Enhancing Preventive Measures in Hydrogen Industries: Root Cause Identification Using NLP Techniques

Plínio Ramos[a*], July Macedo[a], Caio Souto Maior[b], Márcio Moura[a], Isis Lins[a]

[a]Universidade Federal de Pernambuco, Recife, Brazil
[b]Universidade Federal de Pernambuco - CAA, Caruaru, Brazil

Abstract: The European Commission has identified hydrogen as a key solution to reduce greenhouse gas emissions, with the International Energy Agency reporting a significant rise in low-emission hydrogen production projects. However, safety concerns remain a major obstacle due to hydrogen's unique properties, including its high flammability and potential for material degradation. Thus, hydrogen-induced material degradations pose additional risks to equipment integrity and accidents involving hydrogen can lead to catastrophic effects. Therefore, correctly detecting the cause of failure is crucial for implementing preventive measures. By documenting and analyzing incidents, companies can identify patterns or trends that may indicate systemic issues requiring attention. This knowledge can inform training programs, safety protocols to minimize the likelihood of similar accidents in the future. However, the sheer volume of accident reports makes human review impractical. Thus, this study proposes employing Natural Language Processing (NLP) techniques to automate the detection of the root causes in hydrogen-related accident narratives. To that end, this study partially automates the creation of a labeled dataset and build a classifier based on Bidirectional Encoder Representation from Transformers (BERT) to identify accidents' causes. The model's effectiveness is tested on the Hydrogen Incidents and Accidents Database (HIAD) 2.1, established by the European Commission-funded Network of Excellence on Hydrogen Safety (HySafe). By automating the analysis of accident narratives, this research contributes to enhancing the proposal of preventive associated with hydrogen-related accidents.

Keywords: Hydrogen, Safety, Accident, Natural Language Processing

1. INTRODUCTION

Global climate change and ecological damage caused by fossil energy have garnered attention for sustainable energy transitions in recent years. Hydrogen is distinguished as a clean and highly efficient energy carrier, with the potential to significantly reduce dependence on fossil fuels across all energy sectors (Huan *et al.*, 2024). However, during the production and operation of hydrogen industrial systems, the occurrence of a leakage accident can easily lead to fire and explosion (Lu *et al.*, 2024). In fact, hydrogen possesses a low ignition temperature and a broad explosive range (Adamson and Pearson, 2000). Due to its smaller molecular volume compared to natural gas, hydrogen can easily penetrate pipe gaps, causing leaks. After hydrogen leakage, a jet is formed and gradually transforms into a plume. Thus, depending on the mixture concentration, hydrogen ignition can occur upon encountering the lowest ignition energy (Yang *et al.*, 2021).

Hydrogen leaks are the leading causes of accidents at hydrogen refueling stations, resulting in severe loss of life and/or property. For instance, accidents like the explosion at the Sandvika hydrogen refueling station in Norway resulted in injuries and the temporary closure of several hydrogen refueling stations in Norway, Denmark, and neighboring countries for an extended period (Wang *et al.*, 2024). Similarly, the hydrogen tank explosion in Gangneung, South Korea, resulted in deaths and injuries. Thus, amid growing concerns about recent hydrogen-related accidents in leading hydrogen energy-adopting countries, maintaining public confidence in hydrogen infrastructure is crucial for advancing energy transitions. Therefore, the risk assessment in hydrogen handling facilities has attracted significant attention from many experts and academics (Zhang *et al.*, 2022).

In this context, risk analysis (RA), which involves identifying and managing risks associated with specific activities, emerges as an important strategy to mitigate, and reduce these risks to acceptable levels. Concurrently, accident analysis plays a crucial role in accident prevention. By learning from accidents and extracting insights, accident prevention efforts become more targeted, encompassing the formulation of regulations, risk management, and knowledge training (Jia *et al.*, 2024).

In parallel, the advancement of computer technology and linguistics has led to the growing application of natural language processing (NLP) technology across various domains. As the number of accident analyses reaches a certain threshold, patterns can be discerned from the causes of accidents, enabling the implementation of universal preventive measures, even on an industry-wide scale. Therefore, analyzing accident causes based on a large dataset of accidents proves to be an effective approach to accident prevention, with manual analysis being the primary method. However, accidents often yield vast amounts of unstructured text data, making manual analysis time-consuming and labor-intensive (Yan *et al.*, 2021). Thus, the sheer volume of accident reports makes human review impractical (Macêdo *et al.*, 2022; Ramos *et al.*, 2022).

