 doi:10.20944/preprints202106.0042.v1

Article

Near Miss Archive: a Challenge to Share Knowledge Among Inspectors and Improve Seveso Inspections

Silvia Maria Ansaldi [1]*, Patrizia Agnello [1], Annalisa Pirone [2] and Maria Rosaria Vallerotonda [2,]

[1] Inail, Dipartimento di Innovazioni Tecnologiche, Monte Porzio Catone, Rome, Italy; s.ansaldi@inail.it; p.agnello@inail.it
[2] Inail, Dipartimento di Innovazioni Tecnologiche, Rome, Italy; a.pirone@inail.it; m.vallerotonda@inail.it
* Correspondence: s.ansaldi@inail.it

Abstract: In European Seveso Legislation for the control of the hazard of major accidents (Directive 2015/12/UE), the Safety Management System SMS is an essential obligation for managers and the authorities are required to periodically verify its adequateness through periodical inspections at Seveso sites. One of the pillars of the SMS is the collection and analysis of documents on accidents, near misses and possibly anomalies, in order to identify weaknesses and implement continuous improvement. In Italy, for a few years, the documents, gathered from all Italian Seveso sites by the inspectors, have been archived and used for research purposes. The archive currently contains some 4000 reports, collected in five years by some 100 inspectors throughout Italy. The paper discusses in the detail the challenges faced to extract the knowledge hidden in the documents and make it usable through the design of a robust model. For this aim, Machine Learning techniques have been used as a preprocessing of the reports for extracting the concepts and their relations, organized into an entity-relation model. The effectiveness of this methodology and its potentiality are pointed out by investigating a few hot topics, exploiting the information contained in the repository.

Keywords: near miss modelling; sustainability; Machine Learning.

1. Introduction

The analysis of near misses is a need of those sectors, including aeronautics, nuclear, chemical and petrochemical, where accidents are not frequent but when occur, have catastrophic consequences [1]. Thus, the study of near misses and anomalies provides an opportunity to recognize unsafe conditions or situations and prevent incidents [2]. For this reason, in the literature there are several definitions of *near miss*, depending on the context, but all of them have a general meaning that is intuitive, understandable and suitable, that a *near miss* is unplanned and unexpected event without consequences in injury, illness, damage, and environmental problems, but had the potential to do so.

A source of knowledge is the analysis of near misses collected in industries under the European Seveso Directive on the control of major accident hazards. The Seveso III Directive 2012/18/EU [3] requires the implementation of a Safety Management System SMS for establishments handling hazardous substances. One of the pillars of SMS is to gather and analyze reports of accidents, incidents, near misses, and anomalies occurred in the establishment, with the aim to point out the weakness in the SMS for its reviewing and improvement.

In Italy, there are about a thousand establishments under Seveso legislation, over half of them are upper tier. They cover many industrial sectors, including oil and gas, petrochemical, chemical, metal processing, pharmaceuticals, explosives, and pyrotechnic.

Since 1999, the Italian Competent Authorities adopted a standard guide for the inspections at establishments under Seveso legislation, based on a detailed checklist. The Italian inspectors are required to discuss the anomalies, near misses, incident reports,

 doi:10.20944/preprints202106.0042.v1

provided by the operators, with the goal of prioritizing the scrutiny of the points in the checklist.

In Italy, the latest implementation of Seveso Directive on July 2015 encloses a form to fill in anomalies, near misses, and accidents by the operator. The set of documents is called operative experience,. Each document contains the description of the events, occurred in the establishments in the last ten years, the analysis that identifies the critical points of the SMS and the solutions adopted for the safety improvement.

In Italy, the practice of analyzing near miss events for improving the efficiency of the SMS, presented years ago by few pioneers [4], has been adopted in a systematic mode by a Inail research group. Since 2015, this group, in which researchers are also Seveso inspectors, started to collect the operative experiences, sent by Inail inspectors who operate throughout the national territory. The reports are organized and uploaded into a central archive and managed for research purposes.

The systematic activity of collecting those reports has significantly increased the amount of documents in the operative experience repository. The reports have heterogeneous content, because describe different types of events (i.e. anomalies, near misses or incidents) and deal with diverse categories of equipment, substances, processes, and working activities.

The definition of near miss already mentioned is very general. In Seveso context, the Italian standard UNI 10617:2019 [5] gives the following definition: "*major near miss is any extraordinary event that could have turned into a major accident. The difference between a major accident and a major near miss does not lie in the causes or modalities of evolution of the event, but only in the different degree of development of the consequences or in the randomness of the presence of items or people*".

In some cases, that definition matches with events that may be precursor of accident [6] i.e. an item in known incidental chain, and their analysis allows the investigation of the causes that triggered the incidental mechanism, usually hidden from relevant consequences. The near misses, reported in application of the Directive aforementioned, include also minor incidents without injuries, anomalies, malfunctions and deviations from the normal operation of the equipment or processes. A few of these occurrences would appear meaningless, but their analysis can contribute to discover safety weakness related to new or emerging risks. This highlights another point of view for near miss definition, as "an opportunity for surveillance and risk reduction" [2], to improve human, environmental, and process safety [7].

Based on the Seveso near miss repository, several types of study may be developed, from the statistical analysis on occurrences to the extraction of information up to learn lessons. The inspector may exploit this archive to improve the inspection, relying, in addition to his/her own expertise, also on operational experiences that have occurred in establishments similar to the one being inspected. On the other hand, researchers can address their research activities to more recurrent issues for improving the solutions, but also face new or emerging risks already occurred although not frequent. The objective of this paper is to describe the methods adopted to extract knowledge from this repository and provide the inspectors and researchers with information or pills of insights.

The challenge, which is also the objective of this study, is to extract the knowledge contained into the documents and make it usable. Thus, it is first necessary to have tools capable of managing texts in natural language. Among the Information Technology systems adopted for processing natural language, Text Mining TM is the most appropriate method to automatically analyze and classify the parts of speech. Recently, more advanced techniques of Artificial Intelligence AI, including the Machine Learning ML processes, greatly improve the automation of analyzing large amount of data.

This article is structured as follows. Section 2 describes the background addressing the management of incident and near miss reports and different approaches for extracting knowledge. Section 3 explains the objective of this research. Section 4 details the peculiarities of near miss archive considered and describes the methods, including Text mining, Machine Learning, and Artificial Intelligence functionalities, adopted for

 doi:10.20944/preprints202106.0042.v1

managing unstructured natural language text and extracting knowledge into entity-relation model. In Section 5, a few use cases describe the results of the model application and related discussions. Sections 6 and 7 provide with general discussion and few concluding remarks, respectively.

2. Background

For many years now, the reports of major accidents occurring in the chemical process industries have been systematically analyzed to learn from what happened and avoid the repetition of the same conditions and unfavorable situations. Thus, there are open databases that gather the accidents reports occurred in industrial chemical sectors.

The Major Accident Reporting System (*eMARS*) [8], established by the EU's Seveso Directive 82/501/EEC since 1982, is the official European database, it collects accidents, incidents and near misses occurred in European Seveso establishments. ARIA (Analysis, Research and Information on Accidents) database (in French and English) organized by Bureau for Analysis of Industrial Risks and Pollutions in France, "catalogues incidents or accidents that were, or could have been, deleterious to human health, public safety or the environment" [9]. Other sites collecting industrial accidents are the following: Zema, the German database of Major accidents and incidents [10], the Chemical Safety Board in the United States [11], Tukes Varo registry in Finland [12], and the Japanese Failure Knowledge database (in Japanese and English) [13].

The study of major accidents aims at understanding and above all learning as much as possible from what happened, but, unfortunately, some causes could be covered by relevant consequences.

