*Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)*
*Edited by* Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P694-cd

# A Method based on Natural Language Processing for Periodically Estimating Variations of Performance of Safety Barriers in Hydrocarbon Production Assets

Dario Valcamonico

*Department of Energy, Politecnico di Milano, Italy. E-mail: dario.valcamonico@polimi.it*

Piero Baraldi

*Department of Energy, Politecnico di Milano, Italy. E-mail: piero.baraldi@polimi.it*

Enrico Zio

*MINES Paris-PSL University, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France.*
*Department of Energy, Politecnico di Milano, Italy. E-mail: enrico.zio@polimi.it*

Luca Decarli

*Eni NR, Milan, Italy*

Laura La Rosa

*Eni NR, Milan, Italy*

Giuseppe Nicotra

*Eni NR, Milan, Italy*

This work develops a methodology for estimating the performance of safety barriers of hydrocarbon production assets using reports of Process Safety Events (PSEs). We address the challenge of dealing with assets that evolve, e.g., due to degradation of components or maintenance interventions, so that also the performances of the safety barriers vary in time. The proposed methodology combines a taxonomy of the words used in the reports, a method of Natural Language Processing (NLP) for the extraction of the keywords, which takes into account the time of occurrence of the PSEs, and a technique for estimating the barrier performance from the number and severity of PSEs involving the barrier. The proposed methodology is validated using a repository of reports of PSEs in hydrocarbon production plants.

*Keywords*: Process safety events, Natural Language Processing, Taxonomy, Hydrocarbon Production Plants.

## 1. Introduction

Hydrocarbon production plants are threatened by the occurrence of Process Safety Events (PSEs), i.e. unplanned or uncontrolled events during which a Loss Of Primary Containment (LOPC) of any substance, including non-toxic and non-flammable material, occurs ("API RP 754" 2021). The availability and safety of the systems is achieved with multiple and independent barriers, which prevent the occurrence and mitigate the consequences of PSEs (Decarli et al. 2020). Assessing the performance of the safety barriers is a fundamental step of Quantitative Risk Assessment (QRA), which aims at modelling and quantitatively estimating the risk of PSEs in hydrocarbon production assets (Zhang et al. 2018).

After the occurrence of a PSE, system operators write a textual report with free-text and multiple-choice fields to describe the event, its causes and consequences, and the level of severity of the LOPC with reference to predefined tier levels ("API RP 754" 2021). A methodological framework based on the combination of techniques of Natural Language Processing (NLP) and a Bayesian Network (BN) has been proposed for the systematic analysis of PSE reports in (Valcamonico et al. 2022). The present work consolidates and extends the previous work (Valcamonico et al. 2022) with the final aim of periodically estimating the performance of the safety barriers.

Since hydrocarbon production assets are continuously evolving, e.g., due to degradation of components, maintenance interventions, installation of new equipment, adoption of new training procedures of the plant personnel and implementation of countermeasures to improve safety, the methodology should be able to follow the modifications of the assets and to periodically update the estimations of the performance of the safety barriers.

The proposed methodology combines a taxonomy of the words used in the reports, a method of NLP for the extraction of the keywords, and a technique for estimating the barrier performance from the number and severity of PSEs involving the barrier.

The taxonomy of the words associated with the design, operation and failure of the safety barriers is built by experts of the hydrocarbon production assets and used to organize the words into a top-down structure of sub-levels, where each sub-level represents a specific safety barrier (Mackie, Welsh, and Lee 2006). Then, the methodology developed in (Valcamonico et al. 2022) to identify the keywords associated to the reports, which is based on the NLP technique of Term Frequency Inverse Document Frequency (TFIDF) (Weiss et al. 2005), is modified to account for possible modifications of the asset. The idea, inspired by (Marwah and Beel 2020), is to multiply the TFIDF of a word in a report by an additional term inversely proportional to the age of the report, so that words written in recent PSEs are more important than words written in old PSEs. Finally, a method for estimating the performance of the safety barriers is developed. It builds up on the method proposed in (Di Maio et al. 2021), which estimates the performance of a given barrier using dedicated key performance indicators computed considering the number and severity of the PSEs occurred in a given interval of time. The novel method exploits the capability of the developed NLP technique of identifying the PSE events containing keywords associated to a specific barrier. This allows estimating the performance of a given barrier considering only the PSEs related to that barrier.

