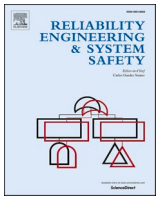
Reliability Engineering and System Safety 257 (2025) 110834

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



A systematic procedure for the analysis of maintenance reports based on a taxonomy and BERT attention mechanism

Dario Valcamonico [a], Piero Baraldi [a,*], July Bias Macêdo [b], Márcio Das Chagas Moura [b], Jonathan Brown [c], Stéphane Gauthier [c], Enrico Zio [d,a]

[a] Energy Department, Politecnico di Milano, Milan, Italy
[b] Department of Industrial Engineering, Center for Risk Analysis and Environmental Modeling, Universidade Federal de Pernambuco, Brazil
[c] Alstom Group, France
[d] MINES Paris-PSL, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France

ARTICLE INFO

Keywords:
Maintenance
Natural Language Processing
BERT
DBSCAN
Freight transport trains

ABSTRACT

This work proposes a systematic procedure for analyzing maintenance reports to support maintenance decision-making for a fleet of similar systems. The proposed procedure allows achieving three objectives: (1) grouping maintenance interventions, (2) identifying common characteristics in the maintenance interventions, and (3) recognizing occurrences of rare events of maintenance intervention. Specifically, the attention mechanism of Bidirectional Encoder Representation from Transformer (BERT) and the Density Based Spatial Clustering Applications with Noise (DBSCAN) methods are combined to group maintenance interventions according to their similarity of stated features. A taxonomy of the words used in the textual reports to state the maintenance interventions is developed to systematically identify common features of the clusters, such as the involved components, their working state, the occurred failures or malfunctions, the performed maintenance actions and the personnel that has performed the intervention. The proposed procedure is applied to a repository of reports of maintenance interventions performed on mechanical and electric components of traction systems of a fleet of trains. The obtained results show that it can effectively support decision-making on the maintenance of traction systems.

1. Introduction

Accidents and failures threaten safety, availability and productivity of industrial systems throughout their lifecycle. Adequate maintenance planning counteracts aging and degradation, minimizing or preventing the occurrence of adverse events and their consequences.

Learning from past accidents, failures, and maintenance activities can lead to more effective maintenance [1]. Maintenance interventions are typically reported by operators in documents containing free-text descriptions of the occurred failures and malfunctions, the performed inspection and replacement activities, and multiple-choice structured fields with other information, such as the involved components, the event severity and the date of the intervention. The use of these sources of information for safety assessments, reliability analyses, and maintenance scheduling is typically challenging [2]. The main reasons are:

(1) The complexity of the language of the documents, including technical difficulties such as the use of domain-related acronyms, abbreviations and codes [3];
(2) The heterogeneity of the reports, which are typically characterized by a multitude of formats, a lack of standardization and the subjectivity of operators [4];
(3) The large number of reports to be considered creates a significant workload for experts, increasing the risk of errors in analysis and possibly leading to misguided maintenance decisions [5].

Natural Language Processing (NLP) offers a promising solution [6]. NLP allows transforming the text of the reports into a set of numerical features, which can be used as input of machine learning methods for extracting knowledge, e.g., in the form of clusters. Section 1.1 reviews the works that have applied NLP methods to repositories of maintenance reports for supporting maintenance decisions. The most promising

* Corresponding author.
E-mail addresses: dario.valcamonico@polimi.it (D. Valcamonico), piero.baraldi@polimi.it (P. Baraldi), july.bias@ufpe.br (J.B. Macêdo), marcio.cmoura@ufpe.br (M.D.C. Moura), jonathan.brown@alstomgroup.com (J. Brown), stephane.gauthier@alstomgroup.com (S. Gauthier), enrico.zio@polimi.it (E. Zio).

https://doi.org/10.1016/j.ress.2025.110834
Received 2 April 2024; Received in revised form 10 January 2025; Accepted 14 January 2025
Available online 15 January 2025

works combine NLP methods with expert knowledge about the system, its failure modes, degradation mechanisms and maintenance activities represented using taxonomies, ontologies and Knowledge Graphs (KGs). However, the combined use of state-of-the-art NLP methods, such as Pretrained Language Models (PLMs) and Large Language Models (LLMs), with methods for the representation of expert knowledge, has not yet been investigated [7]. In NLP, PLMs and LLMs are emerging for their capability of understanding (and reproducing) the content and form of textual data by considering the semantic, positional and logic relations among the words in the text, using a large number of parameters and training the models with extensive datasets. LLMs can generate semantically coherent answers to arbitrarily complex user queries by using a number of parameters one or two orders of magnitude larger than that of PLMs methods [7]. Notice, however, that the most successful applications of PLMs and LLMs have been obtained in applications such as social networks [8] and biology [9], where the number of textual documents available for model training is remarkably larger than the number of maintenance reports typically available in industrial applications.

