LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

Quantum NLP for Predicting Injury Leaves in Occupational Accident Reports

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

Work accidents can threaten workers' health and impose significant costs on organizations, such as the need for work restructuring and the direct or indirect costs associated with employee absence. Accident investigation reports offer valuable information that can help companies propose preventive measures and identify the causes and consequences of injury events. However, this information is often complex, redundant, or incomplete, making thorough human review challenging due to the large volume of reports. In this study, we applied Quantum Natural Language Processing (QNLP) techniques to predict injury leave based on accident reports from a hydroelectric power company. Our cross-validation results showed a median accuracy of 73.33%. We also analyzed reports with high and low classification accuracy to explore potential reasons for these outcomes. This study demonstrates that accident investigation reports can provide essential insights to enhance workplace safety.

KEYWORDS. Quantum Computing. Quantum Natural Language Processing. Occupational Safety.

Paper topics (indicate in order of PRIORITY the paper topic(s))

1. Introduction

Occupational accidents have significant negative consequences for employees, the environment, and the surrounding community. Factors that influence human error and contribute to accidents are typically documented and managed through accident reports. These reports are crucial for preventing similar incidents in the future, promoting workplace safety, and predicting the outcomes of past events [Swuste et al., 2020].

Investigating injury leaves caused by accidents is essential for evaluating employee unavailability and work time loss. Employee absences can incur substantial financial costs for companies. According to the Brazilian National Institute of Social Security [INSS, 2022], an organization must pay an employee's salary for the first 15 days after an accident, after which the government covers the social security benefit for the remainder of the leave period. However, the company remains responsible for the worker's guarantee fund and may need to hire a temporary replacement. Significant research has been dedicated to analyzing occupational accidents, particularly concerning accident rates and recovery times. Automating the categorization of incident reports can alleviate the burden of manual tasks in safety strategies, leading to more efficient and empirical accident analysis [Koc & Gurgun, 2022; Moura et al., 2016; Sarkar et al., 2018].

Despite the growing use of quantum approaches in various fields, Quantum Natural Language Processing (QNLP) applications in Risk Analysis (RA) are still underexplored, especially in the industrial sector, where risk analysis is crucial. Addressing this gap and building on prior research [Ramos et al., 2022], our study aims to bridge this void. We conduct a exploratory study in order to evaluate the performance of a QNLP model in classifying industrial accidents through binary classification, determining whether an injury leave is expected from a given accident report.

The intersection of quantum computing with operational research in industrial engineering opens new possibilities for optimizing complex decision-making processes and enhancing predictive analytics in workplace safety management. Our approach uses Lambeq's pipeline [Kartsaklis et al., 2021] to convert sentences into quantum circuits. This exploration not only advances academic knowledge but also provides insights into the practical implementation of quantum technologies in enhancing workplace safety and associated costs. We apply our model to a comprehensive occupational accident database from a Brazilian hydroelectrical company, which contains 626 reports describing accidental events over a six-year period.

By automating the classification of accident narratives, our model allows experts to focus on critical aspects without the need to manually read through entire accident narratives. Even with the current limitations, such as hardware constraints, this study lays the groundwork for future applications and improvements in predictive analytics and decision-making processes in industrial settings.

2. Methodology

The database analyzed consists of a spreadsheet where each row corresponds to an accident investigation report and columns capture various characteristics of the event (e.g., location, causes, accident agent, consequences, and financial impact) and details about the employee involved (e.g., job position, experience, training, and injury leave). The dataset includes 626 reports, all in Portuguese, describing accidental events that occurred over a six-year period. This dataset, previously utilized by Macêdo et al. (2022) and Ramos et al. (2022), categorizes accidents into nine different "accident agent" classes based on NBR 14,280. The "accident agent" factor refers to the object, substance, or environment related to the unsafe condition that caused the accident. Each class within this category is associated with specific accident-causing agents.

Our analysis focuses on the class with the most samples, "trip or slip", which includes 196 reports, in a similar way to Ramos et al. (2022). Classes with fewer reports (e.g., 10 reports in the "chemical substance" class) were excluded to avoid increasing the vocabulary list without providing sufficient data for pattern extraction. This study aims to predict whether an occupational

accident led to injury leave. Our approach involves converting sentences into quantum circuits. All steps were implemented in Python using Lambeq, Pandas, Numpy NLTK libraries.

Textual data often contains noise, such as varying capitalizations, punctuation, and special characters. To improve the learning process, we applied three preprocessing operations to clean the input sentences: (i) lowercasing, (ii) stopwords filtering, and (iii) tokenization. These preprocessing steps standardize the raw corpus, facilitating the discovery of patterns and information extraction. Given hardware limitations and the number of available qubits, we randomly selected 30 narratives from the database and used OpenAI Summarization to condense each narrative to 20 words. This summarization process allowed us to extract essential information while addressing constraints. In addition, all narratives were manually translated to English, since the grammatical rules considered for mapping sentences onto the quantum circuit available in the Lambeq library consider the grammatical structure of the English language.

The reports were randomly divided into training and test datasets, with 80% used for training (20% of which was set aside for validation) and 20% used for testing. The QNLP model configuration included the XORSentenceModel inherited from PennyLane (PennyLane, 2023), AerBackend for implementing necessary settings, SPSAOptimizer for parameter optimization, and Quantum Trainer as the trainer. The Ansatz type "IQPAnsatz" was used for this stage (Discopy Documentation). The number of epochs, learning rate and batch size were set to 25, 10^{-2}, and 4, respectively. We performed a 5-fold cross-validation in order to assess the models' performance.