The methodology has been verified through its application to a real repository of reports of PSEs occurred in a fleet of hydrocarbon production plants.

The remaining of the work is organized as follows. Section 2 states and formulates the problem. Section 3 describes the developed methodology. Section 4 presents the case study and Section 5 reports the obtained results. Section 6 discusses the work conclusions.

## 2. Problem statement and formulation

We consider a repository of $D$ PSE reports, $R = \{r_i, i = 1, \dots, D\}$, where each report $r_i$ is the free text description of the PSE, which typically contains information on the state of a subset of the $G$ safety barriers, $\{B_1, \dots, B_g, \dots, B_G\}$, installed in the asset. The time $[t_0, t_f]$ during which the reports have been collected is divided into $N_T$ intervals $[t_{h-1}, t_h)$, $h = 1,2, \dots, N_T$, of the same length $\Delta T = \frac{t_f - t_0}{N_T}$. At time $t_h = t_0 + h\Delta T$, $h = 1, \dots, N_T$, the $D_h$ reports collected in the time interval $[t_{h-1}, t_h)$ become available. The set of reports collected in $[t_{h-1}, t_h)$ will be referred to as $R_h = \{r_i, i = \sum_{\tilde{h}=1}^{h-1} D_{\tilde{h}}, \dots, \sum_{\tilde{h}=1}^{h} D_{\tilde{h}}\}$. A label $z_i$, which indicates the severity of the consequences of the PSE and can assume the three possible values $\{tier_1, tier_2, tier_3\}$, with $tier_1$ indicating severe consequences, $tier_2$ moderate consequences, and $tier_3$ minor consequences, is assigned to each report $r_i$, $i = 1, \dots, D$, by the plant operators.

The objective of this work is the development of a framework for the extraction of knowledge on the performance of the safety barriers from the set of reports $R_h$ collected in the sequential intervals of time $[t_{h-1}, t_h)$, $h = 1, \dots, N_T$, with the objective of supporting QRA. This is achieved by:

*i.* Building a taxonomy to systematically structure the language used by the operators to describe the $G$ barriers involved in the PSEs and their performances during PSEs. The taxonomy is composed by a structure of three hierarchical sub-levels, where the $M$

words of the vocabulary of the reports $\{v^1, \ldots, v^M\}$ are assigned by system experts to the corresponding barrier $B_g$, $g = 1, \ldots, G$.;

*ii.* Developing a methodological framework to systematically identify $N^h$ keywords $\{k^h_{l^h}, l^h = 1, \ldots, N^h\}$ from the vocabulary of the reports collected from $t_0$ to $t_h$. The obtained keywords provide a compact representation of the reports, which carries most of their semantic content;

*iii.* A method to quantify the barriers performance exploiting the taxonomy and the extracted keywords.

## 3. Developed methods

Section 3.1 describes the developed taxonomy, Section 3.2 the proposed method for the identification of the keywords and Section 3.3 the method for the estimation of the performance of the barriers.

### 3.1. *Taxonomy*

Taxonomies are top-down hierarchical structures that are built to systematically structure the knowledge in a domain (Dayrat 2005). The taxonomy built in this work organizes the vocabulary made by the $M$ words used by the operators in the PSE reports into three hierarchical sub-levels. The first sub-level distinguishes the type of barrier into “Preventive” and “Mitigative”, the second sub-level specifies the safety barriers and the third sub-level reports the words associated to the barriers. Figure 1 shows a sketch of the taxonomy, where:

*i.* the first sub-level contains the two states $\Psi^1$=”Preventive” and $\Psi^2$=”Mitigative”;

*ii.* The second sub-level contains the $G$ states $\{\Gamma_1, \ldots, \Gamma_{G_1}\}$ and $\{\Gamma_{G_1+1}, \ldots, \Gamma_G\}$ which are associated to the specific safety barriers of $\Psi^1$ and $\Psi^2$, respectively;

