PROACTIVE MANAGEMENT OF OCCUPATIONAL RISKS IN AN ORGANIZATION IN THE OIL SECTOR USING BUSINESS INTELLIGENCE AND TEXT MINING

Jesus David Julio-Parra, Universidad De Manizales
Juan Pablo Giraldo-Rend, Universidad De Manizales

ABSTRACT

Occupational health and safety are strategic variables of great importance for the development of any organization, since it reflects the responsibility towards its workers and the environment where it is developed. In this article, an analysis of the reports of unsafe acts and conditions is carried out through the application of business intelligence and text mining techniques, which led to the construction of an interactive dashboard that includes a word cloud diagram allowing to identify the phrases most reported by workers. In addition, it associates them with risk situations that may be occurring in different projects. This allows timely decision making that leads to intervene in the occurrence of incidents. Finally, the limitations of the analysis are presented and suggestions for improvement of the method used are proposed.

Keywords: Occupational Health and Safety, Business Intelligence, Text Mining, Incidents, Unsafe Acts.

INTRODUCTION

Occupational health and safety are strategic components for the development of any organization, so it is no longer a secondary or support process (Gan et al., 2020). In the hydrocarbon industry, people are exposed to risks and hazards that can cause accidents, generate permanent injuries, occupational diseases and even death. According to Hamalainen et al. (2017), 2.78 million workers die each year from occupational accidents and occupational diseases (of which 2.4 million are disease-related), and 374 million workers suffer non-fatal occupational accidents. According to the Colombian Ministry of Health and Social Protection and Labor Ministerios (2019) in 2018, out of every 100 workers affiliated to the General System of Occupational Risks (SGRL), 6.2 suffered an occupational accident in Colombia. The economic sector with the highest accident rate was mines and petroleum, with 13 accidents per 100 workers.

Data analytics is increasingly applied in business management and in the analysis of occupational accidents (Ragini et al., 2018), because it is a tool that enables greater control of costs and times, and the optimization of processes and resources (Mohammadpoor & Torabi, 2020). In this context, Natural Language Processing (NLP), as part of Artificial Intelligence and Computational Linguistics, gains key relevance, because it is a useful tool to extract valuable information from text records, which helps to improve different processes in companies (Zhai & Massung, 2016). In recent years, there has been a great deal of interest in the automatic classification or tagging of text logs that are made when incidents occur in companies (Pan & Zhang, 2021). The above is in line with what is stated by Zhang et al. (2019), who argue that, in

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.

order to prevent similar accidents from reoccurring in the future, scientific risk control should be done. For this purpose, accident analysis using text mining and natural language processing techniques are essential (Cheng et al., 2020).

This article looks at text logs of unsafe condition reports from workers in an organization in the petroleum industry. Numerous similar studies have been conducted in this regard, with the difference that the text records are analyzed, but from accidents that have already materialized, i.e., they extract valuable knowledge from what has already happened to prevent future accidents, which means that they work with reactive and not proactive data. However, this limitation is overcome in this study. (Goh & Ubeynarayana, 2017), for example, evaluated the usefulness of various text mining classification techniques to classify 4,471 accident reports in the construction industry obtained from the US OSHA website, manually labeled 1,000 records and after dividing them into training (749) and test (251) datasets, evaluated the performance of 6 different machine learning algorithms, resulting as the highest performing Support Vector Machine (SVM) (Verma et al., 2016).

A similar study to the previous one was developed by Zhang et al. (2019), which analyzed accidents in the construction industry. However, unlike the previous study, these authors also relied on natural language processing techniques and ensemble techniques. Ensemble techniques are Machine Learning techniques that combine several base models to produce an optimal predictive model by building a set of classifiers and then classify new data points by taking a (weighted) vote of their predictors. Using Deep Learning and text mining techniques, Zhong et al. (2020) classified and analyzed the narrative surrounding accidents in the construction industry to better understand how they originate. This study used word assembly technique to model the semantic narratives of accidents, and then used a convolutional neural network (CNN) model to automatically extract text features and classify accident narratives without the need for manual feature processing.

Most of the identified studies have been focused on analyzing text records of accidents that have already occurred. However, it would be advantageous to analyze the text records of reports of unsafe acts and conditions, i.e., conditions that did not materialize in an accident and, therefore, no physical or material damage is recorded. This study analyzes the behavior of the data of unsafe acts and conditions, reported through a mobile application called REPORTA, in which, through business intelligence and text mining, those patterns are identified that will serve for decision making to avoid accidents. Additionally, a Dashboard is presented to perform a real monitoring of the different events and thus control the factors that lead to the high number of incidents.

