ASSESSING LOW-QUALITY ACCIDENT REPORTS USING NATURAL LANGUAGE PROCESSING

July Macêdo[a,b], Plínio Ramos[a,b], Caio Maior[a,b], Márcio Moura[a,b], Isis Lins[a,b] and Rômulo Vilela[c],

[a]CEERMA - Center for Risk Analysis, Reliability and Environmental Modeling, Universidade Federal de Pernambuco, Brazil
[b]Department of Production Engineering, Universidade Federal de Pernambuco, Brazil
[c]Companhia Hidrelétrica do São Francisco (CHESF), Brazil

ABSTRACT

Accident investigation reports provide useful knowledge to support companies to propose preventive and mitigative measures. However, the information presented in accident reports databases is normally large, complex, filled out with errors, missing and/or redundant data. In this paper, we propose text mining and natural language processing techniques to investigate low-quality accident reports. We adopted machine learning (ML) to perform exploratory analysis and multi-classification task to detect and investigate inconsistencies on accident reports. The methodology was applied on 626 documents collected from an actual hydroelectric power company and focused on the accident agent categories. The initial ML performances indicated data divergences and concerns related to the report structure. Afterwards, we manually curated and restructured the accident database more properly achieving 73% the ML performances and confirming the initial supposition about the quality of the reports investigated. The proposed approach can be used as a diagnostic tool to improve the design of accident investigation reports to provide a more useful source of knowledge to support informed decisions in the safety context.

1. INTRODUCTION

Work accidents can lead not only to huge financial losses for the organization but also to serious threats to people's integrity and environment. In contrast, these accidents are useful sources of evidence to extract valuable factors that contributed to the occurrence of the event [1]. Thus, systematic accident investigation reports retain knowledge that can be explored to support decision-making. Typically, these reports are written in natural language [2]. Free, textual responses allow describing the event as one perceived it. However, reports' low quality and lack of detail may limit their usefulness because reasonable resources are required for manual analysis, which is a complex and error-prone task [3].

In this context, automatic mining patterns from massive amount of textual data is attractive as the text mining (TM) and natural language processing (NLP) techniques aim to understand, process, and interpret human language allowing to train intelligent models [4], which provides a rapid and trustworthy analysis of large, textual databases comprised of accident investigation reports [5].

Usually, the studies investigate the model's performance to automatically identify patterns and classify causes once this is paramount to accident investigation [3], [4]. However, real investigation reports oftentimes present drawbacks related to the quality of information filled out. In practice, these issues are a significant hurdle for analysts to extract useful insights, and then propose effective preventive measures to improve safety.

Hence, we used a 6-year historical database in a real hydroelectric power company, to investigate and discuss the current characterization of the accident investigation reports, which were structured based on the Brazilian

Standard ABNT NBR 14280 - Workplace accident record [6]. We aim to use TM techniques to recognize, address, and point out possible inconsistencies in the accident investigation reports.

2. DESCRIPTION

A schematic overview of our proposed methodology to extract knowledge from text is shown in Fig. 1. The main idea is to support the diagnosis of the quality content and understanding of raw texts of accident investigation reports using NLP methods, such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Doc2Vec, and Latent Dirichlet Attribution (LDA), and ML classifiers, such as Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). Here, the database is composed of written reports about 626 previous accidental events that occurred in a real company. This can allow experts to obtain meaningful reports that provide valuable and well-structured information. Although we processed texts in Portuguese, we followed generic steps.