*iii.* The third sub-level contains the $M_g$ words $\{v^1_g, \ldots, v^{M_g}_g\}$ assigned to each state $\Gamma_g$, $g = 1, \ldots, G$. Notice that $\sum_{g=1}^{G} M_g = M$

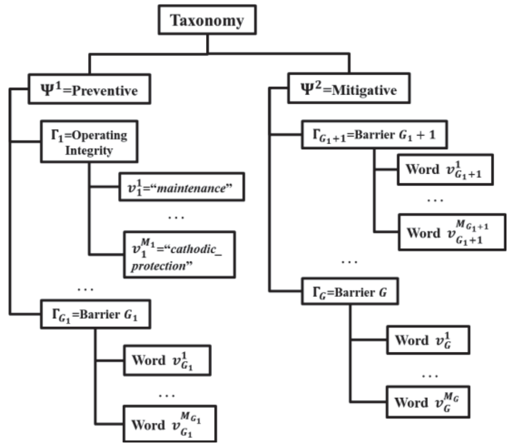


Figure 1: Sketch of the developed taxonomy. For confidentiality reasons the barriers and the words are not specified.

The definition of the sub-levels and the assignment of the words to them is operated by leveraging the knowledge of experts of the assets on the barriers of the system.

### 3.2. *Keywords identification*

The method proposed for the identification of the keywords of the PSE reports extends the method proposed in (Valcamonico et al. 2022) for the same purpose. The main difference is that the new method considers the time of occurrence of the PSEs. The main idea is to represent the current state of the system by giving less importance to words used in old reports that do not appear later, since they typically refer to issues that have been resolved and thus tend to not occur anymore. Conversely, the importance of words used in recent reports is increased, since they are expected to refer to new unresolved issues that need to be addressed or issues that have not been fully resolved despite past efforts.

Specifically, once at time $t_h$, $h = 1, \ldots, N_T$, the reports $R_h$ collected in the time interval $[t_{h-1}, t_h)$ become available, the method proposed for the extraction of the keywords $\{k^h_{l^h}, l^h = 1, \ldots, N^h\}$ is applied to the set of reports $\bigcup_{\tilde{h}=1}^{h} R_{\tilde{h}}$ collected in the time interval $[t_0, t_h)$.

Firstly, the vocabulary of the words used in the reports, $\Delta^h = \{v^h_{j^h}, j^h = 1, \ldots, M^h\}$, is built by collecting the unique words and their contiguous combinations, i.e. $n$-grams, used in the set of reports $\bigcup_{\tilde{h}=1}^{h} R_{\tilde{h}}$. Then, the identification of the $n$-grams is obtained by computing their Normalized Pointwise Mutual Information (NPMI) index (Bouma 2009), which measures the amount of information that a specific $n$-gram carries with respect to the information of the words alone, and by including it in $\Delta^h$ only if its associated NPMI index is larger than a threshold $thresh_{NPMI}$ set using the procedure described in (Valcamonico et al. 2023).

The set of $N^h \leq M^h$ keywords $\{k_{l^h}^h, l^h = 1, \dots, N^h\}$ is extracted from the words of the vocabulary $\Delta^h$ by defining a measure of semantic importance of the words. In (Valcamonico et al. 2022), this is done by using the NLP technique of Term Frequency Inverse Document Frequency (TFIDF) (Weiss et al. 2005). Each report, $r_i \in \bigcup_{\tilde{h}=1}^{h} R_{\tilde{h}}$, with $i = 1, \dots, \sum_{\tilde{h}=1}^{h} D_{\tilde{h}}$, is transformed into the numerical vector $\phi_i = \{\phi_{i,1}, \dots, \phi_{i,j^h}, \dots, \phi_{i,M^h}\}$, whose generic element $\phi_{i,j^h}$, which quantifies the semantic importance of the word $v_{j^h}^h$ in the report $r_i$, is:

$$\phi_{i,j^h} = TF_{i,j^h} log\left(\frac{\sum_{\tilde{h}=1}^{h} D_{\tilde{h}}}{DF_{j^h}}\right) \qquad (1)$$