Looking at near misses as potential accidents intercepted and interrupted by chance, luck or skill provides the opportunity to analyze them with respect to the Safety Management System SMS, for identifying its weakness points, moreover the lack of consequences guarantees a more open narrative without fear of repercussions. Thus, in complex sectors, including chemical and petrochemical industries, where major accidents can be catastrophic events with serious consequences for human health, environment, and assets, the analysis of near misses and anomalies is strongly encouraged. The study and analysis of near misses leads to the rediscovery of hidden or forgotten knowledge and the learning of how to improve safety [14]. Although the analysis of near misses is an adopted and consolidated approach in major hazard industries, however, it is also developing in other sectors where the frequency of severity is very high, e.g. the construction sector. Thus, [15] explores the potentiality of using near miss information for improving construction safety performance; the objective is to fill the gap from the theoretical definition of near miss and its practical understanding for better identify the causes and its process management.

The accidents reports (e.g. those contained into eMARS database) usually contain keywords for better classifying the events, their causes and their consequences; they are written by experts in a technical language that is understandable to all the community with the aim to learn the lessons and avoid recurrence of the same events. By reading the narrative of an accident, however, human experts are able to extract even information and concepts not explicitly represented by the keywords, but contained within the text, including cause-effects relationships.

Natural Language Processing NLP applications aim to extract information contained in the accident story. Single et al. [16] use NLP techniques to extract information from eMARS accident database and insert them into an ontology, used to represent the knowledge base for enquiring purposes. In [17] is described the approach adopted by the authors to extract meaning from multi-lingual free-text safety incident reports in railway transport. The approach is to import the text into a graph database and connect with an ontology for managing multi-languages using NLP techniques. Nakata [18] proposes a method for recognizing typical flow of events in a large set of accident reports, by

 doi:10.20944/preprints202106.0042.v1

adopting text-mining capabilities, focusing on adjacent sentences, and extracting the pair of predecessor and successor words that characterize the flow.

Near miss reports are quite different from accidents reports, although they deal with similar matter, but may lack of systematic view of the events. In fact, near misses are detected by workers and usually registered by supervisor or owner, and tell about the facts and the direct causes, the actions done and those that would be planned. The objective of their analysis is also to learn the lessons, aiming at workers to correct unsafe conditions and situations, and report critical issues of daily operations, as described by Bragatto et al [19].

Although in some reports of large organizations predefined keywords also appear, useful for organizing statistical frequency analyzes more quickly, the textual story always remains the most interesting and complete aspect. For this reason, the analysis of the near misses requires tools able to interpret and process the natural language used for their description. The usual methods adopted were based on pre-defined taxonomies of the most important concepts involved, including substance, equipment, people, and process activity. Thus, the near miss analysis tried to classify the event and representing elements with the items contained in the taxonomies [20].

Using the same items for representing both accidents and near misses has been a method of understanding and measuring the distance of near misses from major accidents. Ansaldi et al. [21] describe how applying similarity techniques to documents, for measuring the semantic distance of near misses with respect to an accident report. This method is applicable and effective when managed data are few and homogeneous, that is coming from the same plant or the same industrial sector, and the taxonomies can be defined a priori because use the same terminology.

Two changes in the near misses management have made this a priori method difficult to adopt. One is the cultural change from blame approach toward safety awareness, from attitude to hide the negative events toward a greater sensitivity to record all anomalies and deviations from normal situations, near misses, also highlighting the positive aspects that stopped their escalation; this has greatly increased the amount of reports. The other aspect is that near misses, because they can be means of identifying weakness in the safety management system, are aimed mainly at the workers of the plant, thus they use the language and jargon common to the sector and plant itself.

On the other hand, for extracting worth information and sharing among the stakeholders, it would be efficient to arrange all reports, recorded from each establishment, into a single repository. In this way, however, the "a priori" definition of taxonomies would require continuous updates, additions and checks of new terms or synonyms used in several industrial sectors and in different jargons, with the risk of omitting important terminologies.

In recent years, Artificial Intelligence AI techniques, including Machine Learning, are becoming stronger to face issues related to huge amount of data, including the improvement in NLP field that involve text data, e.g. technical documentations, and are spreading in many industrial sectors. The AI techniques, especially Machine Learning ML technology, are challenging for extracting information from bulk of data; thus, together with text mining tools would be able to work on amount of documents for eliciting knowledge.

Cheng et al. [22] show a comparison of different algorithms, based on NLP techniques, for extracting knowledge from construction accident reports and classifying the narratives. Arteaga et al. [23] address the issue of analyzing reports on the severity of traffic crashes and extracting meaningful information for developing safety countermeasures. Kurian et al. [24] apply ML and keywords analysis for defining a customized library that more efficiently supports ways to report incidents; special attention is for those with minor consequences that often lack of details useful for understanding the causes. Paltrinieri et al. [25] suggest a risk assessment approach that is based on machine learning techniques, including deep neural network model.

Xu and Saleh [26] provide a detailed overview of different ML categories and corresponding models and algorithms used, reviewing ML applications in reliability and safety

 doi:10.20944/preprints202106.0042.v1

applications. They also give a rough definition of ML, as a data analysis method that iteratively learns from past data and adapts independently when applied to new data. The authors also "believe a most promising application of ML is in unleashing its power to harvest more value from near miss data and other accident databases for ultimately improving accident and occupational injury prevention".

3. Objectives

The first aim of this work is to provide the Seveso inspectors with a valuable knowledge resource, so that to improve the quality of the Seveso inspections. The sharing of what happened in different establishments, for example, to the same type of equipment, or in the same process with a certain substance although with the involvement of different parts of an equipment, has many advantages. Among these, it allows inspections to be carried out according to a more homogeneous criterion throughout the national territory. It also let identify the possible solutions to be adopted by considering those that have given a better outcome or those that are more adequate in similar situations, highlighting which barriers have worked best and which ones proved to be lacking.

A more general goal is to understand the weakness of the safety management in Seveso industries, and addressing efforts in those directions. It has been at least a couple of decades that near miss analysis has been used to better understand whether the risk assessment, that has been carried out in the plant, actually includes all the possible triggers that can lead to top event and then to a possible accident scenario. On the other hand, an event, whose probability of occurrence was considered too low to be included in the incidental chain, has instead proved, by the analysis of operative experiences, to be more plausible. In this case, the most remarkable feedback is the improvement of safety procedures and the update of risk analysis by extending it, where appropriate, with the examination of new critical items.

With the evolution of technologies, operative contexts and working methods, even the risks, whether traditional or emerging, face a change. Indeed, the traditional risk are affected by the new context, while emerging risks represent a novelty. In both cases, an accurate analysis of what is reported in the near miss reports certainly allows an early identification of elements that could represent unexpected hazards up to that moment.

4. Methods

The method adopted is Text Mining with Machine Learning capabilities to extract concepts and modeling them into a knowledge representation.

The aim is not only to identify the terms but also to recognize their semantic and above all their relationships. The semantic recognition of the single words is not sufficient to understand the story; in fact, a term may be present in the document without having a direct role in the event. For instance, in the phrase "*leakage from the drainage valve of the suction pump used for the tank*", the *tank* is not a primary term in the description of the near miss, so it would not be correct to count such an event as an occurrence to *tank* entity.

The definition of the model for representing the knowledge of the near misses is the core of this research. On this model, called *EsOpIA* (Operative Experience and Artificial Intelligence), the text mining and the Machine Learning techniques provide the capabilities for the extraction and classification of the concepts, and modelling them into the knowledge base. The following subsections describe the *EsOpIA* model (entities and relations) and the AI techniques applied for extracting the knowledge.

4.1. Operative Experience reports

The operative experience documents, collected during the Italian Seveso inspections, tell about the anomalies, near misses and minor incidents that occurred at the establishment during the previous decade.