NLP techniques hold promise in supporting RA since they can be applied to extract, organize, and classify information from a text (McDonald, Ade and Peres, 2020). Data-driven approaches are increasingly employed to enhance RA. Studies by (Ahmadpour-geshlagi *et al.*, 2020; Baker, Hallowell and Tixier, 2020; Kutela, Das and Dadashova, 2022; Janstrup *et al.*, 2023) utilize NLP to extract information from accident investigation reports, limitations persist, particularly in the context of aviation accidents. For example, (Kuhn, 2019) utilized Latent Dirichlet Allocation (LDA) topic modeling to patterns in motor vehicle crash records. However, LDA models a document as a Bag-of-Words (BoW), ignoring the contextual information of words within a sentence. (Zhang, Srinivasan and Mahadevan, 2021) used the National Transportation Safety Board (NTSB) texts for building supervised machine learning models for performing the prognosis of adverse events like accidents, aircraft damage, or fatalities. However, these authors did not focus on the identification and analysis of the causes of accidents.

This study addresses the need to apply NLP to detect the cause of accidents inferred from its narratives. The proposed methodology applies contextual word-vector representations derived from pre-trained Bidirectional Encoder Representation from Transformers (BERT) (Devlin *et al.*, 2018) to identify the root causes of accidents contained in the Hydrogen Incidents and Accidents Database (HIAD) 2.1, created by the Joint Research Centre (JRC) of the European Commission as part of the Hydrogen Safety Excellence Network (HySafe) 2004–2009 (Daniele, Jennifer Xiaoling and Moretto, 2019). The remainder of the study is organized as follows, Section 2 provides an overview of hydrogen accidents, highlighting their significance and the challenges associated with their prevention and management accidents. Section 3 provides insights into how NLP techniques can be applied to accident narratives to extract meaningful information; Section 4 describes the methodology used in this study. Section 5 presents the findings of the study, followed by Section 6 which concludes.

2. HYDROGEN ACCIDENTS DATABASE

Previous research has primarily focused on analyzing specific accidents to identify causative factors. For instance, studies highlighted the significance of organizational and personnel factors in hydrogen accidents (Lu *et al.*, 2024). Similarly, (Sakamoto *et al.*, 2016) emphasized design errors and maintenance deficiencies as common causes of accidents at hydrogen refueling stations. Understanding these factors is crucial for formulating effective preventive measures and mitigating the risk of hydrogen leakage accidents. Therefore, the creation of structured databases and repositories, along with thorough risk analysis, is crucial for understanding the causes of hydrogen leakage accidents. This understanding enables the implementation of effective preventive measures to enhance safety in companies involved with hydrogen. These databases play a significant role in reporting and analyzing accidents across different industrial sectors and social activities.

In Europe, databases such as Accident Reporting Information Analysis (ARIA) (BARPI, 2024) and European Major Accident Reporting System (eMARS) (European Commission, 2024) collect incident reports and investigations related to industrial accidents, including those involving hazardous chemicals. Similarly, databases like Relational Information System for Chemical Accidents Database (RISCAD) (AIST, 2024) in Japan and those maintained by regulatory bodies in the United States compile information on accidents and their causal factors.

However, only two databases specifically focus on hydrogen-related accidents: HIAD 2.1 and Hydrogen Tools Lessons Learned (H2TOOLS), developed by the Pacific Northwest National Laboratories (PNNL) and

financed by the U.S. Department of Energy (PNNL, 2024). These databases aim to provide extensive information on hydrogen-related accident events. While H2TOOLS offers detailed lessons learned from previous hydrogen-related events, HIAD 2.1 allows for large-scale statistical evaluations, providing valuable insights into the risks associated with hydrogen production, handling, and storage.

Despite these advances, the development of risk prediction models for hydrogen leak accidents remains a challenge. Addressing this challenge requires the integration of scientific methods to analyze causal factors and the development of efficient risk prediction models. Leveraging NLP techniques offers a promising avenue for analyzing accident reports and extracting valuable information to improve risk prediction and accident prevention efforts in hydrogen-related industries.

3. NATURAL LANGUAGE MODEL

Transformers-based models have demonstrated their efficacy in NLP by acquiring universal language representations through training on extensive text corpora. However, training transformers from scratch requires significant computational resources and time (Han and Wang, 2021). To address this challenge, transfer learning enables the utilization of knowledge acquired from source tasks (i.e., pretraining tasks) to facilitate downstream tasks. Additionally, the availability of labeled datasets for NLP tasks can be limited. To overcome this hurdle, self-supervised learning empowers transformers models to learn through pseudo-supervision, utilizing one or more pretraining tasks to extract valuable language information (Kalyan, Rajasekharan and Sangeetha, 2021).

