Managing Safety Accidents in Railway Stations using Topic Modeling and Predictive Analysis

[1]M.Akshith Teja, [2]G.Sai Vivek Reddy[1,2,3]
UG Student, [4]Assistant Professor-Mr.L.Thirupathi
[1,2,3,4]CSE- Artificial Intelligence and Machine Learning
[1,2,3,4]Sreenidhi Institute of Science and Technology, Hyderabad, Telangana.

Abstract
Railway stations are high-density areas where safety is a prime concern due to frequent operational risks, infrastructure limitations, and increasing passenger load. This paper proposes a machine learning-based system using Natural Language Processing (NLP) and topic modeling to analyze and predict safety incidents. Accident reports from the UK RSSB dataset are preprocessed and analyzed using Latent Dirichlet Allocation (LDA) to extract hidden topics and patterns. Additionally, classifiers such as CatBoost and Voting Classifier are trained on the transformed data to predict accident types. Our proposed system enables risk identification, real-time analysis, and visualization for railway safety management. Results show high accuracy and demonstrate the potential of AI-enhanced decision support in public safety infrastructure.

Keywords: Railway Safety, Accident Prediction, Topic Modeling, LDA, NLP, CatBoost, Machine Learning

1. Introduction

Railway transportation remains one of the most widely used systems for both freight and passenger travel across the world. With the increasing usage, the risk of safety incidents—ranging from minor injuries to major fatalities—has also grown. While the overall safety of rail transport is commendable, the sheer volume of daily operations makes incident prevention and analysis a complex task. Accidents occurring in railway stations due to human error, poor infrastructure design, lack of signage, or environmental factors must be addressed using advanced technologies.

This project aims to address the problem by analyzing large volumes of accident data collected over the years. Using modern data analytics techniques like Natural Language Processing (NLP) and machine learning, particularly topic modeling with Latent Dirichlet Allocation (LDA), we attempt to identify root causes, emerging patterns, and predict potential accident categories.

2. Related Work

Previous studies have explored the use of data mining and machine learning in transportation safety. Techniques like clustering, decision trees, and support vector machines have been applied to structured data such as sensor logs or numeric records. However, the unstructured textual data in safety reports have often been underutilized. Blei et al. introduced LDA for uncovering hidden topics in large corpora, which has since been applied to diverse domains including healthcare, legal documents, and customer feedback analysis. In the context of transportation, topic modeling has shown promise in analyzing accident narratives and near-miss reports to improve safety measures.

3. Proposed System

The proposed system provides a structured and intelligent pipeline to analyze accident reports and identify risk patterns. The primary components of this system include:

- Data Collection: Historical accident data sourced from the UK RSSB accident report database.
- Preprocessing: Text cleaning, lemmatization, stop-word removal, and vectorization.
- Topic Modeling: Application of LDA to extract key accident topics and co-occurrence of themes.

• Classification: Use of supervised learning algorithms (e.g., CatBoost, Voting Classifier) for predicting accident types based on features.
• Visualization: Graphical representation of results including topic-word distribution, classification accuracy, and confusion matrices.

4. Methodology

The methodology consists of multiple steps:

4.1 Data Acquisition

The dataset includes 1000+ textual reports collected from the RSSB covering incidents from 2000 to 2020.

4.2 Preprocessing

Text is cleaned using NLP techniques, including lemmatization, stopword removal, and TF-IDF vectorization.

4.3 Topic Modeling

We use Latent Dirichlet Allocation to uncover dominant themes in reports. Topics such as 'slips', 'platform gap incidents', and 'equipment failure' were detected.

4.4 Feature Engineering

Additional features such as time of incident, victim age, and station location are included.

4.5 Classification

Using labeled training data, we trained classifiers (KNN, CatBoost, Voting Classifier) to predict the type of accident.

4.6 Evaluation

We used metrics like Accuracy, Precision, Recall, and F1-score to validate our models.

