Pegem Journal of Education and Instruction, Vol. 13, No. 4 , 2023 (pp. 800-811)

RESEARCH ARTICLE WWW.PEGEGOG.NET

MACHINE LEARNING WITHOUT HUMAN SUPERVISION FOR HANDLING SAFETY INCIDENTS IN TRAIN STATIONS

M.Kavitha[1], Badige Bhagyasree [2], Bandari Veneela [3], Paluchuri Siri [4]

[1] Associate Professor, Department of Information Technology, Sridevi Women's Engineering College, Hyderabad

sweckavitha2414@gmail.com

[2, 3, 4] Department of Information Technology, Sridevi Women's Engineering College, Hyderabad

Abstract: Railroad operations must be reliable, easily accessible, well-maintained, and safe (RAMS) in order to move both passengers and freight. The potential for accidents at railway stations is a major issue with everyday operations in many cities. Damage to the market reputation, injuries, public fear, and expenses are further outcomes of accidents. 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. Using textual data collected from RSSB, which includes one thousand accidents at UK railway stations, this study aims to optimize Latent Dirichlet Allocation (LDA) for fatality accidents in railway stations. Improving station safety and risk management via advanced analysis is the goal of this study, which details the use of a machine learning topic technique for systematic spot accident characteristics. This research looks at the effectiveness of text mining in evaluating large-scale incidents that resulted in deaths in order to get information, lessons learned, and a thorough understanding of the risks involved. Predictive accuracy for important accident data, including hotspots at railway stations and underlying causes, is shown by this Intelligent Text Analysis. Additionally, unlike with limited domain examination of accident reports, the enhanced big data analytics allow us to comprehend the nature of the incidents in ways that would have been impossible with a large quantity of safety history. This technology ushers in a new age of artificial intelligence (AI) safety applications in the railway sector and beyond, with pinpoint precision.

Keywords: Machine learning, Unsupervised learning, Principal Component Analysis, K-Means, Unlabeled data, Labeled data, Text mining.

How to cite this article: M.Kavitha1, Badige Bhagyasree 2, Bandari Veneela 3, Paluchuri Siri 4. MACHINE LEARNING WITHOUT HUMAN SUPERVISION FOR HANDLING SAFETY INCIDENTS IN TRAIN STATIONS.Pegem Journal of Education and Instruction, Vol. 13, No. 4, 2023, 800-811

Source of support: Nil **Conflicts of Interest:** None.

DOI: 10.48047/pegegog.13.04.85

Received: 12.10.2023

1. Introduction

safest form of public transportation. Nonetheless, there are a lot of interrelated issues, including station operations, architecture, and customer

Many people believe that trains are the

Accepted: 22.11.2023 **Published:** 24.12.2023

behaviors, that might put people aboard trains in danger. There are certain hazards during station operations due to the ever-increasing demand, the very crowded society, the current layout of some stations, and the complexity of their designs. 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. Reduce hazards to manageable levels and increase safety with the use of RAMS analysis principles. But that has been a pressing concern, and statistics reveal that a number of people die each year at train stations, with some incidents resulting in serious injuries or even death. One case in point is when 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]. Station accidents are the leading cause of passenger injuries in the UK in 2019 and 2020. Most Excellent there were almost 200 major injuries due to slips, trips, and falls

[2]. Reducing injuries on station platforms and providing a quality, dependable and safe travel environment for all passengers, workers, and the public is of utmost importance. Accidents disrupt operations, harm the industry's brand, incur time and expense, and even if no one is hurt or killed, they nonetheless provoke dread and concern among the public. 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. The goal of this project is to analyze accident data collected from 01/01/2000 to 17/04/2020 in order to provide a smart technique that may improve future safety levels, risk management, and data gathering at railway stations. The participants in the study gave their consent for RSSBS to utilize the data they provided. It is not an easy task to analyze large amounts of data that are stored in a different format. Finding a particular piece of information in today's digital big data, which includes material from the web, videos, and photographs, might seem like searching for a needle in a haystack. This massive data set needs a robust tool to help organize, search, and comprehend. References [3], [4]. In order to extract useful information