Each report, written in natural language, i.e. Italian, adopts the standard format provided by the Italian Seveso legislation, whose fields contain information related to the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

description of the event, the recovery activities undertaken and the follow-up actions. The description is the narrative of the event occurred and highlights the substance and equipment involved, the technical devices or the procedures that failed or misapplied, as well as those that stopped the escalation of the occurrence, avoiding the consequences or mitigating their effect.

In spite of using the same format, the reports are compiled differently for the accuracy of the description and the detailed information recorded. The interpretation of operative experience concept is also different from one establishment to another. At few establishments, just release of hazardous substances is recorded; in other cases, anomalies, unsafe conditions and situations are detected, also those not related to major accident hazards. This diversity represents truthful pictures of the events occurred into establishments and depots, but increases the complexity of extracting knowledge from those reports.

The precious information of those documents is in the story itself, in a few sentences the report tells *what* happened, *what* are the elements involved (equipment, substances, people), *what* failed and *what* succeed. Therefore, for this research, knowledge extraction means to represent the story into a mathematical model.

4.2. EsOpIA Model

4.2.1. Entities

Since the model must represent the story contained in each near miss document, its definition has to reflect the concepts that best describe the facts. The concepts were identified by answering simple questions, including: *what*, *when* and *where* did it happen? *Who* and *what* were involved? *What* stopped the escalation, and *what* failed?

The first question identifies the key elements to identify *what* happened and give it a place in space (*where*?) and time (*when*?). Thus, the entities are, respectively, *event*, *industrial sector*, and *date*. The second question identifies the persons, *who* were involved in the story (entity *people*), as well as the equipment or a part of it (*apparatus*) concerned to the event, the *substance* involved, and, eventually, the type of work (*activity*) undertaken when the event occurred. During the Seveso inspections, the operator provides the inspectors with the list of technical and organizational measures (*barrier*) adopted for preventing the accidents and for mitigating the consequences when undesired events occurred. Thus, the third question refers to an entity *barrier* that failed or succeeded in the near miss occurrence.

The identified concepts are used for classifying the terms extracted from the documents. Indeed, they have a broad meaning and often would require further specification for a more effective knowledge representation. On the other hand, a more detailed specification could make the Machine Learning applicability difficult in the classification process; therefore, a balance between keeping some details and ensuring the success of ML is the strategy adopted.

Therefore, the entities, including *substance*, *people*, and *activity*, are not further specified, while *apparatus*, *barrier*, and *event* are classified into several subclasses. The *apparatus* is subdivided into *equipment* and its parts, i.e. *component*; thus, tank is classified as *equipment*, while flange is a *component*. *Barrier* is split into two subsets, the *technical* and the *organizational* barriers. Thus, the level gauge, a safety physical device, is *technical barrier*, while permit to work and instructions for loading/unloading operations are procedures or technical instructions classified into *organizational barrier*.

The *event* entity is the core of the story of near miss, without it the near miss or anomaly would not exist; therefore, its subdivision into several classes provides greater and useful specification of the concepts. The subclass *loss* collects the terms related to a loss of containment, e.g. leakage, overfilling, overflow; *failure* gathers the mentions dealt with any breakdown, malfunctioning, damage of machineries or devices, but also wrong behaviors or errors in working activities. The defects of equipment that would cause integrity problems (e.g. corrosion, erosion, pitting, holes) or less efficiency (e.g. occlusion, lack

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

of elasticity, fouling) are grouped into *deterioration* subclass. The near miss archive contains also incidents and a few accidents; thus, a subclass defines *major* events.

The groups described above represent negative occurrences, what was wrong, but it is also important to point out the actions or the circumstances that succeed to interrupt and block the event escalation, to notice unsafe conditions in an early phase, or promptly to stop working activities; thus, *success* subclass contains those terms.

4.2.2. Relations

The identification and classification of the terms in a text are not exhaustive for representing the knowledge contained in the document. In a sentence, the words that are parts of a discourse, further to their meaning, may have a role or be irrelevant in the story. This ambiguity can be solved by relating concepts to each other.

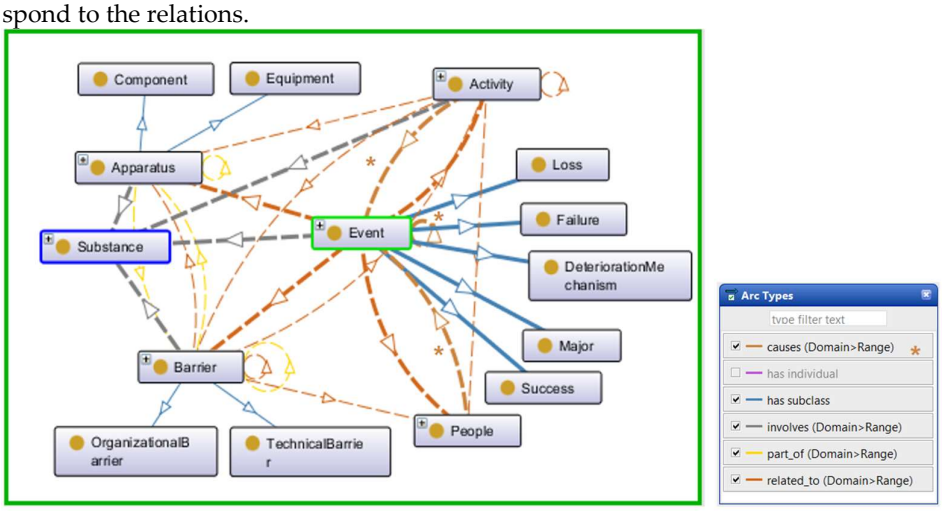
The relations designed in the model are the following: *related_to*, *part_of*, *involves*, and *causes*. The *related_to* is a weak link putting two elements in relation, without adding a specific type. While the other three implicitly provide a meaning to the connections between the terms, *part_of* links a physical component to an *equipment*, *involves* relates an element to a substance, the relationship *causes* points out something (*event*, *activity*) or someone (*people*) that led to an event. These three relations are different from the first one also because they are oriented connections, that is, they have a direction from one concept to another.

The Table 1 shows the model elements, the relations put in correspondence with their entity types in a predefined order.

Table 1. List of relations in correspondence with their entity types.

Relation	First Entity	Second Entity
CAUSES	EVENT, ACTIVITY, PEOPLE	EVENT
INVOLVES	ACTIVITY, EVENT, APPARATUS, BARRIER	SUBSTANCE
PART_OF	BARRIER, APPARATUS	APPARATUS, BARRIER
RELATED_TO	EVENT, BARRIER, ACTIVITY	ACTIVITY, PEOPLE, BARRIER, APPARATUS

The Figure 1 shows the graph, designed with Protégé[1], corresponding to the conceptual model, the rectangles describe the entities (classes and subclasses) and the arcs correspond to the relations.



[1] https://protegewiki.stanford.edu/wiki/WebProtegeUsersGuide

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

Figure 1. The *EsOpIA* model: the rectangles correspond to classes and subclasses, the blue lines indicate the parent-child relation (*has subclass*), the dash lines correspond to specific relationships, including *related_to, involves, part_of,* and *causes*, as indicated in the legend in the right side.

4.3. AI techniques for extracting knowledge

The methods adopted for extracting knowledge by near miss archive are based on text mining capabilities, which is used for analyzing the parts of speech and extracting the tokens (words). Text mining is the process for eliciting information from an unstructured and free form text, by analyzing the parts of speech and classifying them with appropriate meaning.

The token classification and their relationships defined into the conceptual model are processed with the support of Machine Learning techniques.

In this context, *data* is the text describing the narrative of near miss or anomaly, and the task of ML techniques is to learn how to extract, classify the terms and define their relationships by referring to the designed *EsOpIA* model.

The ML application is an iterative process, characterized by two phases: the training for building the machine-learning model and the evaluation of its performance.

