised machine learning algorithms for accident prediction and prevention, Kumar and Kaur demonstrate encouraging outcomes in practical settings. The work of Liang et al. focuses on risk assessment and anomaly identification; specifically, they provide a framework for detecting safety-related abnormalities by combining clustering and outlier detection approaches.

In their case study, Wang and Zhang shed light on how clustering algorithms might be

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

used in the real world to analyze safety incidents. While Park et al. suggest a combined strategy integrating unsupervised learning with Internet of Things devices for safety monitoring, Gupta and Sharma investigate ways to proactively enhance railway station safety.

Anomaly identification in surveillance films is the focus of Yang et al., whereas data-driven safety management utilizing unsupervised learning approaches is the area of expertise of Zhou and Liu. All of these research show that unsupervised machine learning may help make train stops safer and better at handling accidents.

3. EXISTING SYSTEM

However, the railway industry has not made use of any of these methods, and there is a lack of consistency in the terminology used in the literature. In addition, a railway signaling company has used NLP to find mistakes in their specifications papers [13].Additionally, for the purpose of converting railway industry technical standards into contract terms [14]. In order to further understand the causes of railway accidents, a taxonomy framework was suggested that makes use of Self-Organizing Maps (SOM) to categorize the many human, technological, and organizational elements involved [15].In a similar vein, association rules mining has been used to discover possible elements that cause railway accidents [16].

Several ML methods, including support vector machines (SVMs), artificial neural networks (ANNs), extreme learning machines (ELMs), and decision trees (DTs), have been used in the arena of machine learning and risk, occupational safety, and safety accidents [7], [17]. Researchers have studied topic modeling in numerous domains, where it has been shown to be one of the most effective methods for data mining [18]. This method has found applications in software engineering [19], medical and health [21], linguistic science [25], and other areas. Occupational accidents [17], construction [8], [27], [28] and aviation [29], [30], [31] are only a few sectors where this method has been used for predictions in the literature. Cybersecurity and Data Science [36], Understanding Occupational Construction Incidents in the Construction Industry and Predicting Injuries in the Construction Industry [32], Industrial Accidents in Steel Factories [35], and Factors Correlating to Occupational Falls in the Construction Industry [34] have all been studied. Also, for the purpose of safety risk analysis, data, correlations, and variables have been extracted from 156 reports of construction safety incidents involving urban rail transit in China [37]. Literature reviews have shown that, first, there is no silver bullet when it comes to text categorization problems, and second, information extraction from text is an incremental process [38], [11]. There has

been significant success in the railway industry using a semi-automated approach to identifying close call data based on unstructured language. Furthermore, similar technology may soon be required for railway safety management, according to reports [11].

The use of text-analyzing technologies in railway safety is anticipated to address concerns including analysis that is both time-consuming and incomplete. The automated process, high productivity with quality, and efficient system for monitoring safety in the railway system are other benefits that have been shown. In addition, machine learning technologies have been used to avoid railway accidents. Data mining makes use of a wide variety of techniques, such as automated learning, information extraction, natural language processing, and information retrieval. For example, a text mining method (classification) that relies on machine learning has been used to differentiate between secondary crashes according to crash narratives. This approach seems to work well and shows promise for improving secondary crash identification [39]. By assisting decision-makers and allowing them to study accident causes, relevant variables, and correlations, such tools are strong for railway safety [40].There are a number of promising new directions for research and development in text mining with applications in railway safety engineering [41].

In order to understand what elements contribute to train accidents, text mining using probabilistic modeling and k-means clustering is useful. The study has been identifying the factors of lane defects, wheel defects, level crossing accidents, and switching accidents as potential causes of a high number of recurring accidents [42] based on its application analysis to reports about major railroad accidents in the US and Canada. The features of rail accidents may be better understood and safety engineers can be better equipped via the use of text mining, which also provides a wealth of additional detailed information.