3. Results and Discussion

The QNLP model achieved an average accuracy of 73.33% in just over 10 minutes of training. To assess the model's performance in accident classification tasks, we performed a 5-fold cross-validation and summarized the results in a confusion matrix in Figure 1. The rows of the confusion matrix represent the actual accident classes, while the columns represent the predicted classes. Each cell in the matrix contains the percentage of instances where a particular actual class was predicted as a certain class. For example, the cell at the intersection of the first row and the first column shows the percentage of times instances of the first class were correctly classified as 'Injury Leave' (i.e. the accident led to an injury leave) across all folds. Conversely, the cell at the intersection of the first row and the second column shows the percentage of times instances labeled as 'Injury Leave' were incorrectly classified.

		Predicted Label	
True Label			
		20	80
		0	100

Figure 1: Confusion Matrix from 5-Fold Cross Validation (Percentages)

It is noticeable that the model had difficulties in identifying instances that led to the removal of the injured employee. One of the reasons could be the small dataset, due to the limitation on the number of qubits we could model in the simulator and the imbalance between categories. However, the model achieved results superior to those obtained by Ramos et al. (2022) using classical computing techniques for the same sampling.

4. Conclusion

In this study, we demonstrated the effectiveness of the QNLP model in accident classification tasks, achieving an average accuracy of 73.33% through 5-fold cross-validation. While the model showed promising performance, challenges were observed in accurately identifying instances leading to employee absence, potentially due to dataset limitations and class imbalance. By addressing these avenues for future research, we can further advance the capabilities

of QNLP models in accident classification and contribute to improved occupational safety and risk management strategies.

It's important to note that our study leveraged simulation tools for experimentation. Transitioning to quantum hardware holds promise for accelerated processing and heightened precision. QNLP stands out for its ability to extract semantics while preserving grammar, a distinct advantage over classical NLP. Moreover, it possesses the potential to attenuate biases inherent in large language models through transfer learning. With quantum computing, QNLP models can operate efficiently with minimal data, offering greater control and transparency for more accountable outcomes.

Future research will also focus on enhancing the robustness of the model, validating its performance against human expertise, and testing its efficacy on real quantum hardware. These efforts aim to advance the application of QNLP in risk analysis, thereby improving safety measures in various industries.

References

INSS. (2022). *Auxílio-doença*. https://www.gov.br/inss/pt-br/saiba-mais/auxilios/auxilio-doenca

Kartsaklis, D., Fan, I., Yeung, R., Pearson, A., Lorenz, R., Toumi, A., de Felice, G., Meichanetzidis, K., Clark, S., & Coecke, B. (2021). lambeq: An Efficient High-Level Python Library for Quantum NLP. *ArXiv Preprint*, *arXiv:2110*, 1–19. http://arxiv.org/abs/2110.04236

Koc, K., & Gurgun, A. P. (2022). Scenario-based automated data preprocessing to predict severity of construction accidents. *Automation in Construction*, *140*(May), 104351. https://doi.org/10.1016/j.autcon.2022.104351

Macêdo, J. B., Ramos, P., Maior, C. B. S., Moura, M. D. C., Lins, I. D., & Vilela, R. (2022). Identifying low-quality patterns in accidents reports from textual data. *International Journal of Occupational Safety and Ergonomics*, 1–27. https://doi.org/10.1080/10803548.2022.2111847

Moura, M. D. C., Azevedo, R. V., Droguett, E. L., Chaves, L. R., Lins, I. D., Vilela, R. F., & Filho, R. S. (2016). Estimation of expected number of accidents and workforce unavailability through Bayesian population variability analysis and Markov-based model. *Reliability Engineering and System Safety*, *150*, 136–146. https://doi.org/10.1016/j.ress.2016.01.017

NTSB. (2022). *Accident Data*. Download the Census of US Civil Aviation Accidents. https://data.ntsb.gov/avdata

Ramos, P. M., Macedo, J. B., Maior, C. B., Moura, M. C., & Lins, I. D. (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*, 1748006X2211401. https://doi.org/10.1177/1748006X221140194

Sarkar, S., Verma, A., & Maiti, J. (2018). Prediction of Occupational Incidents Using Proactive and Reactive Data : A Data Mining Approach. In Springer (Ed.), *Industrial Safety Management* (pp. 65–79).

Swuste, P., Groeneweg, J., van Gulijk, C., Zwaard, W., Lemkowitz, S., & Oostendorp, Y. (2020). The future of safety science. *Safety Science*, *125*(November 2019), 104593. https://doi.org/10.1016/j.ssci.2019.104593.

PennyLane. (2023, April 26). Quantum NLP with the Lambeq-PennyLane Integration. PennyLane Blog. Retrieved from https://www.pennylane.ai/blog/2023/04/quantum-nlp-with-the-lambeq-pennylane-integration/

Discopy Documentation. (n.d.). IQPAnsatz. Discopy Library Documentation. Retrieved from https://docs.discopy.org/en/main/_api/discopy.quantum.ansatze.IQPansatz.html