801

from the massive amounts of safety data stored in the stations, including textual data, a number of pre-processing methods and approaches are needed. This study delves into the topic of topic modeling in order to discover valuable traits, like accident causes, and factors, which are collections of related words or phrases that summarize and explain the information in accident reports in a concise and accurate manner, saving time. In natural language processing, topic modeling approaches are widely used for topic recognition and semantic mining in unstructured

materials. These algorithms are resilient and clever. As a result, this study proposes the LDA model, a well-known probabilistic unsupervised learning approach for identifying the underlying themes in a set of settings [5]. A smart analysis using topic modeling techniques can be very useful and effective to semantic mining and latent discovery context documents and datasets. This is because there has been an increase in the application of new technologies, a data revolution, technological development, and the utilization ofAI in many fields. We are aiming for unstructured textual data since other types of data (images, videos, and numbers) have already been processed using AI methods that include supervised learning [6, 7]. As a result, we're interested in learning more about the topic modeling methodologies used to security and accident topics in the stations. This study aims to contribute to the future of smart safety and risk management in stations by providing a way of subject modeling based on LDA with additional models for advanced analytics. We study railway safety incidents including fatalities by using the models. This work introduces a novel approach to the field by studying the potential for extracting the underlying causes of accidents from railway station accident reports and doing a comparison between the textual and likely reasons. When a fully automated procedure that can take text as input and produce output is not yet available [8]. Successful implementation of this approach is anticipated to address concerns such as providing real-time assistance to decision-makers and making key information easily understandable for non-experts, improving the ability to thoroughly identify accident details, designing expert smart safety systems, and making effective use of safety history records. A more methodical and intelligent examination of risk management and safety might be aided by these findings. We capture the crucial text information of accidents and their causes using a state-of-the-art LDA method.

2. Literature Review

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 its specifications documentation. Additionally, for the purpose of converting railway industry technical standards into contract terms. The suggested taxonomy framework uses Self-Organizing Maps (SOM) to categorize human, technological, and organizational components in railway accidents, which helps to identify the key contributors to these incidents. The same holds true for railway incidents; association rules mining has helped pinpoint possible causes and their interrelationships. Semantic value matrix (SVM), artificial neural network (ANN), extreme learning machine (ELM), and decision tree are just a few of the ML methods that have found utility in the domain of risk, safety accidents, and occupational safety (DT). Researchers have studied topic modeling extensively, and it has shown to be one of the most effective data mining approaches. It has found applications in software engineering, medicine, linguistics, and other professions. It is also clear from the literature that Several fields have made use of this method for forecasting, including aviation, construction, and occupational accidents. The approach has been used to understand occupational construction events in the construction industry, as well as to anticipate injuries in the construction sector. It has also been applied to assess the elements linked to occupational falls in steel factories, as well as to cyber security and data science. In addition, data, correlations, and variables pertaining to potential dangers have been culled from 156 reports of construction accidents involving urban rail transportation in China. Literature reviews have shown that, first, there is no silver bullet when it comes to text categorization problems, and second, data extraction from texts is an iterative process. There has

been significant success in the railway industry using a semi-automated approach to identifying close call data based on unstructured language. Additionally, it has been stated that such technology may be required for railway safety management in the future. We anticipate that problems like time-consuming analysis and inadequate analysis will be addressed by using text-analyzing approaches to railway safety. 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 techniques have been used to reduce the occurrence of train accidents. Data mining makes use of a wide variety of techniques, such as automated learning, information extraction, natural language processing, and information

retrieval. To better identify secondary crashes, for example, a machine learning-based text mining technique has been used to differentiate between them using crash narratives; this method shows promise for improving secondary crash detection. These approaches greatly enhance railway safety by providing decision-makers with the information they need to evaluate accident causes, important variables, and linkages between them. Evidence suggests that text mining might be useful for railway safety engineering in a number of future contexts. Train accident reasons may be better understood by text mining using probabilistic modeling and k-means clustering. 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 based on application analysis of 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. While additional study is required, a combination of text analysis and ensemble approaches applied to eleven years'

