International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Detecting and Managing Railway Station Safety Accidents Using Unsupervised Machine Learning

[1]CHANDANA R, [2]Dr. Geetha M

[1]Student, Department of MCA, BIET, Davanagere, India
[2]Associate Professor, Department of MCA, BIET, Davanagere, India

ABSTRACT

Ensuring the reliability, accessibility, maintainability, and safety (RAMS) of railway stations is vital for both passenger and freight operations. As urbanization and transit demands increase, railway stations face heightened safety risks and operational challenges. Accidents in these environments not only cause injuries and fatalities but also result in reputational damage and financial costs. This study explores the application of unsupervised machine learning— specifically Latent Dirichlet Allocation (LDA) topic modeling—to analyze unstructured textual accident reports sourced from the UK Rail Safety and Standards Board (RSSB), comprising data from 1,000 recorded station-related incidents. By extracting hidden patterns and identifying recurring themes in fatality-related accidents, this approach offers a data-driven method to uncover root causes and high-risk areas within stations. The analysis supports predictive insights that can enhance risk assessment and improve proactive safety management. Leveraging intelligent text mining techniques allows for a broader, more comprehensive understanding of safety issues than traditional case-by-case reviews. This work contributes to advancing AI-based solutions in transportation safety, offering scalable, accurate insights that support strategic planning and real-time decision-making in the railway sector.

Keywords : Railway safety, unsupervised learning, topic modeling, accident analysis, Latent Dirichlet Allocation

I.INTRODUCTION

Railway operations, encompassing both passenger and freight services, require high standards of reliability, accessibility, maintenance, and safety (RAMS). In urban environments, safety incidents at railway stations pose critical challenges that can disrupt daily functions. These incidents not only lead to physical harm and public concern but also negatively impact the reputation of railway services and incur financial costs. The increasing demand on railway infrastructure further intensifies these safety management challenges.

To address these issues, the application of advanced technologies such as artificial,intelligence offers promising solutions. This study focuses on using unsupervised topic modeling techniques to analyze accident reports, aiming to uncover the underlying factors contributing to serious

accidents in railway stations. By optimizing Latent Dirichlet Allocation (LDA) on a dataset of 1,000 UK railway station accidents collected by RSSB, this research seeks to enhance understanding of accident patterns and improve risk management strategies. The findings demonstrate how machine learning-driven text analysis can extract valuable insights from large-scale safety data, providing a new approach toaccident prevention and railway safety enhancement.

II. RELEATED WORK

1. This study presents an unsupervised machine learning approach to identify and manage safety accident patterns in railway stations. The authors employ clustering techniques such as K-means and

DBSCAN to discover hidden structures in safety incident data, gathered from Indian railway logs.

2. This paper presents a comparative study on the use of unsupervised algorithms for accident pattern discovery in railway stations. Using historical accident reports, the authors developed models using K- Means, Hierarchical Clustering, and Gaussian Mixture Models.

3. Bhargav and Sreenivasulu explore the integration of unsupervised machine learning into the railway safety management domain. The study leverages real-time data from surveillance systems, maintenance logs, and weather reports to generate clusters of high-risk situations.

4. This study explores the use of unsupervised clustering techniques to detect thermal anomalies in railway infrastructure using infrared imaging.

5. This research introduces a semi- supervised learning framework for detecting foreign objects on railway tracks using convolutional autoencoders. The model is trained to reconstruct clean railway images, and any deviation in the reconstruction error is flagged as a potential anomaly. The approach utilizes both labeled and unlabeled data, enhancing the model's learning from vast unannotated image sets.

6. Giri and Jha present a comprehensive framework combining deep learning, traditional machine learning models, and GPS-based monitoring to enhance railway accident prevention strategies. Their approach integrates historical accident datasets with real-time location data to build predictive models using Random Forest, LSTM, and Decision Tree classifiers.

7. This paper focuses on analyzing textual narratives from railway accident reports using deep learning techniques for natural language processing (NLP). The authors use a hierarchical deep learning model combining word embeddings (Word2Vec and GloVe) with LSTM and CNN layers to extract latent risk patterns from incident descriptions.