where $TF_{i,j^h}$ is the number of times that the word $v_{j^h}^h$ occurs in the report $r_i$, $\sum_{\tilde{h}=1}^{h} D_{\tilde{h}}$ is the number of reports collected from $t_0$ to $t_h$ and $DF_{j^h}$ is the number of reports in $\bigcup_{\tilde{h}=1}^{h} R_{\tilde{h}}$ wherein the word $v_{j^h}^h$ is used. The idea of the measure $\phi_{i,j^h}$ is that the importance of word $v_{j^h}^h$ in the report $r_i$ is directly proportional to $TF_{i,j^h}$ (a word repeated several times is expected to be relevant for the report) and inversely proportional to $DF_{j^h}$ (a word present in a large number of reports is expected to provide less specific semantic information about the report, given its scarce specificity, than a word present in few reports).

In the present work, the TFIDF measure in Eq.(1) is modified to consider the age of the reports according to (Marwah and Beel 2020). The semantic importance of the word $v_{j^h}^h$ in report $r_i$ is obtained by multiplying the right-hand term of Eq.(1) for a new factor:

$$\tilde{\phi}_{i,j^h} = TF_{i,j^h} log\left(\frac{\sum_{\tilde{h}=1}^{h} D_{\tilde{h}}}{DF_{j^h}}\right) log\left(1 + \frac{DF_{j^h}}{h - h_i + 1}\right) \qquad (2)$$

where $h_i \in \{1, \dots, h\}$ is the time interval in which the report $r_i$ has been collected. Notice that by introducing the term $log\left(1 + \frac{DF_{j^h}}{h-h_i+1}\right)$, the importance of the word $v_{j^h}^h$ is inversely proportional to the age $h_i$ of the report. This is in accordance with the observation that if a report is recent, it can contain critical issues even if they appear in many other reports or have occurred in the past.

The final result is that the measure of importance in Eq.(2) that is decreasing for words which are progressively not used and increasing for words that are used more recently or reoccur.

Based on Eq.(2), the word $v_{j^h}^h$ is a keywords of the report $r_i$ if its importance $\tilde{\phi}_{i,j^h}$ is larger than a threshold, $thresh_k$, set by using the procedure described in (Valcamonico et al. 2023). Notice that since the words of the vocabulary are organized in the taxonomy, also the keywords are automatically organized in the sub-levels of the taxonomy.

### 3.3. *Estimation of the performance of the barriers*

In (Di Maio et al. 2021), the health state of each barrier is defined as a multistate variable which can assume values "High" ("H"), which indicates that the barrier is in a as-good-as-new condition, "Medium" ("M"), which indicates that the barrier is moderately degraded, and "Low" ("L"), which indicates that the barrier is in a critical condition that may cause it to cease its function. A probability is assigned to each health state of the barrier $B_g$, $g = 1, \dots, G$, by defining barrier-specific Key Performance Indicators (KPIs), $x^g \in [0,1]$, $g = 1, \dots, G$. The relationships between the health state of the barrier $B_g$ and its KPI $x^g$, $g = 1, \dots, G$, is given by an expert-based probabilistic models.

Considering, for example, the barrier "Operating Integrity" (OI), its KPI at time $t_h$, $x_h^{OI}$, $h = 1, \dots, N_T$, is:

$$x_h^{OI} = \frac{\sum_{z=tier_1}^{tier_3} w_z f_1(D_{h,z})}{\sum_{z=tier_1}^{tier_3} w_z} \quad (3)$$

where $D_{h,z}$ is the number of PSE reports in the time interval $[t_{h-1}, t_h)$ with consequences of tier level $z$, $f_1$ is a function that maps $D_{h,z}$ to a value in $[0,1]$, which is typically defined by experts and $w_z$ is a weight:

$$w_z = \begin{cases} \frac{1}{3}, if\ h = 1 \\ exp\left(\frac{f_1(D_{h-1,z}) - f_1(D_{h,z})}{f_1(D_{h-1,z})}\right), if\ h \neq 1 \end{cases} \quad (4)$$

Then, the probabilities $P_h^H$, $P_h^M$ and $P_h^L$ of the healthy states "H", "M" and "L" of the barrier OI in the time interval $[t_{h-1}, t_h)$, $h = 1, \dots, N_T$, are computed as:

$$[P_h^H, P_h^M, P_h^L] = f_2(x_h^{OI}) \quad (5)$$

where $f_2$ is an expert-based model that maps the KPI to the set of probabilities of the barrier states.