4.3.1 Machine Learning model construction

The adopted ML system is based on techniques for annotating the terms contained in the text and defining their relationships. In the jargon of text analysis process, "annotation" is the technique of assigning a type of the entity model to a term or a part of speech that is to provide it with a meaning or semantics.

The goal is for the system to learn to annotate the text correctly, thus a team of experts has supported this operation. At each step of this iterative process, experts manually annotate a set of documents using the ML tool adopted for developing the project (IBM Watson Knowledge Studio [27]).

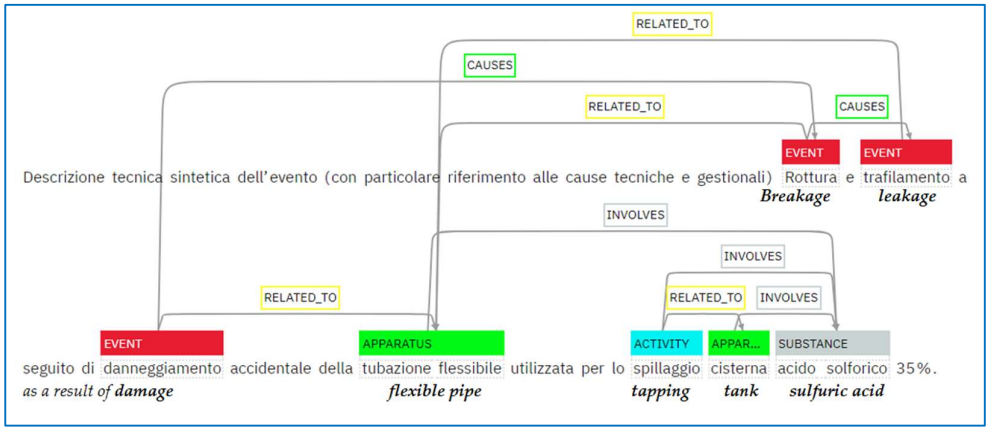


Figure 2. The extraction from the description of a near miss. The words in English indicate the annotated mentions, classified as *event* (red), *equipment* or *component* (green), *activity* (blue), and *substance* (gray), respectively.

The Figure 2 shows an example of annotations and their relationships. The documents are in Italian, but in the figure, the key terms are translated into English to facilitate reading. According to *EsOpIA* model, the colored boxes correspond to the different types of entities, including the classes: *event* (red), *apparatus* (green), *activity* (blue), and *substance* (gray). The other boxes with colored outline represent the relationships and the lines link the mention items.

The statement is a brief description of the event occurred, that is "*breakage and leakage as result of damage to the flexible pipe used for tapping the sulfuric acid from the tank*". The entity annotator classifies *breakage* and *damage* as terms belonging to *failure* subclass of *event* class, while *leakage* is in *loss* subclass. The terms *flexible pipe* and *tank* are annotated as *apparatus*,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

members of *component* and *equipment* subclass, respectively. The *activity* in progress at the time the event occurred is the *tapping* of *sulfuric acid* (entity *substance*).

The relations *related_to* and *involves* are quite simple to be defined, while *causes* relation must take into account the order between the entities. Indeed, reading the statement, for humans is easy to understand that the damage has caused a breakage with a leakage consequence. The challenge is to train the ML system to learn this reasoning in order to classify correctly the cause relations.

4.3.2 Performance of the machine-learning model

Following the overview provided by Xu and Saleh [26] on ML methods, based on their capabilities and features, the ML adopted in *EsOpIA* has the characteristics of a *supervised learning*, since the aim for the system is to learn a target function that can be adopted to predict the values of a class. The annotation process, as described in [28], for training *EsOpIA* model required about fifteen iterations, with sets of documents ranging from 5 to 10 in the starting phase, up to 20 and 50 in the most advanced stages of learning. At each phase, the system evaluates the test model through some metrics, usually adopted by ML techniques [23], including *precision*, *recall*, and *F1 score*.

Defining *TP*, *TN*, FP and *FN* the number of *true positive*, *true negative*, *false positive* and *false negative* outcomes, respectively, each mentioned metrics are computed as follows:

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ score = \frac{2\ (precision * recall)}{(precision + recall)} \tag{3}$$

The measure *precision* (1) is a predictive value computed as the number of correct positive results divided by the sum of true positive and false positive values. While *recall* (2) is a measure of sensitivity, determined by the number of true positive results divided by counting the positive results that should be returned (*TP* + *FN*). The *F1 score* (3) is interpreted as a weighted average of *precision* and *recall* values, whose best value is 1 and the worst is 0.

The Table 2 shows the statistics of test set of the deployed model for each entity.

Table 2. The statistic measures computed for each entity types of the EsOpIA model.

EsOpIA ENTITY TYPES	F1	Precision	Recall
ACTIVITY	0.64	0.78	0.55
APPARATUS	0.76	0.83	0.7
BARRIER	0.64	0.79	0.54
DATE	0.96	0.96	0.96
EVENT	0.79	0.84	0.74
PEOPLE	0.75	0.79	0.71
SUBSTANCE	0.77	0.85	0.7

The formula (1) enhances that a high *precision* value means that all citations that are annotated as a certain type of entity really belong to that classification. The Table 2 shows that *DATE*, *APPARATUS* and *SUBSTANCE* entities are annotated with the highest *precision* values, but also the other mentions have a high level of correctness. Thus, we are quite confident that the system is able to classify correctly the annotated concepts.

A high value of *recall*, the formula (2), means that all citations that should be annotated as a certain type of entity really are. The Table 2 shows satisfactory values (greater than 0.7) for many of the types of entities, but two are just sufficient, i.e. *ACTIVITY* and

 doi:10.20944/preprints202106.0042.v1

BARRIER. One explanation is that both of these types of entities have terms that can be classified as other types (homonyms), and the system is not always able to classify correctly maybe because sentences are too short from which it is difficult to deduce a more explicit context. For example, the same term *block* is used as an event (*FAILURE)*, an action to interrupt something (*ACTIVITY*), or a technical mechanism (*BARRIER*) for preventing undesired events. Thus, in case the sentence is short the system may have difficulty in classifying it correctly.

The Table 2 shows the values of the deployed model, but during the iterative process of ML techniques, lower values have required specific interventions to improve performance, including new documents to be annotated, choosing them from those containing the most ambiguous terms, but also coordinating and making the choices of human annotators converge in the same solutions.

4.3.3 Application of other AI functionalities

This section briefly mentions other AI functionalities adopted for cleaning the document repository and optimizing the management of model and their terms.

One of the critical points faced in managing near miss archive has been to assure the content anonymization. Indeed, proper names of people or companies have to be removed by the text, but this operation, unthinkable to do manually, takes advantages by applying text-mining capabilities for recognizing the undesired terms and remove them.

Other activities deal with organizing the extracted terms by considering the synonymies and lemma, or discarding the parts of speech not useful for the model. All these activities are strictly related to the characteristics of the language used in the documents; in our case is the Italian, so we think it would be tedious to mention all the details that would probably be different for other languages.

It is important to outline, however, that, in this project, the stop words identification is not a pre-process of text mining, but rather a post-process for removing words meaningless with respect to the story; however, this activity requires caution. The prepositions are usually considered stop words, but in Italian, they often are important parts to describe a concept; for example the *loading arm* is literary translated into Italian as "*arm of loading*", thus, if the preposition *of* is removed as a stop word, the concept is meaningless or worst is split into two concepts: *arm* (*component*) and *loading* (*activity*).

The lemmization and synonymy have been manual operations performed on the list of words extracted and classified by the system, whose outcomes have decreased the number of entities; more than 33000 terms have been reduced until to about 1500 words in normal form; the ML model counts also more than 27000 relations.