BERT and Generative Pre-trained Transformer (GPT) were pioneering pre-trained language models that utilized transformer encoders and decoders, respectively (Devlin *et al.*, 2018). In the context of risk and reliability engineering, despite the advances when considering studies related to core NLP tasks, it is common to find applications using classical models such as BoW, TF-IDF, and Doc2Vec. Moreover, BERT implementations in Pytorch and Tensorflow have been available for more than three years (Wolf *et al.*, 2020), in different languages, stably, with no long-term compatibility problems between libraries. Hence, we developed our methodology based on BERT, as it gives us flexibility and robustness.

Multiple variations of pre-trained BERT models are available for download, allowing users to fine-tune these models for specific supervised learning tasks (Nguyen, Le and Le, 2021). This involves adding an untrained layer of neurons on top of the pre-trained BERT model. Overall, during fine-tuning, the pre-trained parameters are adopted to initialize the model, which is then updated using labeled data tailored to the supervised task (Macêdo et al., 2022). For instance, we can adjust BERT's architecture by adding one output layer on top of the pre-trained model to adapt it for performing a classification task. For instance, we can adjust BERT's architecture by adding one output layer on top of the pre-trained model to adapt it for performing a classification task. It's worth mentioning that the parameters related to the additional layer are the only parameters that require random initialization and learning from scratch. This approach enables the construction of state-of-the-art architectures within a reasonable timeframe (Howard & Ruder, 2018). For more details see (Devlin et al., 2018).

4. METHODOLOGY

4.1. Dataset

The study is based on data available on the HIAD, a repository tool that gathers reports of industrial accidents related to hydrogen and its derivatives. The JRC of the European Commission created HIAD as part of the HySafe 2004–2009 (Daniele, Jennifer Xiaoling and Moretto, 2019). New events were regularly provided to HIAD given that JRC experts were responsible for maintaining and updating the database. These events were reviewed and validated by JRC experts before being made public. The HIAD database aimed to facilitate the exchange of lessons learned from hazardous events involving hydrogen to improve the information network and prevent similar unexpected events in the future (Jones, Kirchsteiger and Bjerke, 1999). In 2017, the JRC, together with the Fuel Cell and Hydrogen Joint Undertaking (FCH 2 JU), updated the HIAD database to HIAD 2.1 and integrated it as part of the European Hydrogen Safety Panel (EHSP) 2009-2022 activities (Daniele, Jennifer Xiaoling and Moretto, 2019).

Regarding the information collected in HIAD 2.1, the database contains all the parameters necessary to understand what happened and how the undesirable event can be described in detail. Currently, all records collected have been exported to an Excel workbook that allows users to access and analyze the data according to their needs (the Excel workbook used in this study is updated as of January, 2024). The file contains six sheets:

- Events – main classification, narrative summary, systems involved, date, location, and cause classification;
- Facility – description of applications, storage conditions, type of location, and pre-event conditions;
- Consequences – effects in terms of human and property losses for the affected facilities;
- Lessons Learned – corrective measures adopted;
- Event Nature – quantitative information on emergency action, leakage characteristics, leak type, and fire consequences;
- Reference – the primary source of information.

Reports from the HIAD 2.1 database are intended for public use. The Event spreadsheet, visually demonstrated in Figure 1, contains some parameters allow gathering detailed information from each event among the sheets: Event ID (i.e., the record number in the database), Quality Seal (i.e., information regarding the level of detail of the report), Full Description (i.e., the descriptive summary of the incident, with detailed information) and Causes (i.e., the cause(s) of the accident).

Event ID: 11

Accident information

- **Evente title:** Hydrogen fire from a pipeline
- **Event full description:** "... The fire resulted in flame impingement on the support of a 100-foot high (approximately 30 m) reactor in a hydrocracker unit. ..."
- **Hydrogen system initiating event:**
  ☒ Yes ☐ No
- **Classification of the physical effects:** Hydrogen release and ignition.
- **Nature of the consequences:**
  ☐ Explosion ☐ false alarm ☒ Fire ☐ Leak no ignition ☐ Near miss
- **Region:** North America
- **Country:** United States
- **Date:** 1937-06-05
- **Causes of accident:**
  ☐ Human factors, ☐ Management factors, ☐ Job factors, ☒ installation error, ☒ Material/manufacturing error, ☒ System design error
- **Causes comments:** "The cause of the initial hydrogen leak is believed to have resulted from the failure of an elbow to reducer weld in the 2-inch (51 mm) hydrogen preheat exchanger bypass line."