worth of US accident reports has helped shed light on the causes and traits of these incidents. Additionally, in the United States, themes may be discovered in reports of railroad equipment incidents by comparing text mining approaches such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). 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. Some aspects, including train speed, age, gender, and time, have been examined using the U.S. accidents highway railroad grade crossings database for the period 2007–2013. A new era of big data has dawned on the railway sector, bringing with it exciting new possibilities for data-driven safety analyses. As a result, methods like Natural Language Processing (NLP) analysis have been proposed for the proactive identification of high-risk situations [46]. Starting with a Big Data Use Case an important component of the railway safety oversight system is the newly-introduced oversight System. Text mining techniques applied to railway accident and fault study reports has recently taken place. An NLP framework for analyzing accident data has been described using investigation reports of railway accidents, and there is a chance to employ big data and natural language for processing in the context of analyzing railway safety. Additionally, in order to enhance the fault diagnosis performance in railway systems, it has been suggested to

use the LDA algorithm for the categorization of maintenance text [49]. The Chinese railway has shown promising results in predicting passenger capacity using text data from social networks in conjunction with text mining and deep learning. Natural language processing has also been used by the Chinese Railway to identify and analyze risk variables from accident reports. Regarding deep learning, Data This study looked at train accident reports from the United States from 2001

to 2016 in order to determine the correlations between the causes of railroad accidents and the descriptions of those incidents. Consequently, deep learning has been used to automatically comprehend domain-specific texts and evaluate narratives of railway accidents. This has greatly benefited safety engineers by allowing them to more precisely categorize accident causes, identify significant variations in accident reporting, and more. Additionally, text mining was used to detect and anticipate switch failures. The prior LDA model was applied for fault feature extraction on high-speed trains, while Bayesian networks (BNs) are also used for fault feature extraction on onboard vehicle equipment. The term frequency-inverse document frequency approach has been used in conjunction with the Naive Bayesian classifier for the purpose of automated text categorization of passenger complaints and eigenvalue extraction.

3. Existing System

Many people believe that trains are the safest form of public transportation. Station operations, design, and passenger behaviors are only a few of the numerous overlapping elements that put people aboard trains in danger. There are certain hazards during station operations due to the ever-increasing demand, the very crowded society, the current layout of some stations, and the complexity of their designs. 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. By using the principles of RAMS analyses, we may reduce risks to manageable levels and increase safety. But that has been a pressing concern, and statistics reveal that a number of people die each year at train stations, with some incidents resulting in serious injuries or even death. One case in point is when 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]. Station

accidents are the leading cause of passenger injuries in the UK in 2019 and 2020. The leading cause of serious injury on station platforms is slipping, tripping, and falling, which resulted in almost 200 injuries in 2016 [2]. Creating a safe, dependable, and high-quality travel environment for everyone is our top priority.

3.1 Drawbacks in Existing System

- Problems with Interpretability and Limited Labeling: Since unsupervised learning depends on discovering patterns in unlabeled data, it might be difficult to understand the outcomes. Validating the accuracy of the model's conclusions or understanding the underlying reasons of safety accidents may be tough without labeled data.

- Problems with Scalability: When working with huge datasets, unsupervised learning algorithms may encounter problems with scalability. This is particularly true in the context of railway station safety, where the amount of data might be enormous. To manage massive datasets effectively, algorithms for grouping and anomaly detection are required.

- Unsupervised learning techniques may not naturally comprehend the causal links between variables, but they are great at detecting patterns and correlations. It is essential to identify the underlying causes of safety events in order to put preventative measures in place that really work. Using unsupervised learning on confidential safety data brings up moral questions about personal information protection and confidentiality. It is of the utmost importance to safeguard sensitive information and ensure compliance with privacy requirements.

4. Proposed System

The goals of the safety management system and the kind of the data should inform the selection of the unsupervised learning algorithms to use. Think about using a mix of algorithms such as KMeans, DBSCAN, or Isolation Forest. Incorporate the unsupervised learning system with

preexisting procedures and systems for safety management. Make that the machine learning model and the humans running the show can communicate and share data without any hitches.

- Iterative Model Evaluation: Use suitable metrics to continuously assess the unsupervised learning model's performance. As data patterns change and new safety concerns arise, iterate on the model accordingly.

- Scalability: Make sure the system can handle increasing datasets and changing safety standards. Think about the possibility of connecting to other transit networks or opening up more train stations.

4.1 Algorithm

A well-known clustering technique, K-Means may group occurrences that are comparable together. Railway authorities may learn more about trends and commonalities in safety accidents by grouping them into clusters. This will help them implement more focused preventative measures.