4.4. EsOpIA application

The *EsOpIA* application is a tool to access and query near miss repository, it aims at people who are working on this matter, both experts and trainees for Seveso inspections. The user can express queries in natural language that is Italian, the language used in the reports, with the chance of selecting the entities from the model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

The search mode may have two starting hypotheses: the first is beginning from a consolidated knowledge to verify if there are still operational experiences that confirm it or not; the second supports intuitions or foresights to understand if they match with real cases [29].

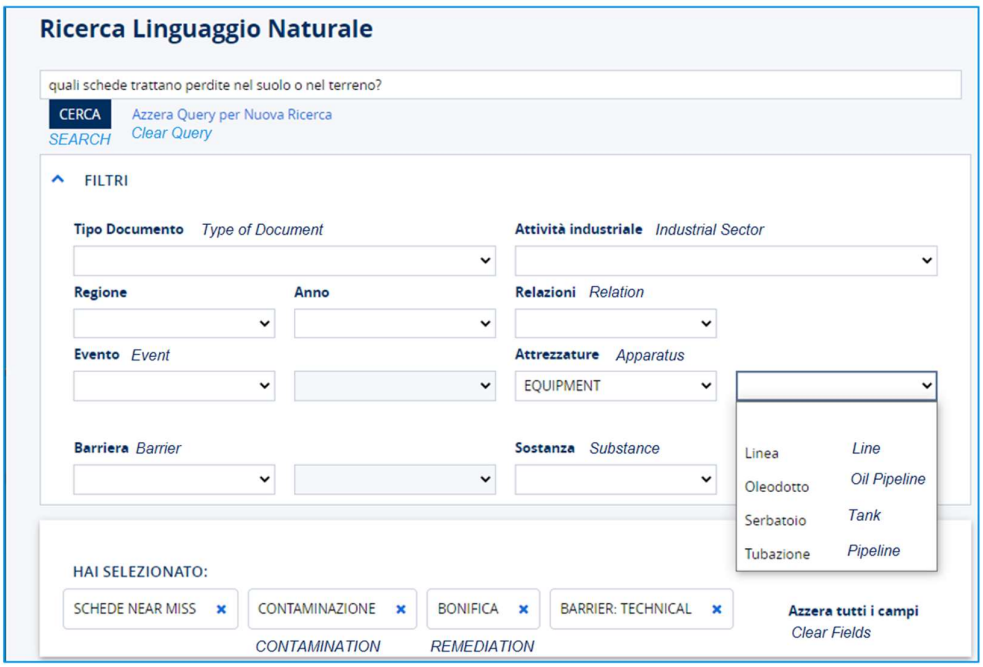


Figure 3. *EsOpIA* –Natural Language Search: functionalities with queries in natural language and filters from the model (Some captions and terms have been translated in English).

The Figure 3 shows the user interface panel of *EsOpIA*: at top, there is the query in natural language, than the filter section containing the terms classified according to the *EsOpIA* model. After running the search, the system updates the lists, loading only the terms contained in the documents found, making the search refinement easier for the user. The picture shows the *EQUIPMENT* combo listing only the items extracted in the outcomes, including *line*, *oil pipeline*, *tank*, and *piping*. The bottom of the panel shows the choices previously selected.

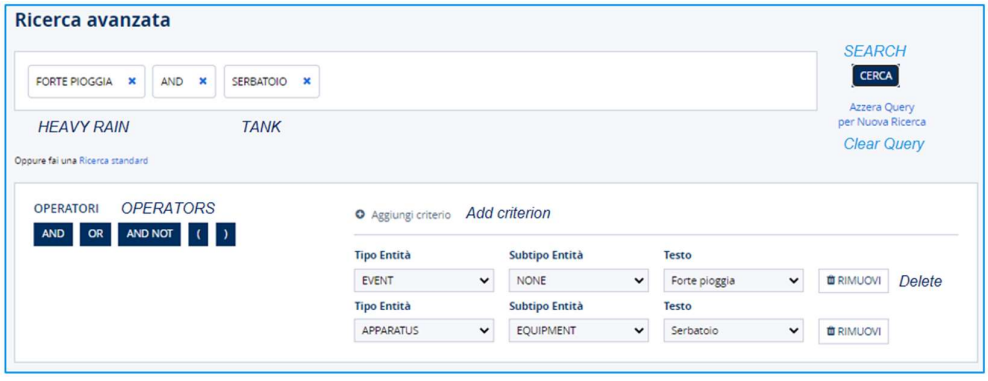


Figure 4. *EsOpIA* – Advanced Search: functionalities based only on the selection of the entities of the model (Some captions and terms have been translated in English).

The *EsOpIA* application also provides the functionalities to query directly the model; the user, therefore, selects the terms from the lists of the entity types and combine them with Boolean operators. Figure 4 shows the panel of Advanced Search, the example looks for the following query:

heavy rain AND tank

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

The items are classified as *EVENT-NONE* and *APPARATUS-EQUIPMENT*, respectively, the *AND* operator means to look for reports that contain both entities.

The search of term is, of course, extended to the synonyms associated to each entity.

5. Results

This section describes a few hot topics by extracting some pills of knowledge contained in this archive using the different approaches, above described, to run the search activities.

The first two case studies deal with known issues, the difficulties involving the permit to work and the loss of containment in ground; the aim is to understand if those problems persist despite the efforts made to ensure work safety and to limit containment losses. The third case study starts from an intuition by looking at some terms contained in the model that are words apparently out of context, but since classified in the *EsOpIA* model by the ML, are therefore interesting for safety purposes.

The following sections describe, for each case study, the most significant searching steps developed with EsOpIA application and their outcomes. The reports are written in Italian, but to make the reader understand the results, the entities are translated into English and listed in tables together with their classes and subclasses.

The editing types, adopted for the tables, have the following meaning: all model components are in *italic*, the names of classes, subclasses and relations are in uppercase, the individuals (i.e. terms) are in lowercase.

Since *EsOpIA* model has an entity-relation structure, each model extracted from the report is a set of (connected or disjoint) triples, i.e. (*entity*, *RELATION*, *entity*); thus, sequences of triples, described in the following sections, provide the representation of the natural language text into a mathematical model. The terms in bold correspond to the words used in the discussion.

5.1. Case Study #1: Risks known in managing working activities

The first case study faces the issue related to the Permit to Work PtW, to check if its management has been directly involved in some events.

PtW is a document addressed to third-party companies or internal workers who must execute activities of maintenance, improvement or changes inside the plant. Agreed between the operator of the establishment and the external company, it is a written document specifying, among other things, responsibilities, means, times, interfaces, intervention limits, precautions (including Personal and Collective Protective Equipment), reports. PtW is a systematic and formalized tool that collects information to carry a work out in full compliance with safety, must take into account all the risks of the working activity but also the conditions and situations in which it takes place, to indicate, therefore, the preventive and protective measures to be adopted.

Searching "permit to work" as free text in *EsOpIA* returns also the reports that refer to PtW as a procedure correctly compiled and followed. Looking for cases in which PtW failed, because misapplied or wrongly filled in, means to check among the model entities extracted from the reports, including *EVENT* and *BARRIER*.

Table 3. List of terms related to "Permit to Work" query.

Class	Subclass	Terms
EVENT	*FAILURE*	***error***, ***lack***, ***missing***
	SUCCESS	*stop*, *protecting from harm*, ***awareness***
BARRIER	*TECHNOLOGICAL*	*gas detector*, ***fire extinguishers***, ***explosivity detector***
	ORGANIZATIONAL	*emergency procedure*, ***maintenance***, ***controls***, *inspections*, ***training***, ***PtW***
PEOPLE	*NONE*	***supervisor***, *maintainer*, ***third-party company***

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

The words listed in Table 3 as failure events summarize general concepts, but the terms extracted from the reports are more detailed. Thus, *error* corresponds to error of *application, intervention, operational, compiling*; *lack* is lack of *analysis*, *supervision*, *preventive and end-of-work checks*, or more seriously, the PtW was *missing*.