In the current work, the possibility of automatically identifying the PSEs containing keywords associated to a given barrier is exploited. For example, in the case of the barrier OI, Equations (3) and (4) are modified considering $D_{h,z}^{OI}$, i.e. the number of PSE reports in the time interval $[t_{h-1}, t_h)$ with consequences of tier level $z$ that contain at least one keyword related to OI, instead of $D_{h,z}$, i.e. number of PSE reports in the time interval $[t_{h-1}, t_h)$ with consequences of tier level $z$. By this way, a more accurate and less conservative estimation of the barrier healthy state is obtained.

## 4. Case study

A repository $R$ of $D$ reports of PSEs occurred in a period of time $T = t_5 - t_0$ in hydrocarbon production plants is considered. The number $D$ of reports and the period of time $T$ are here not reported for confidentiality reasons. Each report is composed by:

*i.* a free text written in English by a system operator and containing the description of the PSE event;
*ii.* the date on which the report has been collected.

The period of time $T$ is sub-divided into five intervals of the same duration ($N_T = 5$) during which the subset of reports $\{R_1, \dots, R_5\}$ are collected.

## 5. Results

The method described in Section 3.2 is applied to extract the keywords from the PSE reports. The hyperparameters $thresh_{NPMI}$ and $thresh_{TFIDF}$ are set to 0.5 and 0.7, respectively, by applying the trial-and-error procedure described in (Valcamonico et al. 2023). The results are here discussed with respect to the barrier "Operating Integrity", which is involved in many PSE reports since it related to maintenance activities.

Figures 2 and 3 show the importance, estimated at time $t_5$, of two keywords related to the barrier "Operating Integrity" in reports occurred before $t_5$. The TFIDF metric $\phi_{i,j^h}$ (Eq.(1)), which does not consider the time dependency, and the metric proposed in this work $\tilde{\phi}_{i,j^h}$ (Eq (2)), which accounts for the time in which the reports were collected, are considered. As expected, $\tilde{\phi}_{i,j^h}$ assigns to the keyword "*cathodic_protection*", which is used only in the reports of two PSEs occurred significantly before $t_5$, a remarkably smaller importance than the TFIDF-based metric $\phi_{i,j^h}$, which does not consider the time of the events. This is in agreement with the intuition that if the word "*cathodic_protection*" has not been used by plant operators to describe events in the time intervals $[t_1, t_2)$, $[t_2, t_3)$ and $[t_3, t_4)$, the correct countermeasures have been taken in the past to resolve problems related to cathodic protection. On the contrary, the importance given by the proposed metric to the keyword "*replacement*" increases as time passes, since it has been used in several PSE reports in all time intervals, which may indicate a persistent problem to address.

