Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)
Edited by Maria Chiara Leva, Edoardo Patelli, Luca Podofillini, and Simon Wilson
©2022 ESREL2022 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-5183-4_S29-02-199-cd

Identification of features of rare risk events in oil refineries using Natural Language Processing (NLP)

July Bias Macêdo[a], Márcio das Chagas Moura[a], Isis Didier Lins[a], Enrico Zio[b,c]

[a]*Center for Risk Analysis and Environmental Modeling, Department of Industrial Engineering, Universidade Federal de Pernambuco*
[b]*MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France*
[c]*Energy Department, Politecnico di Milano, Milan, Italy*
E-mail: julybias@gmail.com, marcio@ceerma.org, isis.lins@ceerma.org, enrico.zio@polimi.it

Accidents in the process industry can be prevented and their consequences mitigated by performing proper risk analyses to inform decision making. Quantitative risk analysis (QRA) is one of the main frameworks used for risk assessment and the identification of all hazards, which a plant is exposed to. There are crucial tasks to ensure the comprehensiveness of QRA that are carried out by the examination of a set of engineering documents by experts during structured meetings. The hazards identified are stored as textual data in documents, which retain valuable information about the risks related to the analyzed system. In this context, Natural Language Processing (NLP) techniques have emerged as a way for extracting, organizing and classifying relevant information from text. However, a challenge arises when we are interested in addressing catastrophic accidents that are rare and for which, thus, only limited information is available. Some accidents are actually postulated as possible in principle, but there are no historical occurrence records. Developing techniques to characterize features about such rare events is a challenging task, yet quite useful for QRA, whose outcomes would guide designing preventive measures and supporting decision making. In this paper, we applied data augmentation (DA) and investigate different configurations to address the rare event issue. DA is applied to obtain a balanced and sufficiently large training set. The final aim of this work is developing a model capable of characterizing relevant features about rare or unseen accidental scenarios in support to hazard and operability study (HAZOP) and preliminary hazard analysis (PrHA).

Keywords: Risk Analysis, text mining, natural language processing, oil refineries, rare accidents, few-shot learning, data augmentation.

1. Introduction

Oil and gas industry is characterized by the presence of flammable, explosive, and toxic materials. The loss of containment of such materials may cause catastrophic accidents with substantial human, environment and financial losses (Pramoth, Sudha, and Kalaiselvam 2020; Soomro et al. 2022).

In this context, Quantitative Risk Analysis (QRA) has been widely used to systematically investigate the risks related to industrial activities. QRA can even be considered mandatory by safety regulators in order to provide license to operate as it provides a thorough picture of the hazards and then enables the organizations to wisely manage risks (Vinnem and Røed 2020; Ramos et al. 2020; Steijn et al. 2020).

The process to perform QRA can be divided into 5 steps: 1) system delimitation and description; 2) identification and assessment of potential hazards associated with the operation of the analyzed system; 3) quantitative analysis of the accident scenarios; 4) evaluation of risks; 5) record and report of the results. Step 2 is a key aspect of the quality of QRA, and a great effort must be made to ensure its completeness. However, the methods adopted to perform this step, such as Preliminary Hazards Analysis (PrHA), strongly rely on experts' knowledge. Hazards identified in the early stages are stored as textual data in documents, which retain valuable information about the risks related to the analyzed system.

Natural Language Processing (NLP) provides helpful techniques for extracting, organizing, and classifying relevant information from text. Indeed, a lot of attention has been put on these techniques in the risk analysis context (Ansaldi et al. 2020; Gulijk and Holmes 2020; Maior et al. 2020; Macedo et al. 2021; J. Macêdo et al. 2020).

For instance, (Srinivasan, Nagarajan, and Mahadevan 2019) adopted NLP techniques to develop a classification system, where the expected output of the system is intended to give information if the accident is likely to happen. (Valcamonico et al. 2020) proposed an approach for modelling accident descriptions using topic modelling and, then, used the resulting representation as input to classify the accident reports. (Ahadh, Binish, and Srinivasan 2021) applied keyword extraction and topic modelling to support the analysis of accident reports. The authors aimed at reducing manual intervention in accident reports in order to extract useful knowledge that can be used to perform classification tasks such as identification root causes of accidents. (J. B. Macêdo et al. 2022) extracted textual data from preliminary hazard analysis reports regarding different systems of an oil refinery. The extracted data were used to train classifiers to predict the possible consequences given the occurrence of a spill and their respective expected frequency and severity level.

