Identification of risk features using text mining and BERT-based models: Application to an oil refinery

July Macedo[1], Diego Aichele[2], Márcio das Chagas Moura[3], Isis Lins[3]

ABSTRACT

Oil and gas refineries involve handling and storing hazardous materials, and the uncontrollable release of these substances may lead to catastrophic accidents. In this context, risk studies are aimed at recommending either preventive measures to avoid the undesired events or designing safeguards to mitigate the consequences in case an accident happens. To that end, risk experts postulate possible leakages, then identify their causes and consequences, and finally evaluate and classify the risks into categories. These analyses rely on examination of a variety of engineering textual documents and attendance to numerous meetings, which is very labor, time consuming, especially for oil refineries that are equipment intensive facilities. Moreover, this qualitative process usually characterizes the first steps to perform a quantitative risk analysis (QRA) and is crucial to ensure its quality. Therefore, we here propose to use text mining and fine-tuned trained bidirectional encoder representations from transformers (BERT) models for reducing some tasks, which are involved in the early stages of QRA. These techniques can be applied to extract, organise, and classify information from past textual risk studies, and then allow for recognizing patterns. Our idea is to identify the potential consequences of accidents related to the operation of an oil refinery and classify each scenario in terms of severity of the consequence and likelihood of occurrence. Thus, we expect to reduce the efforts required for completing the early stages of a QRA. The proposed method was applied to an actual oil refinery and presented very promising results. The models resulting from this research were embedded into an app, HALO (hazard analysis based on language processing for oil refineries), which is available online.

1. INTRODUCTION

Quantitative risk analysis (QRA) is one of the main tools used to manage risks in oil refineries. Overall, in the first steps of QRA (hazard identification and analysis), experts recognize relevant scenarios that may arise, assessing and reporting their likelihood and potential consequences [1]. Qualitative approaches are adopted by a multidisciplinary team of experts on design, operation, and maintenance of the plant as tools to complete these steps such as preliminary hazard analysis (PHA).

Generally, these techniques involve examining different engineering documents that describe the installation and attending numerous meetings to postulate possible leakages, identify hazards and their possible causes and consequences and, finally, evaluate and classify risks [2]. Therefore, this paper proposes the application of text mining (TM) and natural language processing (NLP) to support the initial qualitative steps of a QRA. Indeed, TM and NLP techniques can be applied to extract, organize, and classify information from text, allowing the automatic identification of patterns [3]. For this reason, the application of these techniques seems attractive for the QRA context to reduce necessary efforts to perform it.

This paper applies TM to extract information from text data and fine-tunes pre-trained bidirectional encoder representations from transformers (BERT) [4] to identify risk features in an oil refinery. Each dataset used to fine-tune the models was built based on PHA documents, which contain valuable textual information, and were previously conducted for an actual oil refinery.

We expect that with a model trained based on all the available information garnered from past risk studies, experts could use that entire source of knowledge to reduce uncertainty. Instead of starting the analysis from scratch, risk analysts could reuse knowledge, imbued in the trained models, from previous studies or use QRA performed for similar plants as a starting point to identify potential consequences, qualitatively characterize frequency and severity of accidental scenarios, and prioritize the most critical events.

The remainder of this paper is organized as follows. Section 2 describes the proposed methodology to predict the potential accidents, discusses the text data used, the preprocessing step and the modeling process. Section

1 MS, PhD Candidate -Center for Risk Analysis and Environmental Modeling, Federal University of Pernambuco, Brazil.
2 Mechanical Engineer - Center for Risk Analysis and Environmental Modeling, Federal University of Pernambuco, Brazil.
3 PhD, Professor - Center for Risk Analysis and Environmental Modeling, Federal University of Pernambuco, Brazil.

3 shows an application of the proposed model to an oil refinery. Finally, Section 4 concludes remarks and points out future research directions.

2. METHODOLOGY

This paper aims at reducing efforts required to develop risk studies. To that end, we adopted TM techniques to extract text data from PHA documents, and then perform text classification tasks to identify risk features in oil refinery's subsystems. Our idea is to develop models capable of learning and recognizing risk features, and thus extract useful knowledge about accidental scenarios.

We here developed three models by fine-tuning pre-trained BERT model with the extracted data to perform three tasks: i) identification of possible consequences, given an occurrence of a leakage; ii) classification of the severity of the consequences; iii) classification of the likelihood of occurrence of the accidental scenario. Each model was trained with a specific annotated corpus that was built from the PHA sheets. Indeed, the corpus contains the data extracted from PHA documents and the target related to its corresponding task.