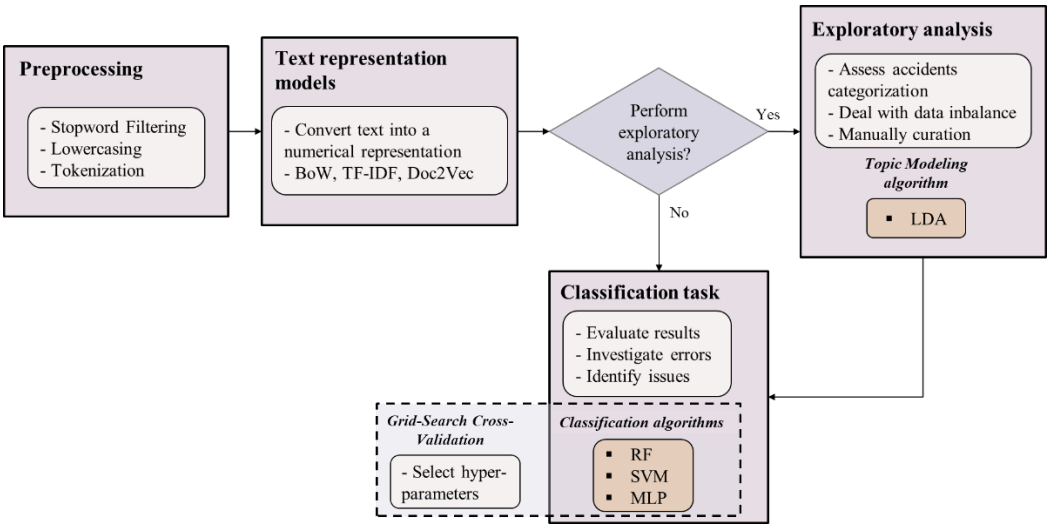

Fig. 1 - Main steps and models used on our proposed methodology.

Firstly, raw reports are pre-processed using operations to remove noise, then are converted into feature vectors through different text representation models. Next, we performed an exploratory analysis, using the resulting representations to summarize their main contents and identify patterns, searching for useful information to perform a further assessment. Then, we trained several classifiers using different combinations of ML algorithms and feature vector representations to categorize the reports into different groups. Finally, we evaluated their predictions and compared their performance. Each step was developed in Python computational language.

3. DISCUSSION

The database we analyzed is presented in the form of a spreadsheet, where rows correspond to an accident investigation report and columns are characteristics (factors) about the event itself and employee involved in the event. Each report contains several factors; thus, the safety technician must fill out almost 60 fields with either a ‘Yes’ or ‘No’ answer or provide a short text. The filling of the long report makes the process of investigating, recording, and documenting the accidental event boresome. In fact, difficulties in filling out the reports are a common problem already mentioned by the safety technician. Moreover, the lack of standardization for describing the factors may hinder the efficient use of the accident investigation reports database for supporting decision-making for risk management.

Here, we focus our analysis on the ‘accident agent’ as it represents a valuable source of information to identify common elements about the cause of accidents and propose preventive measures. The safety technician standardized these categories into nine different classes, which are presented in Tab. 1 (a). However, we identified that, no matter the background and expertise, the safety technician was often confused by the

categories when reporting the accidental event. We manually evaluated the original pre-defined categories, merging the ones that seem to be similar. The restructuring resulted in six categories as shown in Tab. 1 (b).

Tab. 1 - 'Accident agent' categories (a) before and (b) after restructuring.

Label	Original Category
0	Scaffolding
1	Duct, ditches, pipes, tunnels, pressure vessels
2	Building, structure, pole, tower, rope, cable, electrical cable, chair, drums, pulleys, tanks, cylinders, tank protection
3	Manual and automatic tools, drilling machines, sander, polisher, grinder, drill, lathe, electrical discharge machine, electrical equipment, electric arc, hydraulic or pneumatic
4	Engine, pump, turbine
5	Trip or slip
6	Chemical substance and industrialized metal, lead, mercury, zinc, cadmium, chromium, rebar, ferrous alloy
7	Commuting accidents
8	Motor vehicle, motorcycle, tractor scooter, on track, hoisting equipment

(a)

Label	Restructured Category
0	Scaffolding
1	Administrative fall/ injury
2	Equipment/ tools
3	Chemical products
4	Commuting
5	Others

(b)

Moreover, we performed an exploratory analysis using LDA to summarize the main themes of a collection of reports. We found thirteen topics and the results reinforced the idea that there are redundant categories. Indeed, it was possible to notice that many topics overlap. For instance, two of the topics involve '*slipping*'. However, they differ in terms of the 'accident agent': '*stairs*' in one topic, while '*floor*' for the other topic.

After reorganizing the database, we proposed new labels (Tab. 1 (b)). Five groups (0-4) are formed by accident investigation reports with common causes within each group, and the remaining set (5: 'others') is composed by events that have different causes and are not assigned to any specific category.