However, step 2 may be quite challenging when dealing with rare accident events with extreme consequences, since limited knowledge exists for these events (Zio and Aven 2012; Spada, Burgherr, and Hohl 2019). On the other hand, developing techniques to identify features about catastrophic events is quite useful for QRA, whose outcomes would guide designing preventive measures and supporting the related decision making.

In this paper, we investigate how to handle accident datasets to improve BERT-based classifier learning about rare event. To do that we compared different data augmentation (DA) configurations to obtain a balanced and sufficiently large training set. The final aim of this work is developing a model capable of characterizing relevant features about rare or unseen accidental scenarios in support to hazard and operability study (HAZOP) and preliminary hazard analysis (PrHA).

This paper is organized as follows. Section 2 briefly introduces NLP, the current scenario and challenges in this field. Section 3 explains the methods adopted and the data used to fine-tune bidirectional encoder representations from transformers (BERT). Section 4 presents the results and discussion for each configuration, and finally Section 5 concludes remarks.

2. Natural Language Processing

NLP investigates how computers can be used to understand and manipulate texts (Chowdhury 2003). There are numerous applications of NLP that can be used in text mining process, such as topic tracking, clustering, text classification, and so on (Zhang 2012).

Most of the information of a company is stored as text which contains valuable information about the industrial system (Bhardwaj and Khosla 2017). Machine learning techniques allow to build models on body of texts with specific context. Due to the predictability of texts, these language models can reveal mean from text data (Bengfort, Bilbro, and Ojeda 2018).

Pre-trained recurrent or transformer language models have been directly fine-tuned, entirely removing the need for task-specific architectures. In addition, the substantial increase in the transformer language models, from 100 million to 17 billion parameters, has brought improvements in several downstream NLP tasks. However, to achieve strong performance on a desired task typically requires fine-tuning on a dataset of thousands to hundreds of thousands of examples specific to that task (Brown et al. 2020). Thus, a challenge arises when we are interested in rare events, since only limited information is available. For instance, some accidents are postulated as possible in principle, but there are no historical occurrence records.

3. Methods

Here a label dataset, built with information extracted from previous PrHA, is used fine-tuned BERT model (Devlin et al. 2018) to classify the frequency of potential accident scenarios. DA is applied to obtain a balanced training set and different configurations are adopted to investigate the most suitable approach to handle rare risk events.

3.1. Dataset

As described in (J. B. Macêdo et al. 2022; J. Macêdo et al. 2020), each document contains the description of potential accident events from different processing units of an oil refinery and their qualitative assessment of frequency of occurrence and severity of consequences. The events are classified into four categories (Tab. 1) according to (ISO 2018). Note that the categories are defined in terms of the damage to human life.

Tab. 1. Description of the categories of the likelihood of occurrence.

Likelihood of occurrence		
Category		Description
A	Remote	conceptually possible, but there are no records in the literature
B	Unlikely	unlikely to occur in normal conditions
C	Possible	might occur sometime
D	Likely	will probably occur

As one can see in Fig. 1, the labelled dataset is quite unbalanced. There are less than 300 potential accidents classified as Remote, representing under 15% of the dataset.

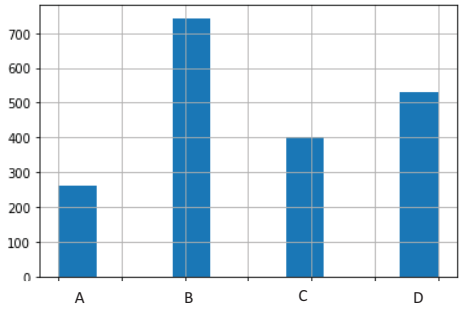

Fig. 1. Number of instances per likelihood category.

As pointed by (Brown et al. 2020), the difficult to build labelled datasets may limit the applicability of NLP to perform text classification tasks. Indeed, this dataset was used in (J. Macêdo et al. 2020; J. B. Macêdo et al. 2022) to build a classifier able to predict the expected frequency, and as pointed in these previous works the worst performance of the NLP classifiers was to categorize remote events (Tab. 1).

3.2. Data Augmentation (DA)

DA is a useful tool to improve the performance of a model; however, in NLP datasets, DA is difficult due to the high complexity of language (Wei and Zou 2020). Thus, we explored different DA configurations to balance our dataset in order to improve the performance of the NLP model in classifying remote events.

