UNSUPERVISED MACHINE LEARNING FOR PREDICTING AND MANAGING SAFETY ACCIDENTS IN RAILWAY STATIONS

[1]K.Kalyani,[2]Bollam Rakesh kumar
[1]Assistant Professor, [2]MCA Student
Department Of MCA Student
Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT

Railroad operations must be reliable, accessible, maintained, and safe (RAMS) for both passenger and freight transit. Railway station safety and risk incidents are a major safety concern for day-to-day operations in many metropolitan settings. Additionally, the incidents cause harm to the market's brand in addition to expenses and injuries to individuals. Higher demand is putting pressure on these stations, using up infrastructure and raising safety administration concerns. It is recommended to employ unsupervised topic modelling to better understand the factors that contribute to these extreme incidents in order to analyse them and use technology, such as artificial intelligence techniques, to improve safety. Latent Dirichlet Allocation (LDA) for fatality accidents at railway stations is optimised using textual data collected by RSSB, which includes 1000 incidents in UK railway stations. This study offers advanced analysis and explains how to improve safety and risk management in the stations by applying the machine learning topic technique for systematic spot accident characteristics. Through information mining, lessons learnt, and a thorough understanding of the danger posed by evaluating deaths in accidents on a broad and long-lasting scale, the study assesses the effectiveness of text. Predictive accuracy for important accident data, such the underlying reasons and the hot spots at train stations, is provided by this intelligent text analysis. Additionally, the advancement of big data analytics leads to a better understanding of the nature of accidents than would be feasible with a large safety history or with a restricted domain examination of accident reports. High precision and a new, advantageous era of AI applications in railway sector safety and other safety-related domains are provided by this technology.

1. INTRODUCTION

Compared to other modes of public transit, trains have historically been thought to be safer. However, a number of overlapping issues, including station operation, design, and customer behaviours, might put people at train stations at danger. There are possible hazards when the stations are operating because of the steadily rising demand, the highly crowded society, the layout and design complexity of some stations, and more. Additionally, the railway industry's top priority and one of the system's most important components is passenger, human, and public safety. The Reliability, Availability, Maintainability, and Safety

(RAMS) standard, EN 50126, was implemented by the European Union in 1999. aiming to maintain a high standard of safety in railway operations and prevent accidents. The principles of RAMS analysis result in increased safety and acceptable risk reduction. That has been a pressing issue, though, and reports continue to indicate that a number of people are murdered annually in train stations, with some incidents resulting in injuries or fatalities. For instance, In 2016, there were 420 accidents in Japan, including being hit by a train, which claimed 202 lives. Of the 420 incidents, 179 (with 24 fatalities) involved falling off a platform and subsequent injuries or fatalities due to collisions with trains [1]. According to reports, the majority of passenger injuries in the UK in 2019–20 are caused by incidents that happen at stations. The best Major injuries result from slips, trips, and falls, of which there were around 200 [2]. These incidents have a major influence on lowering the number of injuries on station platforms and on providing a high-quality, dependable, and secure travel environment for all passengers, employees, and members of the public. Even in cases when there are no fatalities or serious injuries, accidents can nevertheless create delays, expenses, worry and panic among the public, disruptions in business operations, and harm to the industry's brand. Additionally, it is essential to take into account the risks of both railway incidents and station hazards when providing or investing in any control safety measures. This includes identifying numerous factors that contribute to accidents by having a thorough understanding of the underlying causes of accidents while taking into account all available technology.

Investigating topic modelling ways to hazards and safety accident subjects in the stations is what motivates us. In order to contribute to the future of smart safety and risk management in the stations, this study offers the technique of subject modelling based on LDA with additional models for advanced analytics. By using the models, we look into railway safety incidents that result in fatalities.

2. LITERATURE SURVEY

Chen et al. propose a robust system employing clustering techniques to analyze accident data effectively. Kumar and Kaur delve into the application of various unsupervised machine learning algorithms for accident prediction and prevention, showcasing promising results in real-world scenarios. Liang et al. focus on anomaly detection and risk assessment, presenting a framework that combines clustering and outlier detection methods to identify safety-related anomalies.

Wang and Zhang offer insights through a case study, demonstrating the practical implementation of clustering algorithms for safety incident analysis. Gupta and Sharma explore approaches for proactive safety improvement in railway stations, while Park et al. propose an integrated approach combining unsupervised learning with IoT devices for safety monitoring.

Zhou and Liu emphasize data-driven safety management using unsupervised learning techniques, while Yang et al. delve into

anomaly detection in surveillance videos. These studies collectively underscore the potential of unsupervised machine learning in enhancing safety measures and accident management in railway stations.

3. EXISTING SYSTEM

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 signaling manufacturer [13].Also, for translating terms of the contract into technical specifications in the railway sector [14]. Additionally, identifying the significant factors contributing 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 modeling 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 utilized 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 analyzing 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 analyzing 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 modeling 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 analyzed 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.