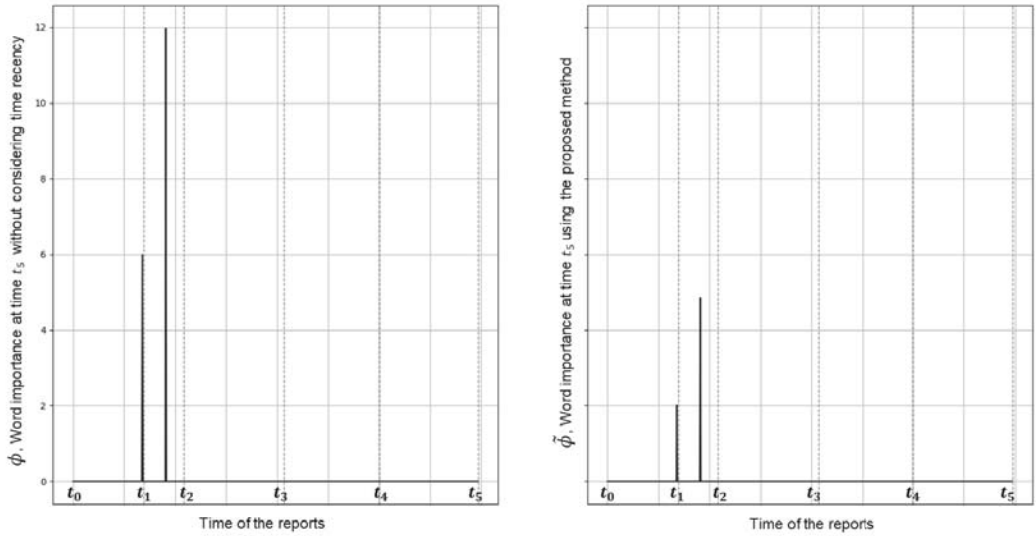


Fig.2: Importance of the keyword "*cathodic_protection*" computed with the measure in (Valcamonico et al. 2022) (left) and with the measure proposed in this paper, which considers the time dependency (right). The performance is computed at time $t_h = t_5$.

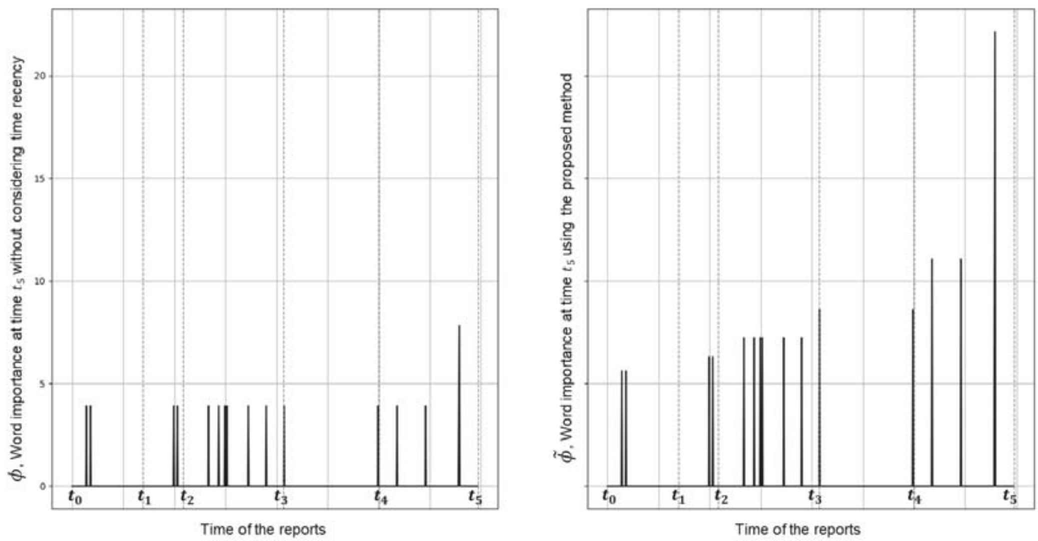


Fig.3: Importance of the keyword "*replacement*" computed with the measure in (Valcamonico et al. 2022) (left) and with the measure proposed in this paper, which considers the time dependency (right). The performance is computed at time $t_h = t_5$.

Then, the performance of the barrier "Operating Integrity" (OI) is estimated applying the method described in Section 3.3 using as inputs the number $D_{h,z}$ of PSEs of tier level $z$ occurred in $[t_{h-1}, t_h)$ (method in (Di Maio et al. 2021), Fig.4), and the number $D_{h,z}^{OI}$, of the PSEs with tier level $z$ occurred in $[t_{h-1}, t_h)$, whose reports contain at least one keyword related to the barrier OI (proposed method, Fig.5).