Quality Seal: 3

Figure 1 - HIAD 2.1 example

HIAD 2.1 considers six cause categories. Three of the accident causes pertain to human factors, encompassing job, individual, and organizational aspects according to the Health and Safety Executive (HSE) definition:

- Individual/human factors: Encompass inadequate skill and competence levels, fatigue, disengagement, and individual medical issues.
- Management system factors: Encompass poor planning resulting in overworked staff, inadequate safety systems, failure to learn from previous incidents, one-way biased communication, lack of coordination and defined responsibilities, poor management of health and safety, and deficient safety

culture. Some incidents highlighted outdated or lacking operative and maintenance guidelines, particularly concerning external contractors.
- Job factors: Include inappropriate equipment and instrument design, design flaws, missing or unclear instructions, poorly maintained equipment, high workload, noisy or unpleasant working conditions, constant interruptions, and disturbances.

The remaining three accident causes relate to system design, material, manufacturing, and installation:

- Installation error: Despite correct component selection and implementation, a malfunction occurs due to improper installation or maintenance. Examples include the absence of a thermally activated pressure relief device (TPRD) on a gas bottle or cylinder or disregard for installation instructions of a safety device.
- Material/manufacturing error: Despite correct component selection and implementation, malfunction arises due to material failure or manufacturing error.
- System design error: Occurs when the system is inadequately designed for hydrogen use or operating conditions. Examples include incompatible components, absence of ATEX components, when necessary, unexpected hazardous gas mixture, unforeseen pressure or temperature loads, and incorrect selection of solenoid/electromechanical valve type.

It is noteworthy that not all these fields are consistently filled, and the quality of the descriptions depends entirely on the information provided by the primary sources and their level of detail. Quality seals are provided and range from 2, if most of the quantitative descriptors are missing, to 5, if lessons learned and root cause analyses are available with good technical detail. For approximately 1.4% of the total events, a final Quality Seal assessment is still missing. Additionally, 48.7% were classified as “Low quality” since most quantitative descriptors are not provided; for 28.7% of the total events, the information source is considered “Good quality”. Furthermore, 9.9% and 10.9% of the total events have “High quality” and “Very high quality” reports, respectively, in which root cause analyses and lessons learned are available, and quantitative technical details are provided

4.2. Pipeline

Using the HIAD 2.1 database, this research develops a BERT-based model to learn and identify the causes of hydrogen-related accidents. The following pipeline (Figure 2) aims to preprocess and filter the data from the database to facilitate its input into the BERT model.

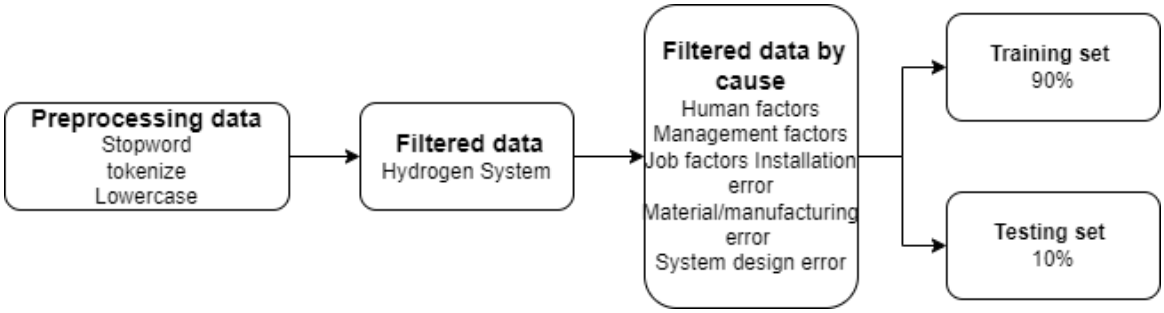

Figure 2 - Pipeline for filtered data

4.2.1 Preprocessing Dataset

In this subsection, we preprocess the dataset to prepare it for analysis. The preprocessing steps involve cleaning and transforming the text data to remove noise and irrelevant information. First, we import the necessary libraries, including pandas for data manipulation and NLTK for NLP preprocessing tasks. We download the stopwords and word tokenization modules from the NLTK library. Next, we define a preprocessing function that converts text to lowercase, removes numbers, punctuation, and stopwords, and

tokenizes the text into words. Then, we apply the preprocessing function to a column in the dataset containing the full description of each event. We create a new column to save the preprocessed text.