A frequent error, extracted from the reports, is the miss *application* of PtW for delivering the plant, after the maintenance operations, in correct operating conditions. In one case, in fact, there was a solvent leakage due to lack of blind disc on the end of the line, in another the operator opened a wrong valve connected to a provisional line, on which the blind flange had not been mounted, both cases occurred after maintenance works. The extracted model highlight the relation *incorrect application RELATED_TO permit to work*.

Another report describes the release of product from the manometer detachment on the pump when the systems restarted after a general stop, since the threaded plug was not applied correctly. In this case, verifying the restoration of standard conditions following a maintenance intervention failed. The model representing the follow-up actions are:

awareness RELATED_TO workers;
training RELATED_TO workers;
training RELATED_TO use RELATED_TO PtW;
review RELATED_TO PtW.

In this report, the revision of PtW foresees to add explicitly a section on verifying the restoration of standard operation after maintenance works. Checking the plant restoration to standard operation after maintenance or change activities should be part of the PtW procedure, but the above relation (review of PtW) highlights that in some cases is still an open issue.

In many reports, however, the PtW compilation is correct, but its execution failed. As described at the beginning of this section, the PtW contains the list of collective and personal protective equipment that must be adopted; one report tells that during a check, a supervisor found the lack of fire extinguishers and explosivity detector foreseen in the PtW, as listed in Table 3 as technical barriers. This case also highlights the importance of presence, in the working area, of the supervisor, whose role is to accompany external workers and oversee all their activities from the beginning; when this checks fail, problems can occur.

Due to other concurrent works, a supervisor postponed the issuance of the PtW, but did not control the working area and third-party worker started the welding activity without waiting the PtW, and, therefore, without making the planned cleaning operations. The result was a fire start, promptly extinguished by other workers. The model contains the following relations:

third-party worker CAUSES not waiting RELATED_TO permit to work; training RELATED_TO worker.

5.1.1 Discussion on Case Study #1

The Italian Seveso legislation foresees that the Permit to Work contains all necessary information related to maintenance activities, including authorizations and responsibilities, preventive checks of conditions and materials adopted, workers' qualifications, instructions for safe working, list of safety equipment, scheduling, communication, verification of correct execution and restarting.

The inspectors already check if those points are addressed in the maintenance procedures provided by operators, as well as in those for information and training of third-party companies, and for procurement of goods and services.

The results of this study, therefore, show that the attention of inspectors towards the management of work permits and their contents, including the role of supervisor and scheduling of activities, is still strongly motivated by the near-miss events that continue to occur.

5.2 Case Study #2: Environmental risks for leakages of hazardous substances

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

The second case study is to verify if there are still situations of dispersion of hazardous substances in the ground, despite the safety measures certainly adopted and controlled in recent years. Below are the search steps that address this issue, as depicted in the Figure 3.

Table 4. List of terms

Class	Subclass	Terms
EVENT	*DETERIORATION*	*corrosion, cracking*
	FAILURE	*failure, damage, break*
	LOSS	*loss, leakage, release*
	MAJOR	***contamination***
	SUCCESS	*stop, containing, protecting from harm*
APPARATUS	*EQUIPMENT*	***tank***
	COMPONENT	*bottom*
SUBSTANCE	*NONE*	*diesel, hydrocarbon, diathermic fuel oil, gasoline, product, solvent, water*
BARRIER	*TECHNOLOGICAL*	***containment basin**, low level gauge, sump, temperature gauge*
	ORGANIZATIONAL	*emergency procedure, **remediation**, maintenance, controls, inspections*
PEOPLE	*NONE*	*supervisor*

At the first step, the question to the NLQ system is: *Which documents deal with losses in the ground?*

What we are looking for is on which situations the leaks of harmful substances required land remediation. Indeed, looking at the list of events classified as *MAJOR*, the term *contamination* suggests that some events deal with polluting conditions, as well as the item *remediation* contained in the list of organizational barriers.

Filtering the search with the above terms (i.e. *contamination* and *remediation*) reduces the number of documents, and, as expected, the types of equipment mainly involved in those events are tanks and pipelines. The third column of Table 4 lists the entities extracted from the occurrence related to tanks.

The *containment basin* is a technological barrier for gathering hazardous substances accidentally released by tanks and avoiding the ground contamination. Thus, the goal is to understand if such a barrier worked or failed. The results of the search describe loss of hazardous substances from tanks into *containment basin*. The model representing such reports contains the following relations:

(EVENT-LOSS) release INVOLVES (SUBSTANCE) product

(EVENT-LOSS) release RELATED_TO (BARRIER-TECHNOLOGICAL) containment basin INVOLVES (SUBSTANCE) product

Where *product* is a general term that indicates the hazardous substances involved, including *hydrocarbon, diathermic fuel oil, gasoline*.

In many cases, this barrier have worked and therefore only cleaning operations of the basins were required, but two reports describe the cases where this type of barrier failed and trace of hazardous substances residues have been found in the soil under the basins. One event involved a tank no longer used, while in another case, accidental release of gasoline into the basin occurred during preliminary reclamation activities for maintenance operation of a tank, the loss required the removal of part of the ground. The latter document does not specify the reason why the basin was not able to contain the loss; it could be due to an inadequacy of itself basin or to its cracking. This example points out some limits of these reports, which do not always describe in detail the reasons why an event occurred.

The number of reports relating to leaks from piping is greater than the events occurred for tanks. The Table 5 lists some of the entities extracted from those documents.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

The losses are mainly due to deterioration mechanisms that in few cases have caused serious soil contamination. The list of organizational barriers contains several types of procedures, including maintenance, checks, and controls, sometimes referred to specific Not Destructive Tests NDT.

Table 5. List of terms

Class	**Subclass**	**Terms**
EVENT	*DETERIORATION*	*corrosion, thinning, aging, mechanical stress*
	FAILURE	*failure, damage, break, defect, collision*
	LOSS	*loss, leakage, release,* ***infiltration***
	MAJOR	***contamination****, fire*
	SUCCESS	*stop, interception, protecting from harm*
APPARATUS	*EQUIPMENT*	***pipeline, oil pipeline, piping***
	COMPONENT	*flange, valve*
SUBSTANCE	*NONE*	***diesel****, hydrocarbon,* ***diathermic fuel oil****, gasoline, product, solvent*
BARRIER	*TECHNOLOGICAL*	***coating****, detector*
	ORGANIZATIONAL	*emergency procedure,* ***remediation****, maintenance, controls, inspections*
PEOPLE	*NONE*	*supervisor, worker*

Among the results, one report describes a release of diesel fuel from an abandoned pipeline, which has caused a contamination of the underlying soil. The document outlines the lack of controls on those types of equipment, and the follow-up actions relate to remediation and subsequent removal of pipeline not used.

Another case describes a leak from a pipe for which its replacement had already been planned. The loss superficially affected a portion of underlying land that was covered with a waterproof sheet in order to avoid the washout of contaminants due to rain, before transferring the polluted soil (waste) to an appropriate disposal facility.

5.2.1 Discussion on Case Study #2

The loss of containment of hazardous substances and the possible dispersion into the environment is one of the cornerstones of the Seveso directive, on these hazards the operator develops the quantitative and qualitative risk analysis, whose results address the operator to implement the measures necessary to prevent them and those to mitigate the consequences. The search activities on near miss repository, however, highlight that there are still several reports related to this topic.

The main problem is the deterioration of the equipment that can increase with its aging. The reports often describe the lack of controls and verifications and, sometimes, the ineffectiveness of some specific tests.

