Transforming Railway Accident Data with NLP

Enhancing Safety through Text Mining and AI

~ Sridhar Reddy Maram
LA32371

Research Problem

- 35% of railroad accidents are human-caused (2009-2020 data).

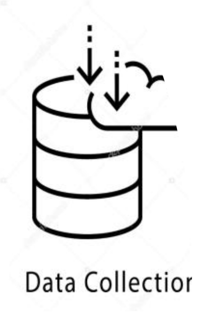
- Current systems like Positive Train Control (PTC) cannot address all human errors.

- Need to identify factors that contribute to human-caused accidents beyond just speed and derailment.

Reference Dataset :

- The FRA dataset contained more th accident records from January 1, 2009, 2020, each containing 145 fields (FRA, 20

Literature Review



- **ML -** to predict accidents based on fixed attributes like train speed, weather condition, track conditions, etc..(Structured Data)

- **NLP -** compared latent semantic analysis (LSA) and latent Dirichlet allocation (LDA) to classify accident narratives (Unstructured Data).

- **Combining ML with Text Mining -** to p severity of train accidents, and the use of Sha theory to rank the contribution of features.

Methodology

- Data Handling
  - FRA Dataset (2009 to 2020)
  - Text Cleaning
    - Tokenization
    - Stop words removal
    - Noise removal
    - Normalization
  - Feature Extraction
    - One Hot Encoding
  - Data Filtering
    - Remove irrelevant data
    - Handle Missing values
  - Data Balancing
    - Sampling technique
- EDA
  - Visualizations
  - Word cloud
  - Clustering
- Model Deployment
  - ML Algorithms
  - NLP Models
- Model Evaluation
  - Performance Metrics
  - Feature Rankin

Results Analysis and Interpretation

- ✓ Human-caused accidents are often not associated with high speeds or derailments.
- ✓ Key Features identified by Shapley Values were strong indicators of human-caused accidents.
- ✓ NLP Uncovered Operational Risks tha Structured Data Missed.

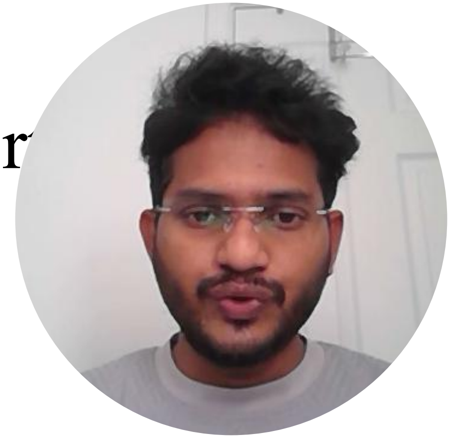
Future of NLP in Railway Safety

Challenges:

- Handling complex language in accident reports.
- Large datasets needed for accurate model training.

Opportunities:

- Real-time data integration for dynamic risk prediction.
- Advanced NLP techniques like BERT and GPT for deeper analysis of accidents.
- Integration with IoT for automated safety aler



Conclusion

Conclusion

- The combination of ML and NLP provides a comprehensive understanding of human-caused railroad accidents.

- Shapley game theory provides a powerful tool for understanding feature importance for both structured and unstructured data.

- Expected Outcomes:
    - i) Policy Implications
    - ii) Management Decisions
    - iii) Future Research

- Bridgelall, R., & Tolliver, D. D. (2023). Railroad accident analysis by machine learning and natural language processing. Journal of Rail Transport Planning & Management, 29, 100429. https://doi.org/10.1016/j.jrtpm.2023.100429

- Syeda, Kanza & Shirazi, Syed Noorulhassan & Naqvi, Syed & Parkinson, Howard & Bamford, Gary. (2019). Big Data and Natural Language Processing for g Railway Safety: Analysis of Ra Incident Reports. 10.4018/978- 1.ch040.

धन्यवाद
THANK YOU
धन्यवाद