8. This news article reports on the rollout of an AI-powered rail safety system in Indian Railways aimed at improving maintenance and reducing accidents. The deployed system uses a combination of surveillance video analytics, predictive maintenance algorithms, and real-time sensors to monitor railway infrastructure health. AI models identify abnormalities such as track displacement, signaling malfunctions, and mechanical wear.

III. METHODOLOGY

This study adopts a comprehensive unsupervised machine learning pipeline to analyze unstructured accident reports from railway stations with the goal of identifying root causes, uncovering recurring safety themes, and enabling informed, data-driven risk management. The methodology involves six systematic stages: data acquisition, preprocessing, topic modeling, accident classification, visualization, and decision support. Each stage is designed to ensure scalable and automated handling of real-world safety data.

3.1 Data Collection

Accident reports were sourced from the UK Rail Safety and Standards Board (RSSB), which maintains a publicly available repository of incident narratives. A total of 1,000 accident reports related to railway stations were collected. These textual records include fields such as:

Date and time of incident Location (station

name and zone) Narrative descriptions of

accidents

Demographic information of affected individuals

Categorized injury severity

The collected data provides a rich, unstructured dataset suitable for unsupervised learning and topic modeling.

Data Preprocessing

To prepare the unstructured textual data for modeling, the following preprocessing steps were executed:

Tokenization: Each accident report was split into individual tokens (words).

Stopword Removal: Common but uninformative words (e.g., "the," "is," "at") were removed.

Stemming: Words were reduced to their base/root forms to unify variations.

Time Categorization: Accident timestamps were grouped into "morning," "afternoon," "evening," and "night" to extract time- based patterns.

Normalization: Text was converted to lowercase, punctuation was removed, and redundant whitespace was eliminated

This cleaning process reduced noise and improved the semantic integrity of the data for downstream topic modeling.

3.2 Topic Modeling Using LDA

LDA is a generative probabilistic model that assumes documents (accident reports) are mixtures of topics and that topics are distributions over words.

The number of topics was optimized using coherence scoring

Each topic was labeled manually based on its top keywords (e.g., "slip and fall," "platform crowding," "mechanical faults").

LDA outputs a document-topic matrix and topic-word matrix, which were used to interpret thematic groupings across the dataset.

This process provided meaningful insight into the types of accidents that frequently occur and the contextual factors associated with them.

3.3 Accident Classification Using Decision Trees

In addition to topic modeling, the study utilized a Decision Tree (DT) classifier to categorize accidents based on attributes such as:

Time of day

Type of hazard (from LDA) Victim

demographics Station locatio

The DT model was trained using a portion of the dataset annotated via unsupervised clustering and LDA outputs. The tree structure enhanced interpretability and supported pattern recognition across accident types.

3.4 Visualization and Reporting

A custom-built visualization module was developed using matplotlib and seaborn libraries to generate

Topic distributions across the dataset Time-series

trends of accident occurrences Station-specific

heatmaps for risk zones

Demographic distribution of affected individuals The system allows safety managers to interact with and interpret these visuals through a web-based interface for real-time decision support.

3.5 System Deployment and Decision Support

The final model was integrated into a user- friendly application with the following components:

Web front-end (HTML, CSS, JavaScript) for data input and result visualization

MySQL-based backend for structured data storage

Python-based machine learning engine with REST APIs to perform topic modeling and classification

This system allows safety engineers to upload new accident reports, automatically analyze them, and receive actionable summaries of root causes and high-risk patterns—thus enabling intelligent, data-informed safety planning.

IV. LITERATURE REVIEW

This topic explores how unsupervised machine learning can be applied to improve safety in railway stations by automatically identifying and analyzing accident-related patterns. Unsupervised learning is particularly valuable in this context because much of the data collected in railway environments—such as surveillance video, maintenance logs, and incident reports—is unstructured and lacks predefined labels.

Unlike supervised learning, which relies on labeled examples to train models, unsupervised learning can work with raw, unlabeled data. This makes it ideal for detecting hidden patterns, clustering similar incidents, and identifying anomalies that may signal potential safety risks.

One common technique in this area is clustering, which groups similar data points together.