More study is required to fully understand the causes and features of these incidents, however a combination of text analysis and ensemble approaches has been used to evaluate U.S. accident reports data spanning 11 years [41]. In the United States, for example, comparative text mining techniques like Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are used to reports of railroad equipment incidents in order to extract themes [43]. Data mining techniques including an ordered probit model, association rules, and classification and regression tree (CART) algorithms have also been used to determine the primary variables linked to the severity of injuries.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

With regard to deep learning, Data The purpose of this study was to identify patterns in the causes and descriptions of railroad accidents in the United States from 2001 to 2016. Consequently, deep learning has been used to automatically comprehend domain-specific texts and evaluate narratives of railway accidents. This has allowed for the correct classification of accident causes, the identification of significant discrepancies in accident reporting, and the provision of benefits to safety engineers [53].Additionally, text mining was used to detect and anticipate switch failures [54]. The prior LDA model was presented for fault feature extraction in high-speed railway fault diagnostics of vehicle onboard equipment [55], and the Bayesian network (BN) is also employed for fault feature extraction [56].

With the use of a Naive Bayesian classifier and the term frequency-inverse document frequency approach, passenger complaint material may be automatically classified and its eigenvalues extracted [57].

Disadvantages

- The system has never employed machine learning (ML) methods, despite their greater accuracy and efficiency, such decision trees (DT), SVM, ANN, and extreme learning machines (ELM).
- The Self-Organizing Maps (SOM) model was not used by the system to categorize organizational, technological, and human aspects in railway accidents.

4. PROPOSED SYSTEM

This article presents a novel approach to the study of how the textual source of data from accident reports from railway stations might be effectively utilized to identify the underlying causes of accidents and create a relationship between the textual and potential reasons. where a fully automated mechanism that can receive text input and produce not-yet-ready outputs is located. By using this approach, problems like helping the decision-maker in real time and extracting the important information that non-experts can understand, better identifying the accident's details in detail, expertly designing a smart safety system, and making efficient use of safety history records should be resolved. An These findings may help to encourage more methodical and intelligent study of safety and risk management. Modern LDA algorithms are used in our method to extract important textual data about accidents and their causes.

Advantages

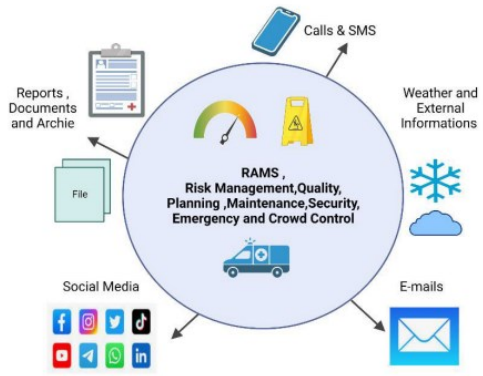
- A decision support tool known as a DT utilizes a tree-like structure to apply choices and their anticipated results [40], [53]. Safety analysis may be approached from a variety of (ML) angles. To be more precise, we teach a DT to identify the types of incidents that happen in the stations

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

as well as the patterns that emerge from them.

- Strong, useful information may be found in the textual data, including the victim's age range, location, description of the accident, and time of the incident. To get more precise timings for mining, the time of incidents has been separated into halves of the day.

5. ARCHITECTURE



6. ALGORITHIM

Gradient boosting

Machine learning techniques like gradient boosting are used to applications like regression and classification. An ensemble of weak prediction models—usually decision trees—is what it provides as a prediction model.[1][2] Gradient-boosted trees is the name of the resultant method when a decision tree serves as the weak learner; it often performs better than random forest.Similar to other boosting techniques, a gradient-boosted trees model is constructed step-by-step; however, it differs from them in that it permits optimization of any differentiable loss function.

K-Nearest Neighbors (KNN)

- Using a similarity metric as a basis for categorization, this simple but very effective technique
- Non-parametric Lazy learning Doesn't start learning until the test case is presented
- Whenever we encounter fresh data to be classified, we use the training data to identify its K-nearest neighbors.

Logistic regression Classifiers

The relationship between a collection of independent (explanatory) factors and a categorical dependent variable is examined using logistic regression analysis. When the dependent variable simply has two values, such as 0 and 1, or Yes and No, the term logistic regression is employed. When the dependent variable, such as married, single, divorced, or widowed, has three or more distinct values, the term multinomial logistic regression is typically reserved for that situation. While the dependent variable's data type differs from multiple regression's, the procedure's practical application is comparable.