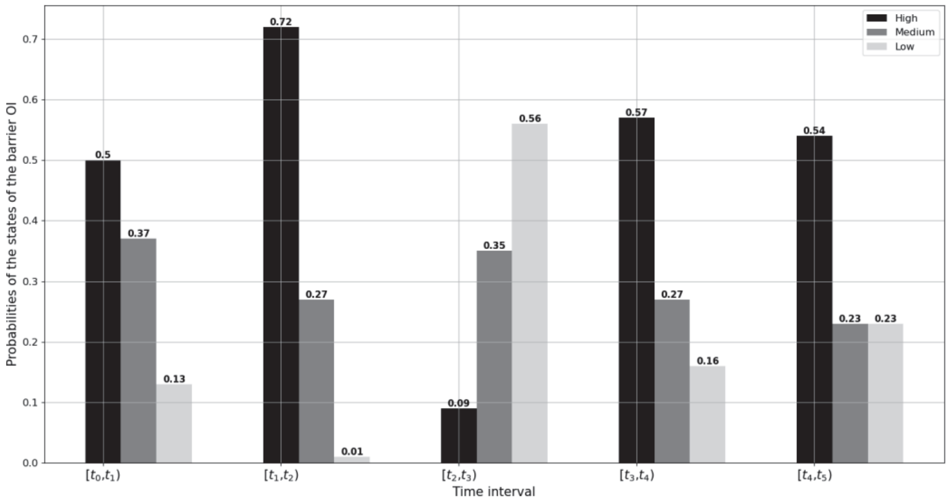


Figure 4: Probabilities of the healthy states of the barrier OI in different time intervals considering as inputs of the method the number $D_{h,z}$ of all PSEs of tier level $z$ occurred in $[t_{h-1}, t_h)$.

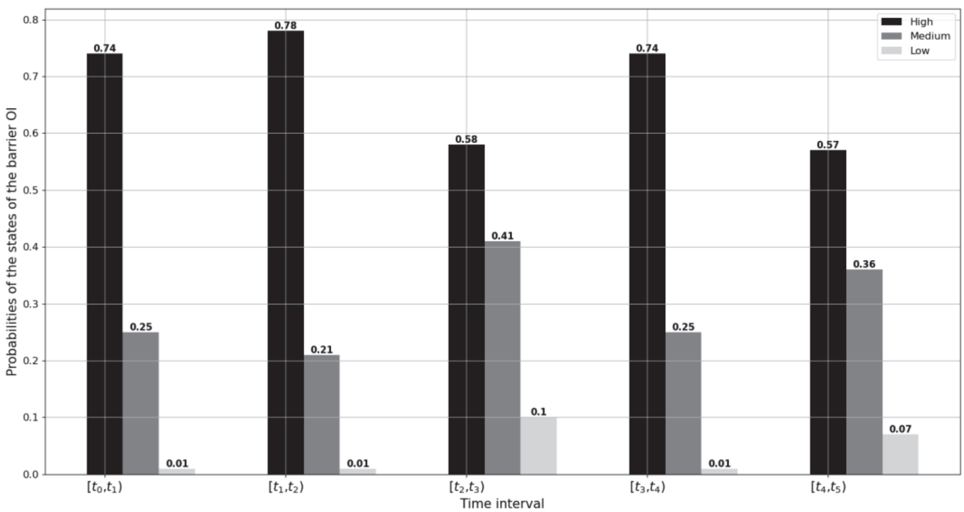


Figure 5: Probabilities of the healthy states of the barrier OI in different time intervals considering as inputs of the method the number $D_{h,z}^{OI}$ of the PSEs related to the barrier OI of tier level $z$ occurred in $[t_{h-1}, t_h)$.

The obtained performances of the OI barrier are shown by reporting the probabilities that the barrier state is “High” (“H”), “Medium” (“M”) or “Low” (“L”). It can be noticed that, when $D_{h,z}$ is used as inputs of the method, the performance of the OI barrier remarkably decreases between the time intervals $[t_1, t_2)$ and $[t_2, t_3)$ (P(H) decreases from 0.72 to 0.09 and P(L) increases from 0.01 to

0.56). This is due to the fact that the number of PSEs occurred in time interval $[t_2, t_3)$ is significantly larger than those occurred in $[t_1, t_2)$. The proposed method, which receives in input the number $D_{h,z}^{OI}$ of the PSEs involving the barrier OI, estimates a smaller drop of the barrier performance (P(H) decreases from 0.78 to 0.58 and P(L) increases from 0.01 to 0.10). Therefore, considering only the PSEs related to the barrier of interest instead of all PSEs allows reducing the overconservativeness of the method, so that the decision-making can be better informed on the individual barrier contribution to the asset risk.