The approach adopted consists in using contextualized word embeddings to generate a vector under a different context. DA was implemented using *nlpaug*, a library dedicated to textual augmentation in ML experiments (Ma 2021). More specifically, we used *ContextualWordEmbsAug* function that is designed to perform insertion and substitution. Simply put, the surrounding words are provided to BERT model to find out the most suitable word for augmentation and used as a feature to predict the target word.

In this paper, under sampling and over sampling strategies are combined; we performed DA on the training data (80% of the dataset) for the classes with less samples (A and C) in order to balance the number of samples. In addition, we assumed that categories B and A have characteristics in common, which may hinder the model's ability to separate these events. Thus, we removed training samples from the category with more samples (B), since we are interest in extract features of rare events that presents few samples available. The following configurations were adopted:

(i) Generate 200 training samples labelled as A
(ii) Generate 200 training samples labelled as A and 100 labelled as C
(iii) Under sampling B to reduce to 399 samples
(iv) Removing all instances labelled with B
(v) Generate 100 training samples labelled as A and remove all instances labelled with B

As mentioned, these configurations were selected to balance the dataset in order to find the best approach to extract features about rare events. To that end, the training sets obtained according to each configuration are used to fine-tune BERT model resulting on different classifiers.

The impact of each configuration on the classifiers' performance is evaluated by comparing the results on the test set to the baseline result, which is brought in Tab. 2 as the performance on test data of a classifier obtained by fine-tuning of BERT using the original training set.

Tab. 2. Baseline results; performance on test data of the classifier using the original training set.

Likelihood of occurrence			
Class	**Recall**	**Precision**	**F_1-score**
A	70.67	81.54	75.72
B	88.89	86.02	87.43
C	90.10	79.82	84.95
D	85.04	91.53	88.19

Moreover, we adopted the Pytorch implementation of BERT available on *Hugging Face* library (Wolf et al. 2020). All classifiers were trained for 25 epochs, with batch size of 16 and learning rate of 5×10^{-5}. The results and discussion are presented in the next section.

4. Results and Discussion

The performance on test data of the classifier obtained using configuration (i) (200 augmented samples labelled as A) is presented in Tab. 3.

Tab. 3. Performance of the classifier using the training set obtained with configuration (i).

Likelihood of occurrence			
Class	**Recall**	**Precision**	**F_1-score**
A	67.53	75.36	71.23
B	88.32	85.25	86.76
C	82.76	88.89	85.72
D	95.08	89.92	92.43

Overall, comparing Tab. 2 to Tab. 3, the performance improved considering categories C and D. There was a 10% increase on the recall for class D and the precision for class C. However, as one can see, the performances for category A and B decreased. This result may indicate that the description of accident scenarios related to A and B are similar. Thus, simply augmenting the data related to remote events is not enough to tackle this issue, because it makes the classifier label scenarios B as A.

Tab. 4 shows the classifier's performance on test data with configuration (ii). One interesting result was the increase of the recall metric regarding classes B and D. This result may indicate that configuration (ii) improved the features extracted for accident scenarios B and D, since the number of instances from these classes that were misclassified decreased.

On the other hand, although there was a significant improvement on the precision metric for class A, there was also a decrease on the recall. This is particularly undesirable since we want a model to extract valuable features about such rare events. With configuration (ii) the extracted features were worse than the baseline classifier to represent class A since more accident scenarios were misclassified into a different class.

Tab. 4. Performance of the classifier using the training set obtained with configuration (ii).

Likelihood of occurrence			
Class	**Recall**	**Precision**	**F_1-score**
A	60.26	95.92	74.02
B	94.89	83.5	88.83
C	83.91	76.84	80.22
D	89.44	91.37	90.39

In addition, we randomly removed some instances belonging to class B. Tab. 5 shows the classifier's performance on test data with configuration (iii). After under sampling class B, most metrics worsened, as expected since we are reducing the number of samples used to train the model. One can see that the model begun to classify more instances as A and consequently the recall for class A increased and the precision decreased. There was also an improvement regarding the recall for class D, because the classifier did not misclassify accidents from class D as B.

Tab. 5. Performance of the classifier using the training set obtained with configuration (iii).