We performed ML classification tasks using the original (Tab. 2 (a)) and restructured categories (Tab. 2 (b)). As one can see, the manual labeling increased the accuracy by about 15% in some cases (e.g., SVM-TF-IDF). The significant improvement in performance confirms the assumption of mislabeling in filling out accidents' investigation report based on the original categories. Indeed, the new categorization apparently is more coherent, aligned with the safety technician view; thus, being beneficial for the identification of patterns.

Tab. 2 - Median accuracy (%) of the classification task using (a) nine and (b) six categories

	Original Categories		
	BoW	**TF-IDF**	**Doc2Vec**
SVM	51.94	51.94	11.24
RF	54.65	56.20	13.18
MLP	58.91	57.76	10.86

(a)

	Restructured Categories		
	BoW	**TF-IDF**	**Doc2Vec**
SVM	64.29	68.26	36.51
RF	67.86	67.06	34.52
MLP	70.63	69.44	37.30

(b)

These results strengthen the idea that TM and NLP provides auspicious techniques to identify poor datasets

once their performance relies on the quality of the database. The proposed approach was able to identify issues on the filling out of the reports as well as in the safety technician grasp on the standard NBR 14280 [6]. Thus, adjustments are necessary to provide documents that are more capable of retaining the knowledge acquired from the events, and then could be reused by the company and improve the current safety environment.

4. CONCLUSION

We analyzed a dataset of accident investigation reports of a real company through different TM and NLP approaches. We were able to identify the usefulness of several categories already adopted, but there were also existing ones that we found out to be ineffective in terms of their descriptions. The results obtained in the exploratory analysis suggested that a lower number of categories would be more suitable for this specific database. This is probably due to a lack of standardization and understanding of the pre-defined categories.

Moreover, the improvement on the performance of the classifiers due to the manual curation may indicate the presence of inconsistencies in the original classification among the 'accident agents'. We showed the importance of the company's safety culture to keep safety technician engaged in constructing a well-structured database. In fact, the safety technician must have the correct understanding of what and how to fill out each required field in the report, which is only achieved by continuous training. This would improve the quality of the reports and allow keeping the original categorization. In addition, a well-designed database provides useful information for risk management and decision-making.

Finally, the quality of filling in the reports was poor, hindered the classifiers' performance, which explains the improvement observed after reclassification. As a future research goal, we aim to analyze other factors (i.e., accidents common causes) in the current database, investigating consequences and possible costs of accidents, for example, associated with injury leave.

5. ACKNOWLEDGEMENT

The authors thank the National Agency for Research (CNPq), the Foundation of Support for Science and Technology of Pernambuco (FACEPE), *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001, and the 'Human Resources Program (PRH) da National Oil Company (ANP) and Finep (Brazilian Innovation Agency) - PRH-ANP 38.1: "Risk Analysis and Environmental Modeling in Exploration, Development and Production of Oil and Gas" for the financial support through research grants.

6. REFERENCES:

[1] H. Hao, Y. (Eric) Li, A. Medina, R. B. Gibbons, and L. Wang, "Understanding crashes involving roadway objects with SHRP 2 naturalistic driving study data," *J. Safety Res.*, vol. 73, pp. 199–209, Jun. 2020, doi: 10.1016/j.jsr.2020.03.005.

[2] C. Pimm *et al.*, "Natural Language Processing (NLP) tools for the analysis of incident and accident reports To cite this version : HAL Id : halshs-00953658 Natural Language Processing (NLP) tools for the analysis of incident and accident reports," no. April, 2014.

[3] T. Madeira, R. Melício, D. Valério, and L. Santos, "Machine Learning and Natural Language Processing for Prediction of Human Factors in Aviation Incident Reports," *Aerospace*, vol. 8, no. 2, p. 47, 2021, doi: 10.3390/aerospace8020047.

[4] J. I. Single, J. Schmidt, and J. Denecke, "Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing," *Saf. Sci.*, vol. 129, no. May, p. 104747, 2020, doi: 10.1016/j.ssci.2020.104747.

[5] M. F. Ballesteros, S. A. Sumner, R. Law, A. Wolkin, and C. Jones, "Advancing injury and violence prevention through data science," *J. Safety Res.*, vol. 73, pp. 189–193, Jun. 2020, doi: 10.1016/j.jsr.2020.02.018.

[6] 14280 NBR, "NBR 14280:2000. Cadastro de acidente do trabalho - Procedimento e classificação.," *Nbr*, p. 94, 2001.