## 6. Conclusions

A NLP method and a taxonomy have been combined to estimate the performance of the safety barriers of hydrocarbon production systems from textual reports of PSEs. The proposed methodology is based on the automatic identification of the textual reports involving the safety barriers of interest. The developed methodology has been applied to a repository of reports of PSE occurred in hydrocarbon production plants collected in sequential intervals of time. The obtained results have shown that the proposed methodology allows evaluating the performances of the safety barriers in sequential intervals of time. By this way, QRA is facilitated in following the modifications of the system in time.

Further developments of the methodology will be devoted to support the identification of the causes of the possible unsatisfactory performances of the barriers and propose countermeasures.

### References

"API RP 754." 2021. American Petroleoum Association (API) Recommended Practice Process Safety Performance Indicators for the Refining and Petrochemical Industries. 2021. api.org.

Bouma, Gerlof. 2009. "Normalized ( Pointwise ) Mutual Information in Collocation Extraction." In *Proceedings of GSCL 30*, 31–40.

Dayrat, Benoît. 2005. "Towards Integrative Taxonomy." *Biological Journal of the Linnean Society* 85 (3): 407–15. https://doi.org/10.1111/j.1095-8312.2005.00503.x.

Decarli, Luca, Luciano Scataglini, Paolo Cherubin, and Laura La Rosa. 2020. "Barrier Integrity Management : An Innovative Approach for Managing Process Safety Risks in Oil and Gas Assets." In *SPE International Conference and Exhibition on Health, Safety, Environment, and Sustainability*. https://doi.org/https://doi.org/10.2118/199472-MS.

Mackie, S I, M B Welsh, and M D Lee. 2006. "An Oil and Gas Decision-Making Taxonomy." In *SPE Asia Pacific Oil & Gas Conference and Exhibition*. https://doi.org/https://doi.org/10.2118/100699-MS.

Maio, F Di, O Scapinello, E Zio, C Ciarapica, S Cincotta, A Crivellari, L Decarli, and L Larosa. 2021. "Accounting for Safety Barriers Degradation in the Risk Assessment of Oil and Gas Systems by Multistate Bayesian Networks." *Reliability Engineering and System Safety* 216 (March). https://doi.org/10.1016/j.ress.2021.107943.

Marwah, Divyanshu, and Joeran Beel. 2020. "Term-Recency for TF-IDF, BM25 and USE Term Weighting." In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, 36–41. https://www.aclweb.org/anthology/2020.wosp-1.5.

Valcamonico, Dario, Piero Baraldi, Enrico Zio, Anna Crivellari, Luca Decarli, and Laura La Rosa. 2022. "A Taxonomy for Modelling Reports of Process Safety Events in the Oil and Gas Industry." In *Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)*, 3071–78. https://doi.org/10.3850/978-981-18-5183-4.

Valcamonico, Dario, Piero Baraldi, Enrico Zio, Luca Decarli, Anna Crivellari, and Laura La Rosa. 2023. "Combining Natural Language Processing and Bayesian Networks for the Probabilistic Estimation of the Severity of Process Safety Events in Hydrocarbon Production Assets." *Submitted to Reliability Engineering and System Safety*.

Weiss, Sholom M, Nitin Indurkhya, Tong Zhang, and Fred J Damerau. 2005. *Text Mining. Predicitive Methods for Analyzing Unstructured Information*. Springer.

Zhang, Laibin, Shengnan Wu, Wenpei Zheng, and Jianchun Fan. 2018. "A Dynamic and Quantitative Risk Assessment Method with Uncertainties for off Shore Managed Pressure Drilling Phases." *Safety Science* 104 (December 2017): 39–54. https://doi.org/10.1016/j.ssci.2017.12.033.