Likelihood of occurrence			
Class	**Recall**	**Precision**	**F_1-score**
A	73.75	56.19	63.78
B	68.98	84.87	76.10
C	82.42	77.32	79.79
D	92.68	88.37	91.00

Configuration (iv) was adopted considering that category B must have many characteristics in common with category A, since class B also represents rare (unlikely) events. Indeed, the results obtained with configurations (i), (ii) and (iii) showed the classifiers often have difficulties differentiating these classes. Tab. 6 shows the classifier's performance on test data with configuration (iv).

Tab. 6. Performance of the classifier using the training set obtained with configuration (iv).

Likelihood of occurrence			
Class	**Recall**	**Precision**	**F_1-score**
A	93.06	97.10	95.04
C	77.23	89.66	82.98
D	92.13	81.25	86.35

Indeed, after training the model only with common accident events (class C and D) and class A (remote events), there was a significant improvement in the performance of the model regarding A. Also, it is possible to see a slight reduction in some metrics related to the classification of C and D, due to the removal of a large number of samples; considering the F_1-score for C and D the performance decreased 2%. In addition to the removal of the samples labelled as B we generated 100 training samples for class A (configuration v). Tab. 7 shows the classifier's performance on test data with configuration (v).

Tab. 7. Performance of the classifier using the training set obtained with configuration (v).

Likelihood of occurrence			
Class	**Recall**	**Precision**	**F_1-score**
A	100	100	100
C	86.35	87.65	86.99
D	93.10	90.00	91.52

After removing the samples that has common characteristics as the Remote events and performing DA, we were able to correctly predict all instances labelled as A of this specific test set. Thus, it seems promising for dealing with rare scenarios that usually are underrepresented in accident and reliability databases.

5. Conclusion

The early stages of QRA involves identifying and assess accident events. Developing techniques to identify features about catastrophic events with extreme consequences is quite useful for QRA. There is a growing trend in NLP applications in risk analysis context; however, it may be quite challenging to develop good models when dealing with rare accident events, since limited knowledge exists for them. Moreover, in NLP datasets, DA is difficult due to language complexity. Thus, we explored different DA configurations to balance our dataset in order to improve the performance of the NLP model in classifying remote events.

The results showed that simply augmenting the training data for the remote accident class is not enough to improve the model's learning for this class. However, fine-tuning only with samples of frequent and likely events (C and D) significantly improved the performance of the model. Thus, combining this approach and performing DA for class A, configuration (v), resulted in a model capable to correctly predict all remote events, which is a promising result.

One possible alternative to DA is to use Few-Shot Learning (Brown et al. 2020); thus, it will be object of future research. In addition, in future works we intend to evaluate whether, through the features extracted through fine-tuning using configuration (v), it is possible to identify other characteristics, such as the severity of the consequences of these rare events.

Acknowledgement

The authors thank the Nacional Agency for Research (CNPq-Brazil) and the Foundation of Support for Science and Technology of Pernambuco (FACEPE) for the financial support through research grants. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Ahadh, Abdhul, Govind Vallabhasseri Binish, and Rajagopalan Srinivasan. 2021. "Text Mining of Accident Reports Using Semi-Supervised Keyword Extraction and Topic Modeling." *Process Safety and Environmental Protection* 155: 455–65. https://doi.org/10.1016/j.psep.2021.09.022.

Ansaldi, Silvia M., Carla Simeoni, Alessandro Di Francesco, Roberto Martini, Luca Di Piramo, and Flavia Fattori. 2020. "Extracting Knowledge from Near Miss Reports Using Machine-Learning Techniques." In *30th European Safety and Reliability Conference*

and the 15th Probabilistic Safety Assessment and Management Conference. https://doi.org/10.3850/981-973-0000-00-0 esrel2020psam15-paper.

Bengfort, Benjamin, Rebecca Bilbro, and Tony Ojeda. 2018. *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. O'Reilly Media, Inc.

Bhardwaj, Priya, and Priyanka Khosla. 2017. "National Conference on Emerging Trends in Information Technology- Advances in High Performance Computing , Data Sciences & Cyber Security." *IITM Journal of Management and IT* 8 (1): 27–31.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Chowdhury, Gobinda G. 2003. "Natural Language Processing." *Annual Review of Information Science and Technology* 37 (1): 51–89.

Gulijk, Coen van, and Violeta Holmes. 2020. "Verification of Safety Rules Using NLP." In *30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*. https://doi.org/10.3850/981-973-0000-00-0 esrel2020psam15-paper.