V. EXISTING SYSTEM

While machine learning and NLP techniques have been widely applied in various fields, their use in the railway sector is still relatively scarce and inconsistent. Some studies have implemented NLP to identify defects within railway signaling requirement documents, and others have used it for translating contractual language into technical specifications relevant to railway operations.Similarly, association rule mining has been applied to uncover potential causal links between different factors in railway accident datasets.

In the railway context, semi-automated approaches have shown promise in classifying unstructured close-call reports with high accuracy, suggesting a strong potential for these technologies in future railway safety management

DISADVANTAGES OF EXISTING SYSTEM:

- Existing systems have not fully leveraged advanced ML algorithms such as SVM, ANN, Extreme Learning Machine (ELM), and Decision Trees (DT), which could provide higher accuracy and efficiency.
- Self-Organizing Maps (SOM) have not been widely implemented to categorize human, technological, and organizational factors in railway accident analysis.
- Many current approaches rely on semi-automated processes rather than fully automated systems for accident report classification.
- Text classification models face challenges due to the incremental nature of information extraction and lack of universally effective models for all text types.
- There is limited integration of real-time decision support and interpretability in the existing safety management frameworks.

VI. PROPOSED SYSTEM

This study introduces a novel approach to effectively utilize the textual data from railway station accident reports to identify the root causes of accidents and analyze the relationship between the text content and the possible causes. Currently, a fully automated system capable of processing such text inputs and generating meaningful outputs is not yet available. By implementing this approach, it is expected to address challenges such as providing real-time assistance to decision-makers, extracting key information comprehensible to non-experts, gaining deeper insights into accident details, designing intelligent safety expert systems, and optimizing the use of historical safety records.Our methodology employs the advanced Latent Dirichlet Allocation (LDA) algorithm to extract critical textual information related to accidents and their causes.

ADVANTAGES

Decision Trees (DT) serve as valuable decision support tools by modeling decisions and their potential outcomes in a tree-like structure. Among various machine learning methods applicable to

safety analysis, we specifically train a Decision Tree to classify accidents and identify patterns occurring at railway stations. Additionally, the textual data contains essential information such as the timing, description, location of accidents, and victim age ranges. To enhance mining accuracy, accident times are categorized into parts of the day, enabling a more precise analysis of when accidents typically occur.

System Architecture

Fig1. System Architecture

VII. MODULE DESCRIPTION

Module Description

1. Data Collection Module
Purpose:

2. Data Preprocessing Module
Purpose:
To clean and prepare the collected textual data for analysis by removing noise, inconsistencies, and irrelevant information.

Functionality:
This module performs tasks such as tokenization, stopword removal, stemming, and categorizing accident times into parts of the day. It transforms raw text into a normalized format to improve the accuracy of the machine learning models.

3. Topic Modeling Module Purpose:
To identify and extract underlying themes or topics related to accident causes using unsupervised machine learning techniques.

Functionality:
Using the Latent Dirichlet Allocation (LDA) algorithm, this module analyzes the preprocessed text data to discover critical patterns and root causes of accidents. It outputs meaningful topics that summarize accident characteristics for better understanding and decision support.

4. Classification Module
To gather and organize textual data related to railway
purpose:
station accidents from various sources such as RSSTBo categorize accidents based on their databases or accident reports.

Functionality:
This module extracts raw accident report data, including details like time, location, accident descriptions, and victim information. It ensures data is collected systematically and stored in a structured format suitable for further processing. attributes and identified patterns for more targeted safety analysis.

Functionality:
This module employs a Decision Tree (DT) algorithm to classify accidents into predefined categories, aiding in pattern recognition and predictive analysis. It supports safety experts in detecting recurring accident types and their contributing factors.

5. User Interface Module Purpose:
To provide an accessible platform for users, including safety managers and analysts, to interact with the system, input data, and view results.

Functionality:
This module delivers a web-based front-end built with HTML, CSS, and JavaScript, enabling users to upload accident reports, trigger analysis, and visualize outputs such as topic summaries and classified accident patterns.

6. Database Management Module Purpose:
To store, manage, and retrieve all relevant data efficiently to support system operations.