METHODOLOGY

This article followed, from a general level, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology Chapman et al.(2000), which implied developing the following stages: (1) business understanding (assess situation; determine data mining goals; and produce project plan); (2) data understanding (collect initial data; describe data; verify data quality); (3) data preparation (select data; construct data; integrate data); (4) modeling (select representation technique and build model); and (5) deploy (produce final report). Each component of the methodology is explained in the following lines.

Business Understanding

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.

- Business case analysis.
- Situation assessment (Inventory of resources, requirements, assumptions, business terminology, among others).
- Establishment of data mining objectives (goals and success criteria).

Data Understanding

- The data set used in this study is based on the reports of unsafe acts and conditions that workers register in the mobile application *"REPORTA"*.
- The sample is made up of records from December 2019 to March 2021, which total 8,743 reports, of which 7,196 have been due to unsafe conditions and 1,547 observations of unsafe behaviors.

Data Preparation

- Data cleaning: columns that contributed little value to our object of study and empty rows were eliminated.
- Data transformation: date format transformation.
- Data Integration: model enrichment with georeferential data from DANE.

Modeling

- The relationship with the Fact Table was built. Creation of the customized measures in the Power BI tool using the DAX language.
- A sentiment analysis of the description in text made by the workers of the reported finding was performed.
- Keyphrase extraction: extraction of the keywords from the *"Find"* field was performed. Analysis of which words were the most repeated in the reports.

Evaluation

Assessment

- Results evaluation.
- Process evaluation.
- Next steps.

Implementation

- Implementation process definition.
- Monitoring and maintenance planning.
- Final report generation.
- Proyect análisis.

Data Collection Process

It is common that the extraction and selection of data for analytics requires multidimensional views as expressed by (Cheng & Chen, 2019), the techniques and forms to reduce the dimensionality allow to improve the classification. Facilitating dimensionality, the data set used in this study is based on the reports of unsafe acts and conditions that workers register in the mobile application *"REPORTA"*. The records generated by the application are mainly texts where an observer describes the observed risk situation. The sample is made up of

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.

records from December 2019 to March 2021, which total 8,743 reports, of which 7,196 have been due to unsafe conditions and 1,547 observations of unsafe behaviors Table 1.

Data Dictionary

Table 1
DATA DICTIONARY

Field name	Data type	Description
Names	Text	Name and surname of the person making the report
Report_Completed	Text	Indicates the category to which the observed finding belongs
Report_Type	Text	Type of report made (unsafe acts, unsafe conditions, near misses or stop work actions - ADT)
Finding	Text	Brief and clear description of the finding identified by the worker
Intervention	Text	Describes the immediate action taken to correct the unsafe act
Proposal_Action	Text	Describes the proposed action to prevent the finding from recurring
Role	Text	Describes the type of position held by the person making the report
Function	Text	State the area of the project where the report was submitted
Latitude	Text	Indicates the latitude of the area where the report was filed.
Longitude	Text	Indicates the longitude of the area where the report was filed.
Date	Date	Date on which the finding was reported
Close_Action	Text	Indicates the action taken to close the reported finding
Polarity_Score	Text	Indicates the polarity score of the sentiment analysis
Compound	Number	Indicates composite polarity in decimals
Polarity	Text	Indicates polarity value (Negative, Neutral, Positive)
Keyphrases	Text	Indicates keywords extracted from the "Findings" field

EXPLORATORY ANALYSIS AND DATA PROCESSING

After having the data collected in Excel, we proceeded to link it with the Power Bi tool. Once linked, an exploratory analysis of the different variables was carried out, eliminating those that contributed little value to the object of study, transformations were performed and the columns and measures necessary to explain the phenomenon were created. The steps carried out are as follows: (1) data cleaning: columns that contributed little value to our object of study and empty rows were eliminated; (2) data transformation: data were transformed according to the type of data previously stated in the data dictionary. The date was in YYYYY/MM/DD format and was transformed to the traditional DD/MM/YYYYYY format; (3) additional column *"Coordinates":* once the *"area"* column was extracted, we proceeded to use it to represent a visualization of a map of the 10 bases distributed in the southwest-east of the country, doing it in this way only with the names of the municipalities where the bases are located, gave us errors in the map, so it was decided to obtain an additional table called *"Coordinate"*, this table contains the latitudes and longitudes of these municipalities, after making the table we proceeded to build the relationship with the Fact Table; (4) measures in DAX language: once the exploratory analysis and the necessary transformations to the data were done, we proceeded to the creation of the customized measures in the Power Bi tool using the DAX language; (5) sentiment analysis: due to the fact that most of the variables we have in the dataset are of text type, a sentiment

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.