When it comes to examining categorical answer variables, discriminant analysis and logistic regression are rivals. Many

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

statisticians believe that discriminant analysis is less flexible and less appropriate for modeling most scenarios than logistic regression. This is so because, unlike discriminant analysis, logistic regression does not make the assumption that the independent variables are regularly distributed.

On both categorical and numeric independent variables, this software computes binary and multinomial logistic regression. Together with the regression equation, quality of fit, odds ratios, confidence intervals, probability, and deviance are also reported. Complete residual analysis is carried out, including with diagnostic residual reports and charts. It has the ability to search for the optimal regression model with the fewest independent variables using an independent variable subset selection search. It offers ROC curves and confidence intervals on expected values to assist in figuring out the ideal cutoff point for categorization. It categorizes rows that are not utilized in the analysis automatically, allowing you to verify your findings.

Naïve Bayes

The naive bayes approach is a supervised learning technique that operates on the basic premise that the existence (or lack) of a certain feature inside a class is independent of the existence (or lack) of any other feature.

But in spite of this, it seems strong and effective. It performs similarly to other methods of guided learning. Numerous explanations have been put forward in the literature. We emphasize a representation bias-based explanation in this lesson. Along with linear discriminant analysis, logistic regression, and linear SVM (support vector machine), the naive bayes classifier is a linear classifier. The technique used to estimate the classifier's parameters—the learning bias—is where the differences reside.

Although the Naive Bayes classifier is commonly used in research, practitioners who want findings that are practical are less likely to use it. On the one hand, the researchers discovered that it is particularly simple to develop and use, that estimating its parameters is a simple task, that learning occurs quickly even on extremely big datasets, and that, when compared to other systems, its accuracy is rather excellent. However, the end users do not get a model that is simple to use and comprehend, nor do they see the benefit of this method.

As a consequence, we portray the learning process' outcomes in a fresh way. Both the implementation and the understanding of the classifier are simplified. This tutorial's initial section covers some of the naive bayes classifier's theoretical underpinnings. Next, we apply the method to a Tanagra dataset. We contrast the outcomes (the model's

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

parameters) with those from other linear techniques, including logistic regression, linear discriminant analysis, and linear SVM. We see a great degree of consistency in the outcomes. This significantly explains why the approach performs well when compared to other methods. In the second section, Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b, and RapidMiner 4.6.0 are among the tools we employ on the same dataset. Above all, we strive to comprehend the outcomes that are accomplished.

Random Forest

Random forests, also known as random decision forests, are an ensemble learning technique that builds a large number of decision trees during the training phase for problems including regression, classification, and other applications. The class that the majority of the trees choose is the random forest's output for classification problems. The mean or average prediction made by each individual tree is provided for regression tasks. The tendency of decision trees to overfit to their training set is compensated for by random decision forests. Although they are less accurate than gradient enhanced trees, random forests still perform better than choice trees in most cases. Their performance, however, may be impacted by the peculiarities of the data.

Tin Kam Ho[1] developed the first random decision forest algorithm in 1995 by using the random subspace technique, which is a means of putting Eugene Kleinberg's "stochastic discrimination" approach to classification into practice.

Leo Breiman and Adele Cutler created an expansion of the algorithm and filed for a trademark for "Random Forests" in 2006; as of 2019, Minitab, Inc. is the owner of this trademark.In order to create a set of decision trees with controlled variance, the extension combines Breiman's "bagging" concept with random feature selection, which was initially presented by Ho[1] and then separately by Amit and Geman[13].

Because they need minimal setup and provide excellent predictions over a broad variety of data, random forests are often employed as "blackbox" models in organizations.

SVM

A discriminant machine learning approach is used in classification problems to discover a discriminant function that can accurately predict labels for newly acquired instances, based on an independent and identically distributed (iid) training dataset. In contrast to generative machine learning techniques, which need calculating conditional probability distributions, a discriminant classification function allocates a given data point (x) to one of the several classes involved in the classification exercise. Discriminant techniques, particularly for a multidimensional feature space and when just posterior probabilities are required, require less computer resources and training data than generative approaches, which are often used when prediction incorporates

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

outlier identification. Finding the equation for a multidimensional surface that optimally divides the various classes in the feature space is the geometric equivalent of learning a classifier.