Fig 1 provides an overview of the proposed methodology. First, we developed two scripts: one that automatically extracts text from a collection of PHA spreadsheets, and another to organize and build an annotated corpus for each supervised-learning task, also referred to as dataset in this paper. Next, the corpus was preprocessed and converted into a manageable format for feeding the learning algorithms. Below, we describe these steps in more details.

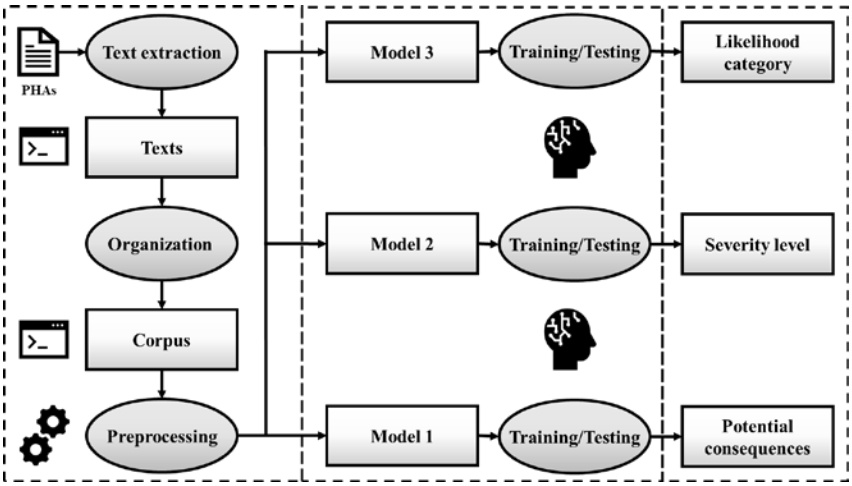

Fig 1 - General overview of the proposed methodology.

As Fig 1 depicts, the target depends on the task performed. For this reason, an annotated corpus (i.e., labelled dataset) was built for each of the three tasks. During the development of the PHA, a set of potential consequences was specified for two initiating events (leakage and rupture) and, thus, we can build a sentence for each set. It is also possible to construct different input sentences using the potential consequences, where the output pair can be either the severity or the likelihood category. Thus, we built corpus 1 with 1,391 instances, i.e., input sentence and label pairs, for Model 1 and dataset 2 and dataset 3 with 2,974 instances for Models 2 and 3 respectively.

Each input sequence provided to Model 1, $x_{1,i}$, characterizes a possible leakage in a specific subsystem of the oil refinery. For the first task, we defined our input as a sentence constructed by joining the following text: unit, system, subsystem description, chemical product, operating conditions, equipment, equipment specifications, and initiating event. The output $y_{1,i}$ is a vector that contains 7 positions, which represent the number of all potential consequences found in the documents. Thus, each position $y_{1,i}^{n}$, $n = 1, 2, 3, 4, 5, 6,$ or 7 (burn injury, vapour cloud explosion, flash fire, irritation, pool fire, toxic vapour cloud, or jet fire respectively), assumes the values 0 or 1, where 1 indicates the presence of the potential consequence n.

For the second and third tasks, the input sentences were similarly constructed by joining the same textual data used for the first task with the addition of the potential consequence; thus, both models are fed with the same input vector $x_{2,i} = x_{3,i}$. The output for the second task may, $y_{2,i}$, be assigned to four possible values (0, 1, 2, or 3), which represent the severity categories (I to IV). Also, the output for the third task, $y_{3,i}$, may be assigned to four possible values (0,1, 2, or 3), which represent the likelihood category (A to D).

Next, three preprocessing operations were performed to transform the input sentences into a cleaner format that can help improve the learning process of the models. The lowercasing and noise removal were implemented in Python using regular expression operations and Pandas library [5] and the tokenization was performed using the tokenizer provided by transformers library [6].

Moreover, we used the Pytorch implementation of pre-trained BERT available at transformers library [6] added one output layer on top of the pre-trained model to adapt it for performing a classification task. Thus, we fine-tuned the pre-trained model three times, using the specific dataset from a given classification task. Each dataset was randomly split into 90% for training and the remaining 10% for test. Finally, we evaluated the model's performance on test data. The results achieved with each model are discussed in the following section.