Another interesting point refers to equipment not currently used, often it is forgotten that decommissioned equipment might arise problems if it is not completely reclaimed and still contains residues.

During the Seveso inspections, at maintenance verification, the inspectors usually check the procedures adopted for managing equipment that are out of service, decommissioning or in demolition, including remediation and disposal of residues. Thus, the outcomes extracted from the operative experience repository and discussed above confirm the need to assess this topic in-depth.

5.3 Case Study #3: Unexpected risks - bad weather conditions

This case study deals with the issue that the occurrence of external factors with unpredictable consequences can put the safety management system in crisis. That is the case of strong and exceptional meteorological phenomena.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

Looking at the entities extracted in *EsOpIA*, some terms, classified as *EVENT – NONE*, relate to weather conditions, including *thunderstorm*, *heavy rain*, *strong wind*, *lightning*, *and ice*. The system was able to classify those terms as *event* and link them to other event items through a *CAUSES* relation.

Table 6. The weather conditions (*EVENT-NONE*) *CAUSES* some types (Subclass) of *EVENT* type

***EVENT - NONE* term**	**Subclass**	***EVENT* terms**
thunderstorm	*DETERIORATION*	*obstruction*
	FAILURE	*malfunction, damage, breakage,* ***electrical blackout***
	LOSS	*spill, release, overflow*
	MAJOR	*fire*
heavy rains	*LOSS*	*overflow*
strong wind	*FAILURE*	*failure, detachment, breakage, fall, detach, damage, vibration*
	LOSS	*leak, spill*
lightning	*FAILURE*	*damage*
	MAJOR	*fire*
ice	*FAILURE*	*breakage, failure*

The Table 6 shows the list of terms related to bad weather conditions, classified as *EVENT*-NONE at the first column, each of them has caused one or more events, contained in the third column, belonging to a certain *subclass* of *EVENT* (second column).

Thus, each row of the table is readable as a triple in the following mode:

(*EVENT – NONE) term - CAUSES - (EVENT – Subclass) terms*.

Starting from the subset of documents that refer to meteorological phenomena, *EsOpIA* application provides the functionalities to look for terms of other entity types.

The *electrical blackout* is one of the events caused by the storm; the interruption of power is usually included in the risk analysis as a possible situation that could occur. However, when this situation is caused by atmospheric events, external to the process and to the establishment, it is interesting to deepen if other elements not foreseen in the risk analysis are involved.

The Table 7 shows the list of classified terms that are inside the reports describing the blackout caused by meteorological effects. There are not terms related to events in sub-classes *MAJOR* and *DETERIORATION*, while there have been losses of containment and device failure. Some events, classified as *SUCCESS*, describe how the development of the event was interrupted by the activation of foreseen safety procedures.

Table 7. Terms contained in the model representing the outcome of the search: meteorological and blackout events.

Class	**Subclass**	**Terms**
EVENT	*LOSS*	*leakage, release, overflow*
	FAILURE	*opening, failure, overload, lack*
	SUCCESS	*stop, flow, reactivation*
APPARATUS	*EQUIPMENT*	***electric generator***, ***co-generator***, *pump, tank, reactor*
	COMPONENT	*valve, switch*
SUBSTANCE	*NONE*	***meteoric water***, *chlorine, diathermic fuel oil, product*
BARRIER	*TECHNOLOGICAL*	*basin of containment,* ***rupture disk***, *alarm, pumping system, drain well*
	ORGANIZATIONAL	*emergency procedure, stop for emergency, maintenance*

 doi:10.20944/preprints202106.0042.v1

PEOPLE	*NONE*	*worker, supervisor*

The scrutiny of the documents can continue by selecting some specific terms. Selecting the equipment *electric generator*, the search result describes two near misses dealing with the opening of *rupture disk*, both occurred in chemical sites.

In one case, the available electric generator was activated manually, but, in the meantime, a reactor has gone into overpressure with consequent opening its *rupture disk* and release of the product. In the other case, the co-generator was out of order due to a previous fault, the supervisor, in accordance with the emergency instruction, tried to restore the power supply, but during this short period, a reactor went into high-pressure causing the opening of the rupture disk. The previous cases, however, represent success stories, since the technical barriers, i.e. *rupture disk*, worked correctly.

The *EsOpIA* model, extracted by the reports, contain the following relationships between entities:

black out RELATED_TO manual activation RELATED_TO electric generator;
opening RELATED_TO rupture disc PART_OF reactor

Another report describes the impact that blackout had on the process activities. During the emergency procedure for stopping the process, an extraordinary supply of liquid was used to neutralize the high concentration hazardous substance that caused the tank overflow. Probably the procedure would have worked in case of a normal power outage, but it was not considered that the pump to dispose of the water without electricity did not work, the additional element is the over quantity of water from the storm was not foreseen.

In the Table 7, at the row *SUBSTANCE*, the list contains two hazardous substances (i.e. *chlorine* and *diathermic fuel oil*), a general term *product* that is meaningless, and the term *meteoric water*. Selecting this last term, the result gives a report that has a blackout condition similar to those described above, but the inability to restore the power was caused by the simultaneous activation of three pumps to empty the basin from the meteoric waters.

Another case is due to the excessive inflow of rainwater due to heavy rains, an overflow from the tank dedicated to their collection occurred. The interesting aspect of this document is that one of the follow-up actions has been the installation of a radar device for monitoring the level of tank gathering the meteoric water. This would suggest that equipment dedicated to services might become critical items as well those involving hazardous substances.

Operative experiences related to adverse meteorological conditions, however, represent deviations from the normal operation of the establishment. Thus, for controlling the containment of rainwater in a tank, in case of lack of technical devices (e.g. level gauge), operative instructions and procedures (i.e. organizational barriers) should include appropriate modes to operate in those specific anomalies and emergency conditions.

5.3.1 Discussion on Case Study #3

The Seveso III Directive describes the minimum information should be contained into a Safety Report SR, including the identification and accidental risk analysis and the causes of accident scenarios also due to the natural causes, for example earthquakes or floods. Therefore, in their SR, owners of establishment have to collect historical information relating to the meteorological, geophysical and hydrogeological events occurred in their site. Interviewing some inspectors, however, it emerges that, during Seveso SMS inspections, usually no one asks questions relating to bad weather problems, that is, what measures have been taken for this issue, unless there are explicit operational experiences in this regard.

The feedback from operational experiences described in 5.3 section would be useful for the authorities who periodically audit the Safety Reports, to understand if the measures taken by the owners address issues related to rapid and worse change of climatic conditions. Is therefore reasonable that auditors of SR assess whether it may occur

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 June 2021 doi:10.20944/preprints202106.0042.v1

extreme meteorological events that go beyond time series. For example, knowing whether the amount of rain fallen in a short time is equal to that recorded over a long period, can help the inspector to assess whether the necessary prevention measures have been taken to cope with extreme conditions. Thus, knowing the events already occurred in similar conditions, the critical aspects, but also the solutions adopted, could be useful to auditors.

6. Discussion

The study presented is the outcome of the collaboration of experts in ML application and text mining techniques, together with the Seveso inspectors whose skills have allowed the definition of the most appropriate knowledge domain, in such a complex sector as the process industry.

The model built and the ways of inquiring events and relationships allow to highlight hidden knowledge that otherwise would have been lost as described in the case study on weather events. While considering the traditional risks, the case study of the loss of hazardous substances into the ground, or the organizational procedures, i.e. work permits, further lessons can still be drawn in addition to confirm those already known. The three case studies reported have shown that the goal has been achieved.

Representing the content of operational experiences and analysing it has the aim of sharing knowledge among all inspectors. Over the years, each inspector, indeed, gains experiences and knowledge that often are more in-depth in some sectors that in others. Thus, to analyse different situations, the inspector should read a lot of documents asking other colleagues for them. Hence, the usefulness of a shared repository and a tool for extracting knowledge.