Ma, Edward. 2021. "Nlpaug Documentation."

Macêdo, July, Diego Aichele, Márcio das Chagas Moura, and Isis Lins. 2020. "A Text Mining and NLP Approach for Identifying Potential Consequences of Accidents in an Oil Refinery." In *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*. Singapore: Research Publishing. https://doi.org/10.3850/978-981-14-8593-0_4527-cd.

Macedo, July, Diego Aichele, Marcio das Chagas Moura, and Isis Didier Lins. 2021. "A Web App to Support Hazard Identification of Oil Refineries." In *31st European Safety and Reliability Conference*. Angers, France.

Macêdo, July Bias, Márcio das Chagas Moura, Diego Aichele, and Isis Didier Lins. 2022. "Identification of Risk Features Using Text Mining and BERT-Based Models: Application to an Oil Refinery." *Process Safety and Environmental Protection* 158 (February): 382–99. https://doi.org/10.1016/j.psep.2021.12.025.

Maior, Caio Bezerra Souto, João Mateus Marques de Santana, Márcio das Chagas Moura, and Isis Didier Lins. 2020. "Automated Classification of Injury Leave Based on Accident Description and Natural Language Processing." In *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, 1276–81. https://doi.org/10.3850/978-981-14-8593-0_4559-cd.

Pramoth, R, S Sudha, and S Kalaiselvam. 2020. "Resilience-Based Integrated Process System Hazard Analysis (RIPSHA) Approach: Application to a Chemical Storage Area in an Edible Oil Refinery." *Process Safety and Environmental Protection* 141: 246–58. https://doi.org/10.1016/j.psep.2020.05.028.

Ramos, M.A., E. López Droguett, A. Mosleh, and M. Das Chagas Moura. 2020. "A Human Reliability Analysis Methodology for Oil Refineries and Petrochemical Plants Operation: Phoenix-PRO Qualitative Framework." *Reliability Engineering and System Safety* 193 (September 2019): 106672. https://doi.org/10.1016/j.ress.2019.106672.

Soomro, Afzal Ahmed, Ainul Akmar Mokhtar, Jundika Chandra Kurnia, Najeebullah Lashari, Huimin Lu, and Chico Sambo. 2022. "Integrity Assessment of Corroded Oil and Gas Pipelines Using Machine Learning: A Systematic Review." *Engineering Failure Analysis* 131 (July 2021): 105810. https://doi.org/10.1016/j.engfailanal.2021.105810.

Spada, Matteo, Peter Burgherr, and Markus Hohl. 2019. "Toward the Validation of a National Risk Assessment against Historical Observations Using a Bayesian Approach : Application to the Swiss Case Toward the Validation of a National Risk Assessment against Historical Observations Using a Bayesian Approach : A." *Journal of Risk Research* 22 (11): 1323–42. https://doi.org/10.1080/13669877.2018.1459794.

Steijn, W. M.P., J. N. Van Kampen, D. Van der Beek, J. Groeneweg, and P. H.A.J.M. Van Gelder. 2020. "An Integration of Human Factors into Quantitative Risk Analysis Using Bayesian Belief Networks towards Developing a 'QRA+.'" *Safety Science* 122 (September 2019): 104514. https://doi.org/10.1016/j.ssci.2019.104514.

Valcamonico, Dario, Piero Baraldi, Francesco Amigoni, and Enrico Zio. 2020. "Text Mining for the Automatic Classification of Road Accident Reports." In *30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*. https://doi.org/10.3850/981-973-0000-00-0 esrel2020psam15-paper.

Vinnem, Jan-erik, and Willy Røed. 2020. *Offshore Risk Assessment*. 4th ed. Vol. 1. Springer, London. https://doi.org/10.1007/978-1-4471-7444-8.

Wei, Jason, and Kai Zou. 2020. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks." In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/d19-1670.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, 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*, 38–45. Association for Computational Linguistics.

Zhang, Muyu. 2012. "Encoding World Knowledge in the Evaluation of Local Coherence."

Zio, Enrico, and Terje Aven. 2012. "Industrial Disasters : Extreme Events , Extremely Rare . Some Reflections on the Treatment of Uncertainties in the Assessment of the Associated Risks." *Process Safety and Environmental Protection* 1 (January): 31–45. https://doi.org/10.1016/j.psep.2012.01.004.