• Principal Component Analysis (PCA): The most significant aspects in the data may be identified using PCA, a dimensionality reduction approach. It is simpler to examine and understand trends connected to safety accidents when the dimensionality is reduced. When it comes to finding outliers or unexpected patterns in data, one anomaly detection tool that has proven useful is Isolation Forest. It is effective for identifying infrequent safety issues because it isolates cases that deviate from the norm.

4.1.1 Advantages

One use of unsupervised learning is anomaly detection, which seeks for out-of-the-ordinary occurrences or patterns. For the purpose of finding safety occurrences that may not have been seen or classified before, this is vital.

- Less Reliance on Labeled Data: Unsupervised learning may be implemented without the need for pre-labeled data, in contrast to supervised learning that makes use of labeled instances for training. When getting labeled safety incident data is difficult or costly, this is a good alternative.

- Monitoring in Real-Time: Safety events may be monitored in real-time using unsupervised learning methods. These algorithms can detect out-of-the-ordinary trends in incoming data and respond or intervene swiftly since they analyze the data continually.

When obtaining labeled data is difficult or costly, unsupervised learning might be a costeffective alternative for analysis. Without the need for lengthy human annotation, enterprises may obtain important insights by exploiting unlabeled data.

4.2 Modules

Service Supplier: A valid username and password are required for the Service Provider to access this module. Once he logs in, he'll have access to features like Train & Test Railway Data Sets, Verify the Precision of Trained and Tested Railway Data Sets with a Bar Chart, See the Trained and Tested Accuracy Results for Railway Data Sets, the Type of Railway Accidents Predicted, and the Ratio of Railway Accidents by Type. Obtain Forecasted Data Collections; Get the results of the railway accident type ratio from Get Access to All Users From Afar. Monitor and Permit Users The admin can get a complete rundown of all registered users in this section. Admins may see user information including name, email, and address, and they can also approve users here.

Work from afar At least n people are active in this module. Prior to doing any actions, users are required to register. Data will be entered into the database after a user has registered. He will need to log in using the permitted username and password when registration is completed. Users will be able to do things like see their profiles, predict the kind of railroad accident, and register and log in after the login process is completed.

5. Conclusion

When it comes to managing risks and ensuring the safety of railway stations via text mining, topic models play a significant role, among other applications. A set of terms that appear in statistically significant techniques is called a topic in topic modeling. Voice recordings, investigative reports,

risk assessments, and other similar types of writing exist. Findings from this study provide credence to the idea that unsupervised machine learning topic modeling may improve industry-level risk management, safety accident investigation, and accident recording

808

and documentation. According to the proposed model and the explanation of the accident's causes, the platforms are the stations' most vulnerable areas. The results show that there are four primary causes of accidents at the station: falls, being hit by trains, electric shock, and other similar incidents. The dangers seem to be greatest at night and on certain days of the week. By enhancing the security of text mining, information may be gathered from many sources and over long periods of time, leading to more efficient RAMS and the development of a comprehensive view for everyone involved. Unsupervised machine learning is a powerful tool for safety applications because it can solve problems, uncover patterns, and handle a wide variety of obstacles, including: • Text data presented in an unstructured format and from several angles Smart labeling, clustering, centroids, sampling, and related coordinates; The ability to discover, handle missing values, and spot safety and risk kyes in data; The ability to capture relationships and causalities; The ability to rank risks and related information; The implementation of measures and risk prioritization .

- Assist with the safety evaluation and gathering lessons from the extensive and extensive experience.

- As configuration choices, they may be weighted and utilized to evaluate risks. This paper does a good job of showcasing the innovative use of unsupervised machine learning in railway accident classification and root cause analyses, but future work should concentrate

on expanding research on huge data topics related to factors like station diversity in terms of size, location, and safety culture, as well as other techniques involving unsupervised machine learning algorithms. In conclusion, this study improves security, but it also brings attention to the value of textual data and proposes a more thorough redesign of data collection methods.

6. References

[1] A. Ahadh et al. ”Text mining of accident reports using semi-supervised keyword extraction and topic modeling” Process Safety and Environmental Protection (2021)

809

[2] X. Zhou *et al.”*Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree” Reliab Eng Syst Saf (2020)

[3] 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.

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

[5] D. M. Blei, ‘‘Probabilistic topic models,’’ *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr.

2012, doi: 10.1145/2133806.2133826.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

810

[12] 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

811