Next, we filter the dataset to include only events related to hydrogen systems and remove entries with unknown causes. Subsequently, we extract subsets of data corresponding to different cause categories, such as human factors, management factors, job factors, installation error, material/manufacturing error, and system design error. Finally, the dataset is divided into training and testing sets using a 90-10 ratio for each cause category, ensuring that each category has a balanced representation in both sets.

4.2.2 Modeling

To build our classifier, BERT model is utilized for sequence classification tasks. Specifically, the 'bert-base-cased' pre-trained model is employed. The training procedure involves data loading, tokenization, model instantiation, optimizer setup, training loop, and validation (as illustrated in the pseudocode in Figure 3).

```
# Define classifier model
Define function create_classifier_model():
    Load pre-trained BERT model
    Add dense layer with sigmoid activation

# Train model
Define function train_model(model, training_data, epochs):
    For each epoch in range(epochs):
        For each batch in training_data:
            Preprocess batch text
            Pass preprocessed text through BERT model
            Compute binary cross-entropy loss
            Backpropagate to update model parameters

# Evaluate model
Define function evaluate_model(model, test_data):
    Initialize evaluation metrics
    For each batch in test_data:
        Preprocess batch text
        Pass preprocessed text through trained model
        Update evaluation metrics with predictions

# Main script
Define main():
    training_data = Load training data
    test_data = Load test data
    epochs = Define number of epochs

    model = create_classifier_model()
    train_model(model, training_data, epochs)
    evaluate_model(model, test_data)

# Run the main script
main()
```

Figure 3 - Pseudocode for building the classifier

First, after preprocessing, the accidents' narratives are tokenized and converted into input sequences suitable for the BERT model. The BERT model is modified for multi-label classification tasks with a specific number of output labels. To do that, the final hidden state corresponding to the [CLS] token ($BERT_{CLS}$) is extracted from BERT model and is passed through a *sigmoid* ($W.BERT_{CLS} + b$) layer, where W and b are the weights and the biases respectively, to obtain the predicted class probabilities. During training, we optimize the model using the cross-entropy loss, which measures the difference between the true labels and the predicted probabilities. Next, training, validation, and testing datasets are loaded using PyTorch DataLoader for efficient processing. Then, the 'AdamW' optimizer is employed with different learning rates to fine-tune the BERT model. The model is trained for a fixed number of epochs. Each epoch involves iterating over the training dataset in batches, computing loss, and updating model parameters. At the end of each epoch, the model's performance is evaluated on the validation dataset to monitor training progress and prevent overfitting.

To evaluate the model's performance under different learning rates and epochs a series of experiments are conducted to identify the most favorable hyperparameters. Then, to ensure that each label category is

represented adequately, training data is filtered to maintain a balanced distribution of labels. Moreover, additional experiments include data augmentation techniques to enhance model generalization and performance. The models tested are described below:

- Experiment 1 (Baseline Training): Initial training runs evaluate performance with varying learning rates, lr, and epochs, e:
  $lr = 10^{-6}$, e = 25, batch size = 4
  $lr = 10^{-5}$, e = 50, batch size = 4
- Experiment 2 (Data Balancing): To address label distribution imbalance, training data is filtered to ensure adequate representation of each label category:
  $lr = 10^{-6}$, e = 25, batch size = 4
  $lr = 10^{-5}$, e = 50, batch size = 4
- Experiment 3 (Data Balancing + Data Augmentation – DA): Further configurations integrate data augmentation techniques to enhance model generalization:
  $lr = 10^{-6}$, e = 25, batch size = 4
  $lr = 10^{-5}$, e = 25, batch size = 4
  $lr = 10^{-5}$, e = 15, batch size = 4

These experiments aim to assess the impact of preprocessing techniques, hyperparameters, and data augmentation on the model's ability to accurately identify the root causes of accidents inferred from accident narratives. All experiments were implemented using Python, employing libraries such as PyTorch and Transformers. PyTorch provided the foundational framework for model development, while Transformers facilitated the integration of pre-trained BERT models, crucial for the NLP tasks undertaken.

This study centers on developing an NLP-based model to analyze aviation accident narratives and determine root causes. Specifically, the model aims to discern whether accidents were attributable to Job factors, Individual/human factors, Management system factors, System design error, Material/manufacturing error, or Installation error. The experiments were conducted on a Windows machine equipped with an Intel(R) Core(TM) i9-9900K processor (CPU @ 3.60GHz 3.60 GHz) and 32 GB of RAM.