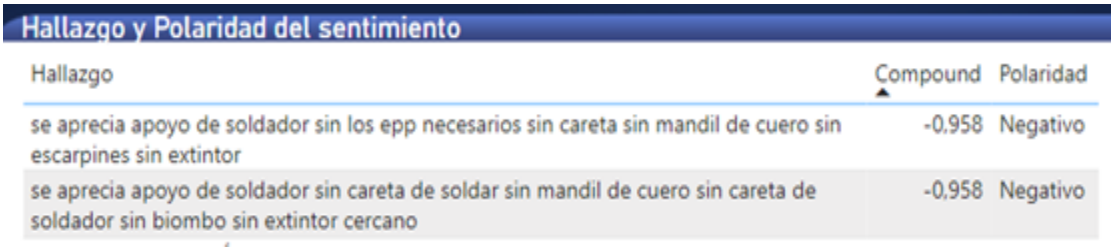
analysis of the description in text made by the workers of the reported finding was performed, for this the "*Finding*" field was selected as the origin and a Python script was executed, which uses the pandas packages for the dataframe, the NLTK package which is the number 1 natural language analysis toolkit in the market; and (6) key phrase extraction: to perform an analysis of which words were the most repeated in the reports made and therefore to determine their incidence for the prevention of the occurrence of accidents, we proceeded to use the Microsoft Azure API, specifically the Cognitive Services service, where through the use of a script and the use of a private key (key) a custom function was created in Power Bi to achieve the connection with Microsoft Azure and extract the keywords from the *"Find"* field (Durbin, 2003).

Dashboard

As a base model of classification and comparison, the proposal of Munoz Caceres (2020) and Raviv et al. (2015) was used based on its structure, an initial classification is segmented for visualization. The result are four scorecards were created that are considered of great relevance in the organization, in which, through graphs, different patterns of behavior can be observed in the occurrence of incidents and accidents at work, which will allow decision making, implementation of training campaigns, provision of physical infrastructure or its repair. With this result in a word map, is necessary advance with result and create diagrams for traceability of data and offer to the company element for decisions, in particular case is priority risks and projects for occupational risks (Work, 2019).

RESULTS

After extracting the *"Keyphrases"* from the description of the *"finding"* identified by the worker, we proceeded to make a word cloud graph, which allowed us to identify which were the words related to the most predominant risk factors; this is very important because according to these words, priority strategies are proposed to attack the factors that cause these risks. On the other hand, the application of sentiment analysis produced good results, since 83.53% were neutral, 16.05% were negative and only 0.42% were positive. These data are in accordance with the type of findings reported, since all findings by their definition are unfortunate events that are always associated with negative things. The findings with a higher negative polarity (-0.958) are shown in Figure 1.



Hallazgo y Polaridad del sentimiento

Hallazgo	Compound	Polaridad
se aprecia apoyo de soldador sin los epp necesarios sin careta sin mandil de cuero sin escarpines sin extintor	-0.958	Negativo
se aprecia apoyo de soldador sin careta de soldar sin mandil de cuero sin careta de soldador sin biombo sin extintor cercano	-0.958	Negativo

FIGURE 1
FINDINGS WITH A HIGHER NEGATIVE POLARITY

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.

In order to analyze in which geographic area more reports are being made and to verify if these findings are being closed with the required timeliness, a geo-referencing graph of the project areas was made (See Figure 2).

FIGURE 2
GEOREFERENCING OF ANALYZED REPORTS

Once the group of words with the key phrases or tokens (Stanford University, 2009) has been made, it can be filtered by type of project, type of report or by the role of the worker who made the report. This type of dynamic filters is extremely important to identify the priority risks according to the type of project, for example: the key phrases that are displayed in the CNPC-LOTE X project (construction and maintenance of oil pipelines), are very different from the CENIT- STATIONS (maintenance of oil pumping stations), which facilitates the implementation of action plans against occupational risks according to the type of project reported.

DISCUSSIONS

The importance of data analytics is crucial in any organization, since the ability to visualize the information, cross it with different factors, interpret it and be able to take actions for improvement One of the practical implications of this project were the scorecards used, because they contribute to prevention through various actions, since being able to detect in time, site, event, occurrence, date and quantity of certain incidents, they manage to establish quick and early measures that directly affect the improvement of performance standards, quality of work and reduction of reports.