3. DISCUSSION

Fig 2 provides a confusion matrix with the prediction on test data for each consequence to evaluate the performance of Model 1. One can see that Model 1 was able to predict accurately most of the potential consequences, even those with the lowest frequency. For instance, there are only eleven instances in test data labelled with irritation and the model correctly classified all instances. Model 1 achieved a mean accuracy of 97.42% to predict the potential consequences of test samples and presented satisfactory results considering all potential consequences and achieved a mean F_1-score above 94.09%.

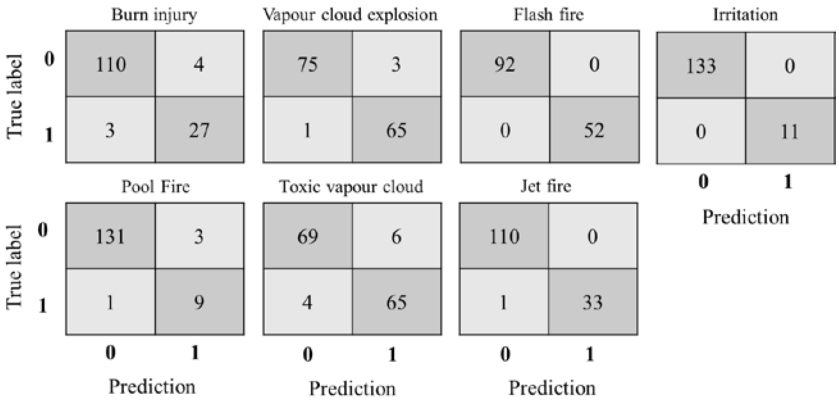

Fig 2 - Confusion matrices for Model 1's classification of test data.

Fig 3a shows the confusion matrix with the Model 2's classification of the test data. The model achieved an accuracy of 86.44% to classify the test data. The results obtained were satisfactory. Indeed, F_1-scores were above 80% for all likelihood categories. Fig 3b shows the confusion matrix with the Model 3's classification of the test data. Model 3 predicted all classes with great precision. Indeed, all performance metrics for category B, C, and D were above 90%. These results of Model 3 suggest a reasonable ability to learn and recognize patterns about all likelihood categories.

Fig 3 - Confusion matrix (a) for Model 2's and (b) for Model 3's classifications of test data

4. CONCLUSION

Overall, this study strengthens the idea that information contained as text data can be automatically extracted and processed by text mining and NLP techniques to support risk studies. The results obtained underscore that TM and NLP can be adopted to support risk analysts in identifying the potential consequences of different scenarios and to describe qualitatively risks in terms of expected likelihood and severity of consequences. Indeed, the proposed method could be a useful tool to support hazard identification and analysis; instead of starting the QRA from scratch, analysts could either reuse knowledge from previous studies or process studies for similar plants. This may be rather useful especially for plants, which are brand new and depend on the approval of the environmental regulators to start the development of the facility design and construction. Then, experts may use that entire source of knowledge to reduce the uncertainty for performing risk analysis based on a model trained with all the available information collected and processed from past risk studies.

5. ACKNOWLEDGEMENTS

The authors thank the National Agency for Research – Brazil (CNPq) 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 - Brazil - Finance Code 001

6. REFERENCES:

[1] W. M. P. Steijn, J. N. Van Kampen, D. Van der Beek, J. Groeneweg, and P. H. A. J. M. Van Gelder, "An integration of human factors into quantitative risk analysis using Bayesian Belief Networks towards developing a 'QRA+,'" *Saf. Sci.*, vol. 122, no. September 2019, p. 104514, 2020, doi: 10.1016/j.ssci.2019.104514.

[2] J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," *Nat. Phys.*, vol. 13, no. 5, pp. 431–434, 2017, doi: 10.1038/nphys4035.

[3] B. Drury and M. Roche, "A survey of the applications of text mining for agriculture," *Comput. Electron. Agric.*, vol. 163, no. February, p. 104864, 2019, doi: 10.1016/j.compag.2019.104864.

[4] J. Devlin, M. Chang, L. Kenton, and T. Kristina, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv Prepr. arXiv1810.04805*, 2018.

[5] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, 2010, doi: 10.25080/majora-92bf1922-00a.

[6] T. Wolf *et al.*, "Transformers : State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.