5. RESULTS

In this section, we analyze the results of the experiments conducted to evaluate the performance of our NLP-based model in identifying root causes of accidents from accident narratives. We explored various configurations and preprocessing techniques to assess their impact on model accuracy and robustness.

- Experiment 1: The best model achieved a test accuracy of approximately 63.7%.
- Experiment 2: Test accuracies varied between approximately 57.8% and 63.1%.
- Experiment 3: The models consistently achieved test accuracies between approximately 60.9% and 63.1%.

Figure 4 presents the confusion matrices for the first experiment, which yielded the best results. Each row in the matrices represents true labels, and each column represents predicted labels. For clarity, the cell (0,0) indicates the true negatives, (0,1) indicates false positives, (1,0) indicates false negatives, and (1,1) indicates true positives.

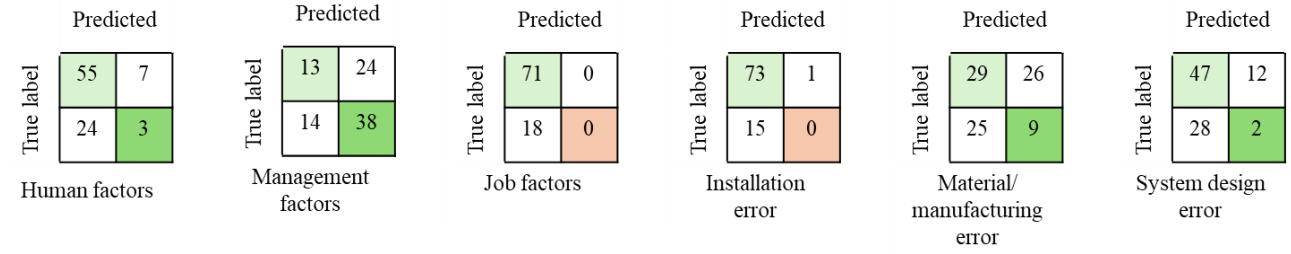

Figure 4 - Confusion matrices with the model's predictions on the test set.

The results reveals that the model exhibits challenges in handling categories with lower training data frequencies, specifically labels 2 ("Material/manufacturing error") and 3 ("Design error"). These labels, with only 68 and 61 training instances respectively, contributed to the model's poor performance in classifying positive instances for these categories. Conversely, the model performed reasonably well in identifying "Installation Error" but showed a tendency towards false positives, suggesting an over-identification of this factor. For "Management Factors", the model demonstrated balanced performance with a notable number of true positives; however, it also faced a significant number of false negatives, indicating missed relevant cases.

The inability to achieve a significant improvement in the model's performance can be attributed to several factors: the effectiveness of NLP models often hinges on the quality and quantity of training data available. If the dataset contains noise or biases, it can hinder the model's ability to learn complex patterns effectively. In addition, the nature of the accident narratives dataset and the complexity of identifying the root causes of accidents present challenges that may require specialized techniques or domain knowledge to overcome.

To address the limitations of the BERT-based model, we developed a BigBird-based classifier designed to handle larger text inputs, potentially enhancing performance. We initially trained BigBird with the same hyperparameters and 80/20 train/test split ratio as the best-performing BERT-based model. Despite the increased complexity and number of parameters, BigBird's performance was initially inferior. We adjusted the train/test split ratio to 90/10 to improve performance.

Overall, while the model shows some proficiency in identifying certain factors, significant improvements are needed to address the high rates of false positives and false negatives, particularly for less frequent categories. We can aim to enhance the model's overall performance and reliability by exploring alternative strategies such as experimenting with more advanced models, incorporating additional data from external sources to provide more context, and ensuring high-quality, detailed, and consistent data entries.

Figure 5 shows the confusion matrix for BigBird's predictions on the test set. The BigBird-based classifier achieved an accuracy of 64.44%. Although there were improvements in specific areas, significant issues persisted with high rates of false negatives and false positives, particularly in less frequent categories. Furthermore, the higher computational cost associated with BigBird did not result in a substantial overall performance gain.