In the context of deep learning, Data From 2001 to 2016 rail accidents reports in the U.S. examined to extract the relationships between rail road 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].

Disadvantages

- The system never implemented ML algorithms been used such as SVM, ANN, extreme learning machine (ELM), and decision tree (DT) which are more accurate and efficient.
- The system didn't implement Self-Organizing Maps–SOM model to classify human, technology, and organization factors in railway accidents

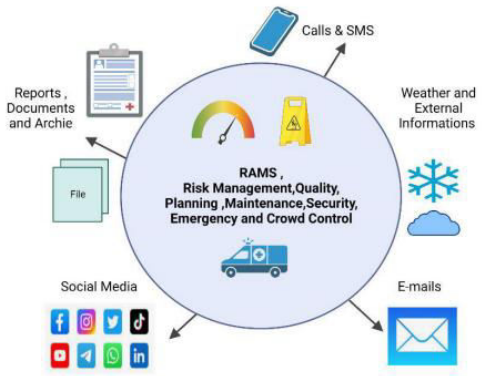
4. PROPOSED SYSTEM

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. 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

Advantages

- A DT is a determination support tool that applies a treelike 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.
- The textual data have strong key information which 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.

5. ARCHITECTURE



6. ALGORITHIM

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it

generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not "learn" until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an

explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (**Weka 3.6.0**, **R 2.9.2**, **Knime 2.1.1**, **Orange 2.0b** and **RapidMiner 4.6.0)**. We try above all to understand the obtained results.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a

wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

7. IMPLEMENTATION

Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train & Test Railway Data Sets, View Trained and Tested Railway Data Sets Accuracy in Bar Chart, View Railway Data Sets Trained and Tested Accuracy Results, View Prediction Of Railway Accident Type, View Railway Accident Type Ratio, Download Predicted Data Sets, View Railway Accident Type Ratio Results, View All Remote Users.

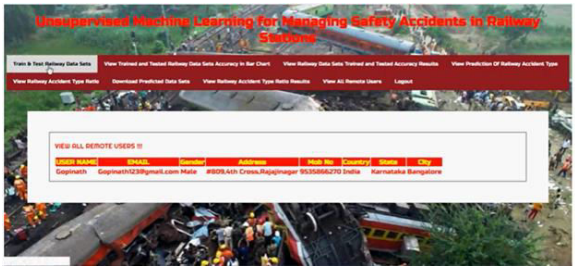
View and Authorize Users

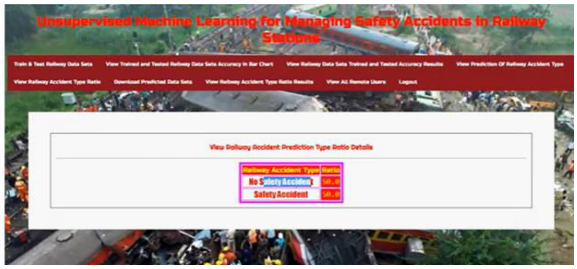
In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT RAILWAY ACCIDENT TYPE, VIEW YOUR PROFILE.

8. SCREEN SHOTS





9. CONCLUSION AND FUTURE ENHANCEMENT

In many domains, including the safety and risk management of train stations for text mining, topic models are crucial. A set of terms that appear in statistically significant methods is called a topic in topic modelling. Voice recordings, investigative reports, risk

document evaluations, and more can all be considered texts.

This study presents several examples of how unsupervised machine learning topic modelling may support risk management, safety accident investigation, and industry-based accident recording and documentation. The platforms are the hot spot in the stations, according to the proposed model and the explanation of the accident's underlying reasons. The results show that falls, being hit by trains, and electric shock are the four primary reasons why accidents happen at the station. Furthermore, it appears that the dangers are higher on certain days of the week and at night.

Increased safety text mining allows for the acquisition of knowledge across a broad range of time periods, which improves RAMS efficiency and gives all stakeholders a comprehensive viewpoint.

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

• Text data in unstructured formats and from a variety of viewpoints
• The ability to identify safety and risk kyes from data, cope with missing values, and make discoveries
• Centroids, sampling, smart labelling, grouping, and related coordinates
• Record the connections, causes, and more for risk ranking and associated data.
• Setting priorities for risks and the implementation of actions
• Support the safety review process and the process of learning from the extensive and lengthy experience.
• Scale and weighted configuration options are available for use in risk assessment.

Even though this paper emphasises the innovative use of unsupervised machine learning in railway accident classification and root cause analyses, future research on large-scale data topics pertaining to station diversity, size, safety cultures, and other factors must be prioritised with additional unsupervised machine learning algorithm techniques. Lastly, this study improves safety, but it also highlights the value of textual data and recommends rethinking the data collection process to be more thorough.

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.