The CNPC-LOT X project has 33% of all reported findings related to unsafe conditions related to *"Equipment and tools in poor condition" and "unsafe work surfaces"*; therefore, strategies are recommended to coordinate with the maintenance department to promptly resolve such maintenance and thus prevent the occurrence of an accident related to these factors. In the Cenit lines project it is evident that the reports made are uniformly distributed among the positions that have more recurrent activities in the field, such as operating personnel, mechanical

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.

supervisors and HSE supervisors, which is a sign that all personnel are committed to identifying risks and have a perception of risk; unlike contracts such as *"UNIPHOS"*, where most of the reports come from the coordinating and supervisory personnel, therefore strategies and awareness and training campaigns should be proposed so that operating personnel have greater involvement in this process. On the other hand, the extraction of key words and their subsequent visualization in the form of a cloud is a good tool to find those recurring words related to the most repeated risks. In the case of the *"Cenit líneas"* project, training strategies should prioritize the use of personal protection elements such as goggles, timely replacement of safety boots and gloves, reinforcing the necessary safeguards in the excavations, such as signage, and finally generating order and cleanliness campaigns in all locations. Analyzing reports of unsafe conditions and unsafe acts is a proactive way to prevent accidents. In this project only business intelligence and an algorithm for extracting key phrases from text fields were applied. In the literature there have been projects of automatic classification of accident reports, for this reason for better risk management it is advisable to apply Machine Learning and natural language processing techniques to process the text field *"finding"* and thus automatically classify each report in a specific category that allows management to propose more effective action plans and early prevention actions.

REFERENCES

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS Inc*, *9*(13), 1-73.

Cheng, C.H., & Chen, H.H. (2019). Sentimental text mining based on an additional features method for text classification. *Plos One*, *14*(6), e0217591.

Cheng, M.Y., Kusoemo, D., & Gosno, R.A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, *118*, 103265.

Durbin, M.T. (2003). *The dance of the thirty-ton trucks: demand dispatching in a dynamic environment*. George Mason University.

Gan, W.H., Lim, J.W., & Koh, D. (2020). Preventing intra-hospital infection and transmission of coronavirus disease 2019 in health-care workers. *Safety and Health at Work*, *11*(2), 241-243.

Goh, Y.M., & Ubeynarayana, C.U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention*, *108*, 122-130.

Hamalainen, P., Takala, J., & Kiat, T.B. (2017). Global estimates of occupational accidents and work-related illnesses 2017. *World*, *2017*, 3-4.

Ministerios. (2019). https://www.minsalud.gov.co/proteccionsocial/RiesgosLaborales/Paginas/indicadores.aspx

Mohammadpoor, M., & Torabi, F. (2020). Big Data analytics in oil and gas industry: An emerging trend. *Petroleum*, *6*(4), 321-328.

Munoz Caceres, P.A. (2020). Design and construction of a security incident classification model using NLP in written text records to automate labeling.

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517.

Ragini, J.R., Anand, P.R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, *42*, 13-24.

Raviv, G., Shapira, A., & Fishbain, B. (2015). Analysing and modeling near misses in crane-work environments. In *ARCOM Doctoral Workshop on Health, Safety and Wellbeing*, 37-49.

Stanford University. (2009). https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html.

Verma, A., Rajput, D., & Maiti, J. (2016). Prioritization of near-miss incidents using text mining and bayesian network. *In International Conference on Advances in Computing and Data Sciences*, 183-191.

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.

Work, O.D. (2019). Safety and health at the center of the future of work. Safety and health at the center of the future of work. *Switzerland: International Labor Organization*.

Zhai, C., & Massung, S. (2016). Text data management and analysis: a practical introduction to information retrieval and text mining. Morgan & Claypool.

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, *99*, 238-248.

Zhong, B., Pan, X., Love, P.E., Ding, L., & Fang, W. (2020). Deep learning and network analysis: Classifying and visualizing accident narratives in construction. *Automation in Construction*, *113*, 103089.

Received: 10-Oct-2022, Manuscript No. JMIDS-22-12660; **Editor assigned:** 12-Oct-2022, Pre QC No. JMIDS-22-12660 (PQ); **Reviewed:** 26-Nov-2022, QC No. JMIDS-22-12660; Revised: 29-Oct-2022, Manuscript No. JMIDS-22-12660(R); **Published:** 05-Nov-2022

Citation Information: Julio-Parra, J.D., Giraldo-Rend, J.P. (2023). Proactive management of occupational risks in an organization in the oil sector using business intelligence and text mining. *Journal of Management Information and Decision Sciences, 26*(1), 1-8.