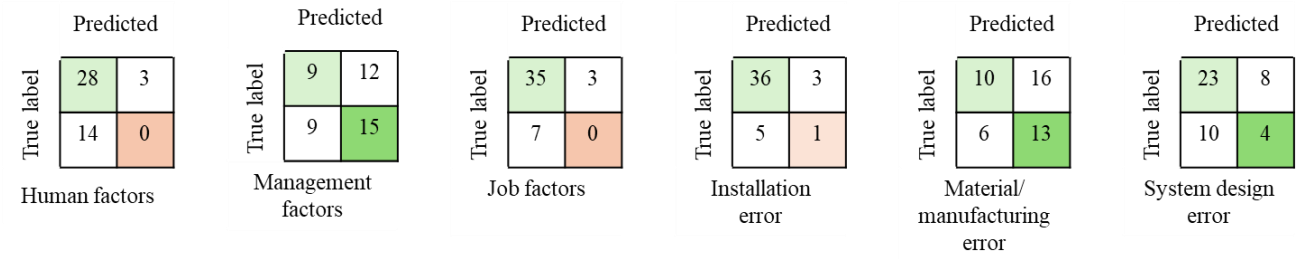

Figure 5 - Confusion matrices with fine-tuned BigBird's predictions on the test set.

While the model demonstrates some proficiency in identifying certain factors, further improvements are necessary to reduce false positives and false negatives, especially for less frequent categories. Future work should focus on exploring advanced models, integrating additional contextual data from external sources, and ensuring high-quality, detailed, and consistent data entries to enhance model performance and reliability.

6. CONCLUSION

The experimental results underscore the model's proficiency in identifying root causes of accidents from narrative descriptions. However, further refinement and optimization since challenges persist in accurately classifying certain label categories. Addressing these challenges may require a combination of methodological refinements, such as feature engineering techniques and model architecture improvements. Additionally, to iteratively refine the model and uncover insights into its behavior continued experimentation

and iteration are essential. Overall, the findings contribute to our understanding of the model's capabilities and lay the groundwork for future research in this domain.

Acknowledgements

The authors thank the following Brazilian research funding agencies for the financial support: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), 310892/2022-8, 402761/2023-5; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001. We also thank the Fundação de Amparo a Ciência e Tecnologia de Pernambuco (FACEPE), APQ-0394-3.08/22 and Human Resources Program 38.1 (PRH 38.1) entitled Risk Analysis and Environmental Modeling in the Exploration, Development and Production of Oil and Gas, financed by Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP) and managed by Financiadora de Estudos e Projetos (FINEP), 044819.

References

Adamson, K.-A. and Pearson, P. (2000) 'Hydrogen and methanol: a comparison of safety, economics, efficiencies and emissions', *Journal of Power Sources*, 86(1–2), pp. 548–555. Available at: https://doi.org/10.1016/S0378-7753(99)00404-8.

Ahmadpour-geshlagi, R. *et al.* (2020) 'Investigating the status of accident precursor management in East Azarbaijan Province Gas Company', *International Journal of Occupational Safety and Ergonomics*, pp. 1–12. Available at: https://doi.org/10.1080/10803548.2020.1770451.

AIST (2024) 'RISCAD Database', *https://www.aist.go.jp/* [Preprint].

Baker, H., Hallowell, M.R. and Tixier, A.J.P. (2020) 'Automatically learning construction injury precursors from text', *Automation in Construction*, 118(June), p. 103145. Available at: https://doi.org/10.1016/j.autcon.2020.103145.

BARPI (2024) 'ARIA Database', *https://www.aria.developpement-durable.gouv.fr/* [Preprint].

Daniele, M., Jennifer Xiaoling, W. and Moretto, P. (2019) 'HIAD 2.0 - hydrogen incidents and accidents database', in *Proceedings of the International Conference on Hydrogen Safety ICHS*.

Devlin, J. *et al.* (2018) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *arXiv preprint arXiv:1810.04805* [Preprint].

Han, J. and Wang, H. (2021) 'Transformer based network for Open Information Extraction', *Engineering Applications of Artificial Intelligence*, 102. Available at: https://doi.org/10.1016/j.engappai.2021.104262.

Huan, N. *et al.* (2024) 'Does accident awareness affect people's risk perception of hydrogen infrastructure and information-seeking behaviour?', *Applied Energy*, 364, p. 123141. Available at: https://doi.org/10.1016/j.apenergy.2024.123141.

Janstrup, K.H. *et al.* (2023) 'Predicting injury-severity for cyclist crashes using natural language processing and neural network modelling', *Safety Science*, 164, p. 106153. Available at: https://doi.org/10.1016/j.ssci.2023.106153.

Jia, Q. *et al.* (2024) 'Enhancing accident cause analysis through text classification and accident causation theory: A case study of coal mine gas explosion accidents', *Process Safety and Environmental Protection*, 185, pp. 989–1002. Available at: https://doi.org/10.1016/j.psep.2024.03.066.