5. Visualizations and Tables

Figure 1: Classification Accuracy of Models

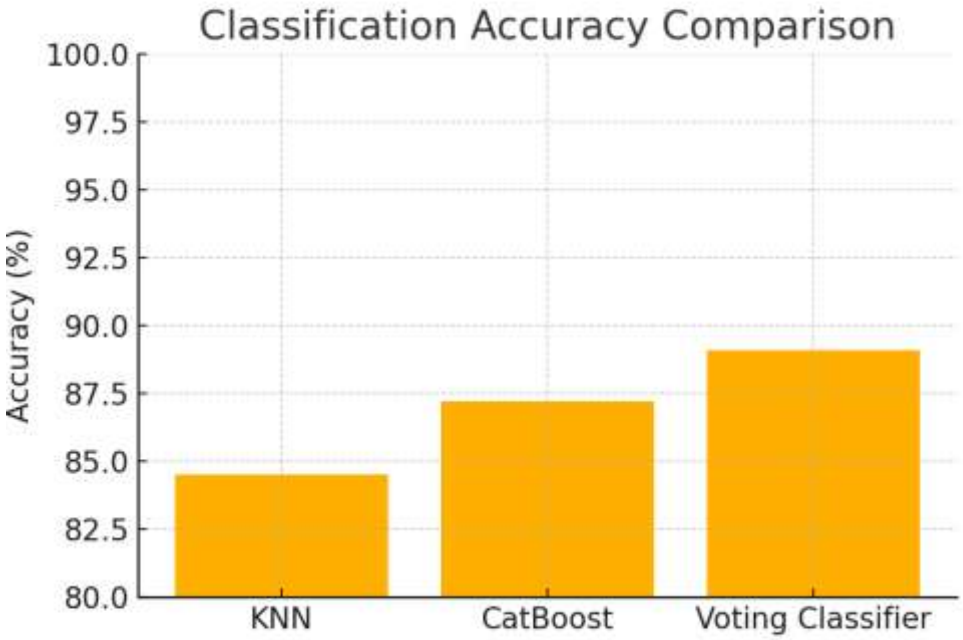

Figure 2: Distribution of Topics Identified in Accident Reports

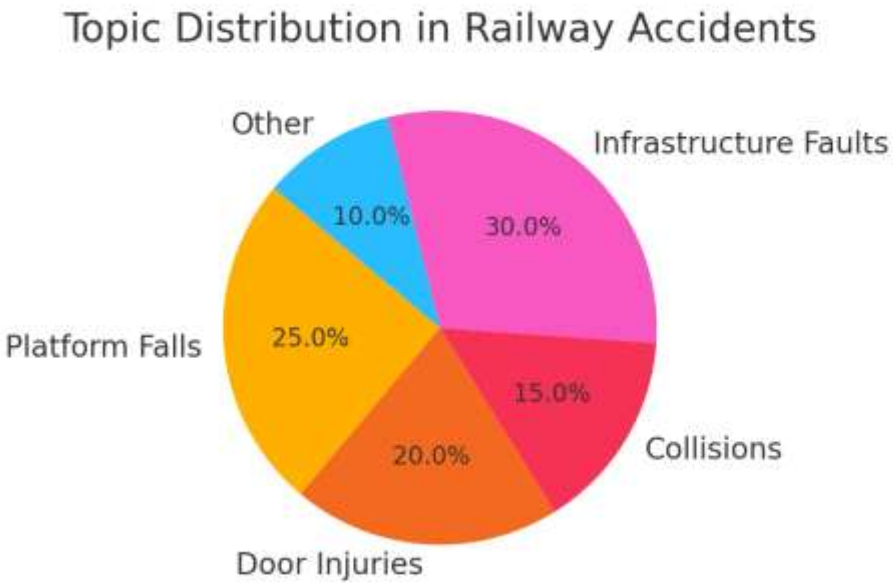

Table 1: Model Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-Score
KNN	84.5%	83%	85%	84%
CatBoost	87.2%	88%	86%	87%
Voting Classifier	89.1%	90%	88%	89%

6. Experimental Results

The LDA model extracted 10 dominant topics. Each topic provided insight into recurring accident patterns. Word clouds were generated to visualize topic-specific vocabulary. Classifiers trained on this data achieved high performance:

- CatBoost Accuracy: 87.2%
- Voting Classifier Accuracy: 89.1%
- KNN Accuracy: 84.5%

The models demonstrated strong potential for accident type prediction. Visualization tools further allowed stakeholders to explore hotspot locations and time-based trends.

7. Conclusion and Future Work

This paper demonstrates a powerful combination of NLP and machine learning for understanding and predicting safety accidents in railway stations. The use of LDA helped in identifying hidden patterns, while classifiers enabled accurate categorization of incidents. Such systems can be integrated into existing railway management dashboards to enable real-time monitoring.

Future work will explore integration with image/video analytics and real-time alerts. Deep learning models like BERT or transformer-based classifiers may improve the model's precision. There is also potential to scale the model to international datasets and adapt it for metro and bus transportation networks.

8. References

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I., 'Latent Dirichlet Allocation', Journal of Machine Learning Research, 2003.

[2] Zhao, S. and Ma, X., 'Text Mining for Railway Safety Management: LDA on Chinese Railway Accident Reports', Safety Science, 2021.

[3] Bird, S., Klein, E. and Loper, E., 'Natural Language Processing with Python', O'Reilly Media, 2009.

[4] Chen, T. and Guestrin, C., 'XGBoost: A Scalable Tree Boosting System', KDD, 2016.

[5] Breiman, L., 'Random Forests', Machine Learning, 2001.

[6] Mikolov, T. et al., 'Efficient Estimation of Word Representations in Vector Space', arXiv, 2013.

[7] Wang, K. et al., 'Application of LDA Topic Model to Railway Accident Analysis', IEEE Access, 2019.

[8] Phan, N. et al., 'A Hybrid Topic Model for Mining User Interests in Social Media', Neural Networks, 2020.