SVM is a discriminant approach that, unlike genetic algorithms (GAs) or perceptrons, both of which are often used for classification in machine learning, always returns the same optimum hyperplane value since it solves the convex optimization issue analytically. The termination and initialization criteria have a significant impact on the solutions for perceptrons. Training yields precisely specified SVM model parameters for a given training set for a certain kernel that converts the data from the input space to the feature space; in contrast, the perceptron and GA classifier models vary with each training set. Only minimizing error during training is the goal of GAs and perceptrons, which translates into several hyperplanes fulfilling this criteria.

7. IMPLEMENTATION

Modules

Service Provider

The Service Provider must provide a valid user name and password to log in to this module. He can do some tasks after logging in successfully, such train and test railway data sets, View the Accuracy of Trained and Tested Railway Data Sets in a Bar Chart. View the accuracy results of the trained and tested railway data sets, the railway accident type prediction, the railway accident type ratio, Acquire Forecasted Data Sets, View All Remote Users and the Railway Accident Type Ratio Results.

View and Authorize Users

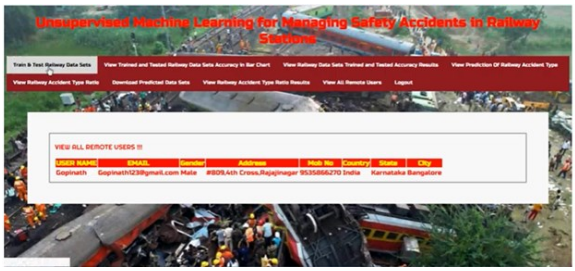
The administrator may see a list of all enrolled users in this module. The administrator may see user information here, including name, email address, and address, and they can also approve people.

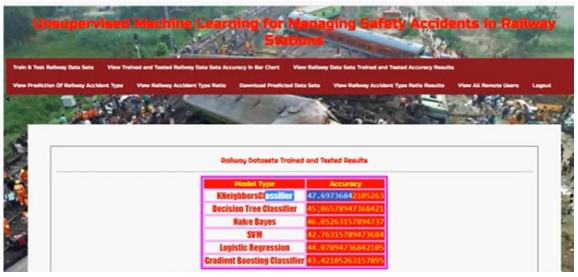
Remote User

There are n numbers of users present in this module. Prior to beginning any actions, the user must register. The user's information is saved in the database when they register. Upon successful registration, he must use his permitted user name and password to log in. Upon successful login, the user may do several tasks such as registering and logging in, predicting the kind of railroad accident, and seeing their profile.

8. SCREEN SHOTS

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86





9. CONCLUSION AND FUTURE ENHANCEMENT

At several domains, including text mining for safety and risk management at railroad stations, topic models play a significant role. A set of words that appear in statistically significant ways is referred to as a topic in topic modeling. Texts may be used to evaluate risk papers, record investigative findings, and more.

This study presents many examples of how unsupervised machine learning topic modeling may support industry-based risk management, safety accident investigations, and accident recording and documentation restructuring. The recommended model and the explanation of the accident's underlying causes have shown that the platforms in the stations are the hotspots. The findings indicate that there are four primary reasons for the incidents at the station: falls, being hit by trains, and electric shock. Furthermore, it seems that there are greater dangers throughout the night and on certain days of the week.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Enhanced safety in text mining allows for the acquisition of information across several domains and time periods, leading to better RAMS efficiency and the development of a comprehensive viewpoint for all parties involved.

The use of unsupervised machine learning is beneficial for safety since it can solve problems, uncover hidden patterns, and address a variety of issues like:

• Textual information in unstructured formats and from a variety of angles

Capture the relationships, causations, more for ranking risks and related information; Smart labeling, clustering, centroids, sampling, and associated coordinates; Power for discovery, handling missing values, and spotting safety and risk kyes from data; Prioritizing risks and measures implementations

• Support the safety review process and draw lessons from the extensive and protracted experience.

• Scale and weighted configuration options are available for use in risk assessment.

While this paper demonstrates the innovative use of unsupervised machine learning in railway accident classification and root cause analyses, more research on large-scale data topics pertaining to station diversity, size, safety cultures, and other factors is imperative, with the goal of developing advanced techniques for unsupervised machine learning algorithms in the future. In conclusion, this study improves safety, but it also highlights the significance of textual data and proposes a more thorough redesign of the data collection process.

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.