Jones, S., Kirchsteiger, C. and Bjerke, W. (1999) 'The importance of near miss reporting to further improve safety performance', *Journal of Loss Prevention in the Process Industries*, 12(1), pp. 59–67. Available at: https://doi.org/10.1016/S0950-4230(98)00038-2.

Kalyan, K.S., Rajasekharan, A. and Sangeetha, S. (2021) 'AMMUS : A Survey of Transformer-based

Pretrained Models in Natural Language Processing', *arXiv preprint arXiv: arXiv:2108.05542v2*, 2108.05542, pp. 1–42. Available at: http://arxiv.org/abs/2108.05542.

Kuhn, K.D. (2019) 'Using structural topic modeling to identify latent topics and trends in aviation incident reports', *Transportation Research Part C*, 87(December 2017), pp. 105–122. Available at: https://doi.org/10.1016/j.trc.2017.12.018.

Kutela, B., Das, S. and Dadashova, B. (2022) 'Mining patterns of autonomous vehicle crashes involving vulnerable road users to understand the associated factors', *Accident Analysis & Prevention*, 165, p. 106473. Available at: https://doi.org/10.1016/j.aap.2021.106473.

Lu, Y. *et al.* (2024) 'Causative factors and risk prediction model of hydrogen leakage accidents: Machine learning based on case evidence', *International Journal of Hydrogen Energy*, 63, pp. 294–307. Available at: https://doi.org/10.1016/j.ijhydene.2024.03.158.

Macêdo, J.B. *et al.* (2022) 'Identifying low-quality patterns in accidents reports from textual data', *International Journal of Occupational Safety and Ergonomics*, pp. 1–27. Available at: https://doi.org/10.1080/10803548.2022.2111847.

McDonald, A.D., Ade, N. and Peres, S.C. (2020) 'Predicting Procedure Step Performance From Operator and Text Features: A Critical First Step Toward Machine Learning-Driven Procedure Design', *Human Factors* [Preprint]. Available at: https://doi.org/10.1177/0018720820958588.

Nguyen, M.T., Le, D.T. and Le, L. (2021) 'Transformers-based information extraction with limited data for domain-specific business documents', *Engineering Applications of Artificial Intelligence*, 97. Available at: https://doi.org/10.1016/j.engappai.2020.104100.

PNNL (2024) 'H2TOOLS Lessons learned', *https://h2tools.org/lessons?search_api_fulltext=* [Preprint].
Ramos, P.M. *et al.* (2022) 'Combining BERT with numerical variables to classify injury leave based on accident description', *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, p. 1748006X2211401. Available at: https://doi.org/10.1177/1748006X221140194.

Sakamoto, J. *et al.* (2016) 'Leakage-type-based analysis of accidents involving hydrogen fueling stations in Japan and USA', *International Journal of Hydrogen Energy*, 41(46), pp. 21564–21570. Available at: https://doi.org/10.1016/j.ijhydene.2016.08.060.

Wang, C. *et al.* (2024) 'Modeling and performance analysis of emergency response process for hydrogen leakage and explosion accidents', *Journal of Loss Prevention in the Process Industries*, 87, p. 105239. Available at: https://doi.org/10.1016/j.jlp.2023.105239.

Wolf, T. *et al.* (2020) 'Transformers : State-of-the-Art Natural Language Processing', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 38–45. Available at: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yan, P. *et al.* (2021) 'Quantum probability-inspired graph neural network for document representation and classification', *Neurocomputing*, 445, pp. 276–286. Available at: https://doi.org/10.1016/j.neucom.2021.02.060.

Yang, F. *et al.* (2021) 'Review on hydrogen safety issues: Incident statistics, hydrogen diffusion, and detonation process', *International Journal of Hydrogen Energy*, 46(61), pp. 31467–31488. Available at: https://doi.org/10.1016/j.ijhydene.2021.07.005.

Zhang, X. *et al.* (2022) 'Hydrogen Leakage Simulation and Risk Analysis of Hydrogen Fueling Station in China', *Sustainability*, 14(19), p. 12420. Available at: https://doi.org/10.3390/su141912420.

Zhang, X., Srinivasan, P. and Mahadevan, S. (2021) 'Sequential deep learning from NTSB reports for

aviation safety prognosis', *Safety Science*, 142, p. 105390. Available at: https://doi.org/10.1016/j.ssci.2021.105390.