A further strength is to allow inspections to be conducted in a homogeneous manner throughout the national territory through the lessons learned from which inspectors can take inspiration to make their activity more effective. These suggestions to reflection and study are as *warnings,* which may arise other queries from the inspectors to the operator, following a slogan such as "*Make you think of!*".

The discussions made in the previous section, for each case study, can be considered answers to general questions, some of which are listed in Table 8.

Table 8. Some suggestions from the discussions of case studies.

Case Study	*Make you think of!*
#1 Permit to work	*How is the supervisor role managed?* *How did the activity schedule?*
#2 Loss into ground	*Look at the decommissioned equipment!* *Are they in the maintenance scheduling?*
#3 Weather events	*Did extreme weather events occur in the region?* *Are there barriers implemented for heavy rains?*

7. Concluding remarks

The importance of near misses analysis has been known for decades. This research has shown how NLP and ML capabilities enforce the power of near miss management extracting hidden knowledge and highlighting both wrong conditions or situations and success stories.

The model implemented is strongly based on events and relationships with other entities that have been identified to express the concepts contained in the operational experiences; thus, it is able to represent the story contained in the text. The model is also suitable for searching purposes: through *EsOpIA* functionalities, the inspector can browse on the entity-relation graph.

The robustness of the model is tested by evaluating its applicability on other types of documents of the same domain of interest, including accident reports and equipment failure data sheets. From the first results of this test, it emerged that the conceptual model is

valid, while it may be necessary to update the term sets on specific fields chosen with further training, i.e. accidents and equipment failures.

Even if the application works on documents in Italian, the conceptual model is usable with text archives in other languages after appropriate training.

8. Patents

Author Contributions: Conceptualization, S.M.A. and P.A.; methodology S.M.A.; data curation A.P. and M.R.V.; writing—original draft preparation, S.M.A. and P.A.; writing-review and editing, all authors.

Funding: This research received no external funding.

Data Availability Statement: The data are currently available only for Seveso inspectors internal to Inail.

Acknowledgments: The authors gratefully acknowledge: Flavia Fattori, Manuel Raimondi and Luca Di Piramo for their precious involvement in the project development, Paolo Bragatto for his constant and precious support, all Inail Seveso inspectors for collecting the near miss reports, without their collaboration, indeed, this research would be poor of data.

Conflicts of Interest: The authors declare no conflict of interest.

Acronyms: SMS – Safety Management System; TM – Text Mining; AI – Artificial Intelligence; ML – Machine Learning; NLP – Natural Language Processing; PtW – Permit to Work.

References

1. Gnoni, M.G.; Saleh, J.H. How near miss management systems and system safety principles could contribute to support high reliability organizations. *Safety and Reliability - Theory and Applications* - Proceedings of the 27th European Safety and Reliability Conference, ESREL 2017, pp. 3099-3104.
2. Taylor, J.A., Lacovara, A.V., Smith, G.S., Pandian, R., Letho, M. Near-miss narratives from the fire service: A Bayesian analysis. *Accident Analysis and Prevention*, 2014, 62, pp. 119-129.
3. https://ec.europa.eu/environment/seveso/legislation.htm
4. Agnello; P.; Ansaldi; S.M.; Bragatto; P.A. Plugging the gap between safety documents and workers perception; to prevent accidents at seveso establishments. *Chemical Engineering Transactions*, 2012, 26, pp. 291-296, doi: 10.3303/CET1226049.
5. UNI 10617:2019. Establishments with major-accident hazards – Safety Management Systems – Essential requirements. UNI Ente Italiano di Normazione. 2019.
6. Saleh, J.H., Saltmarsh, E.A., Favarò, F.M., Brevault, L. Accident precursors, near misses, and warning signs: Critical review and formal definitions within the framework of Discrete Event Systems. *Reliability Engineering and System Safety*, 2013, 114, pp. 148-154, doi: 10.1016/j.ress.2013.01.006.
7. Phimister, J.R., Oktem, U., Kleindorfer, P.R., Kunreuther, H. Near-miss incident management in the chemical process industry. *Risk Analysis*, 2003., 23, 3, pp. 445-459.
8. https://emars.jrc.ec.europa.eu/en/emars/content
9. https://www.aria.developpement-durable.gouv.fr/
10. https://www.infosis.uba.de/index.php/en/site/13947/zema/index.html
11. https://www.csb.gov/
12. https://varo.tukes.fi/
13. http://www.sozogaku.com/fkd/en/
14. Bragatto; P.A.; Agnello; P.; Ansaldi; S.; Pittiglio; P. Exploiting near misses to revive safety knowledge in process plants. *AIChE Annual Meeting; Conference Proceedings*, 2010.
15. Zhou Z., Li C., Mi C., Qian L. Exploring the potential use of near-miss information to improve construction safety performance. *Sustainability*, 2019, 11, 1264; doi: 10.3390/su11051264.
16. Single, J.I., Schmidt, J., Denecke, J. Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. *Safety Science*, 2020, 129, doi: 10.1016/j.ssci.2020.104747.
17. Hughes, P., Robinson, R., Figueres-Esteban, M., van Gulijk, C. Extracting safety information from multi-lingual accident reports using an ontology-based approach. *Safety Science*, 2019, 118, pp. 288-297.
18. Nakata, T. Text-mining on incident reports to find knowledge on industrial safety. *Proceedings - Annual Reliability and Maintainability Symposium*, 2017, 7889795, doi: 10.1109/RAM.2017.7889795.
19. Bragatto, P.A., Ansaldi, S., Antonini, F., Agnello, P. Bow-tie approach for improved auditing procedures at "Seveso" establishments. *Safety, Reliability and Risk Analysis: Beyond the Horizon - Proceedings of the European Safety and Reliability Conference, ESREL 2013*, 2014, pp. 1447-1455.
20. Bragatto, P., Agnello, P., Ansaldi, S., Artenio, E., Delle Site, C. Reviving knowledge on equipment failures and improving risk management at industrial sites. *Journal of Applied Engineering Science*, 2015, 13, 4, pp. 271-276, doi: 10.5937/jaes13-9573.

21. Ansaldi, S.M., Agnello, P., Bragatto, P.A. Incidents triggered by failures of level sensors. *Chemical Engineering Transactions*, 2016, 53, pp. 223-228. doi: 10.3303/CET1653038.
22. Cheng M.-Y., Kusoemo D., Gosno R.A. Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, 2020, 118, doi: 10.1016/j.autcon.2020.103265.
23. Arteaga C., Paz A., Park J. Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach. *Safety Science*, 2020, 132, doi: 10.1016/j.ssci.2020.104988.
24. Kurian D., Sattari F., Lefsrud L., Ma Y. Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. *Safety Science*, 2020, 130, doi: 10.1016/j.ssci.2020.104873.
25. Paltrinieri N., Comfort L., Reniers G. Learning about risk: Machine learning for risk assessment. *Safety Science*, 2019, 118, pp. 475-486, doi: 10.1016/j.ssci.2019.06.001.
26. Xu Z., Saleh J.H. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliability Engineering and System Safety*, 2021, 211, doi: 10.1016/j.ress.2021.107530.
27. https://www.ibm.com/it-it/cloud/watson-knowledge-studio
28. Ansaldi, S.M., Simeoni, C. Di Francesco, A., Martini, R., Di Piramo, L., Fattori, F. Extracting knowledge from near miss reports using machine-Learning techniques. *Proceeding of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and management Conference*, 2020.
29. Ansaldi S.M., Pirone A., Vallerotonda M.R., Bragatto P.A., Agnello P., Delle Site C. How inspections outcomes may improve the foresight of operators and regulators in Seveso industries. *Chemical Engineering Transactions*, 2019, 67, doi: 10.3303/CET1867062.