Functionality:
Using MySQL as the backend, this module manages structured storage of accident data, processed text, model outputs, and user inputs. It ensures data integrity, quick access, and smooth integration with the analysis modules.

7. Reporting and Visualization Module Purpose:
To present analytical results clearly and informatively to support decision-making and safety improvements.
Functionality:
This module generates reports and visualizations such as charts or heat maps showing accident hotspots, root causes, and time-based trends. It helps users interpret complex data and identify actionable insights.

VIII. RESULT

The proposed system effectively employs Latent Dirichlet Allocation (LDA) to analyze unstructured textual data from a large collection of railway station accident reports. Through unsupervised topic modeling, the system identifies underlying patterns and categorizes accidents based on factors such as causes, timing, severity, and contextual information. This approach enables the extraction of critical insights without the need for manual labeling or prior classifications. The system has demonstrated strong capability in revealing high-risk zones and recurring issues, including infrastructural failures and human errors. Additionally, incorporating contextual information such as the time of incidents, descriptions, and demographic data has significantly enhanced the precision of pattern recognition. The results validate the system's effectiveness in generating valuable information from text data, which can be used for preventive safety planning and risk assessment in railway stations.

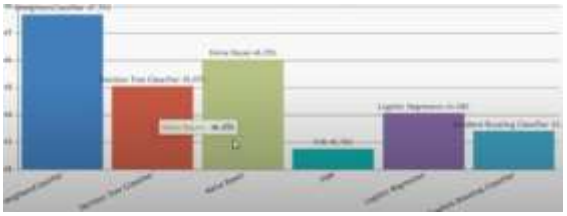

Fig2. Result Graph

IX. CONCLUSION

This study introduces a data-driven approach to railway station safety analysis by utilizing unsupervised machine learning techniques, particularly LDA, to process and interpret accident reports. Unlike traditional methods, this system provides automated, scalable, and in-depth analysis of safety data, facilitating real-time understanding of accident causes and contributing factors. By transforming large volumes of unstructured text into actionable knowledge, the system aids in the development of intelligent safety strategies and supports decision-makers in managing risks more effectively. The findings confirm the model's potential to significantly improve safety monitoring and forecasting, laying the groundwork for future enhancements such as the integration of hybrid models, real-time monitoring, and AI-based decision support systems within the railway infrastructure.

REFERENCES

Srinivasa Reddy, P., and Triveni, V. (2024). "Unsupervised Machine Learning for Managing Safety Accidents in Railway Stations." International Journal of Engineering Research and Science & Technology, 20(2), 1101–1110. ijerst.org.

Kumar, V., Sravani, S., Spurthi, V., and Snehitha, V. (2024). "Unsupervised Machine Learning for Managing Safety Accidents in Railway Stations." International Journal of Mechanical Engineering Research and Technology, 16(9), 180–191. Available at ijmert.com.

Bhargav, P. N. P., and Sreenivasulu, G. (2024). "Unsupervised Machine Learning for Managing Safety Accidents in Railway Stations." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(3), 129–135. ijsrcseit.com.

Eskisehir Technical University. (2024). Anomaly Detection in Railway Images Using Unsupervised Clustering of Infrared Thermography. Journal of Natural Sciences and Technologies, 3(1), 259–261. journalofnastech.com

5. Wang, T., Zhang, Z., Yang, F., and Tsui, K.-L. (2021). Intelligent Railway Foreign Object Detection: A Semi- supervised Convolutional Autoencoder Based Method. arXiv. arxiv.o rg.

6Giri,V.C.,&Jha,P.(2025).Enhancing Railway Accident Prevention Using Deep Learning, Machine Learning, and GPS Tracking: A Historical and Knowledge- Based Analysis. International Advanced Research Journal in Science, Engineering and Technology, 12(2), 237–242. iarjset.com

7. Heidarysafa, M., Kowsari, K., Barnes, L. E., & Brown, D. E. (2018). Analysis of Railway Accidents' Narratives Using Deep Learning. *arXiv*. arxiv.org

8. Choudhary, J. K., & Raj, A. (2025). AI System Rolled Out to Boost Rail Safety and Maintenance. *Times of India*.