In this context, this work develops a systematic procedure for analyzing maintenance reports to support maintenance decision-making for a fleet of similar systems. The proposed procedure allows achieving three objectives: (1) grouping maintenance interventions, (2) identifying common characteristics in the maintenance interventions, and (3) recognizing occurrences of rare events of maintenance intervention.

Given the lack of works combining the promising NLP methods of PLMs and LLMs with methods for representing expert knowledge for the analysis of repositories of maintenance interventions, the procedure developed in this work is based on:

(1) Representing the textual reports of the maintenance interventions in the form of numerical vectors by using the PLM method Bidirectional Encoder Representation from Transformers (BERT), which has proven successful in many NLP applications [10,11];
(2) Clustering the vectors by using the Density Based Spatial Clustering Applications with Noise (DBSCAN) method [12];
(3) Identifying the characteristics of the obtained clusters by using a taxonomy defined by field experts, specifically a top-down structure of the words in sub-levels representing specific elements of domain knowledge [13].

Considering (1), BERT was preferred to other NLP methods, such as Term Frequency Inverse Document Frequency (TFIDF), topic modelling and Word2Vec, because it generates vectorial representations of the words that allow accurately catching their semantic meaning [14]. Specifically, we consider the attention mechanism of BERT to transform each report into a numerical vector where the elements are the attention given by BERT to the words when performing the classification of the event severity [15]. BERT is selected also because it is a PLM pretrained on a large corpus of textual data, which allows obtaining informative representations of the reports also in the case of this work, wherein the number of available maintenance reports is limited to few hundreds.

Concerning (2), once the numerical representations of the reports have been obtained in the form of attention vectors, the categorical information of the component involved in the maintenance interventions, which is provided by the maintenance operators using a label, is appended to the attention vectors. Then, the DBSCAN method is used to cluster the obtained vectors [12]. DBSCAN is selected because it has outperformed other clustering methods in applications involving clustering long vectors containing numerical and categorical data [16]. Also, since DBSCAN is a density-based clustering method, it is not forced to assign all vectors to a cluster. Consequently, if a vector is remarkably dissimilar to the others, the method does not assign it to any cluster. This allows identifying rare events of maintenance intervention that do not belong to any cluster.

Concerning the identification of the common characteristics of the reports assigned to a cluster (3, above), the increasing ability of recent NLP methods to accurately represent the semantic content of textual data comes at the expenses of making their outcomes less interpretable [17]. To address this issue, eXplainable Artificial Intelligence (XAI) methods have been developed with the objective of explaining the output of machine learning models [17]. According to [18], the attention mechanism of transformer-based methods, such as BERT, can be exploited directly for explainability. Specifically, the attention mechanism enables feature importance-based explanations, i.e. it allows assigning a measure of the importance of a word in a text [18]. In this work, the words with largest associated attention are those that are more relevant for the BERT fine-tuning task, which corresponds to the classification of the event severity. Considering the objective of identifying the common characteristics of the clusters found by DBSCAN, this work analyzes the statistical distribution of the attentions of the words of the reports. A challenge is that the reports are written using many synonyms and technical terms to describe the same concepts. Expert knowledge is used to address this challenge by developing an ad-hoc taxonomy that organizes the words in sub-levels representing the main elements of the reported maintenance process (e.g. the involved components, their working state, the occurred failures or malfunctions, the performed maintenance actions and the personnel that has performed the intervention). As a result, the use of sub-levels, to which semantically similar words are assigned, allows identifying similar concepts expressed using different words in reports of the same cluster. In practice, the characteristics of the clusters are identified by considering the words of the reports with largest total attention associated in each sub-level of the taxonomy and, as such, identifying the common characteristics of the clusters. Notice that the taxonomy developed in this work differs from the taxonomies traditionally used in industrial practice. For example, the taxonomy developed in [19] classifies standard equipment of oil and gas plants location, use and type. Since maintenance reports typically contain also information about the causes, the consequences and the factors that influenced the occurrence and severity of the event, the available industrial taxonomies cannot be directly used for the purpose of this work. Also, industrial taxonomies are typically proposed for specific sectors, e.g., the one proposed in [20] for the analysis of accidents and risks in floating oil and gas systems. As a consequence, they cannot be applied to other industrial sectors characterized by different maintenance issues.

The proposed procedure is applied to a repository of real maintenance intervention reports on mechanical and electric components of traction systems of a fleet of trains. The reports contain the narrative of the maintenance events, the involved components and the descriptions of the actions performed. The obtained results have been validated by the experts of the company owning the data.

The main contribution of this work is the development of a systematic procedure for analyzing maintenance reports to:

i. Find homogeneous clusters of maintenance interventions;
ii. Identify common physical/engineering characteristics of the clusters, such as the types of components, their working state, the specific failures or malfunctions occurred, the maintenance actions performed and even the personnel that performed the intervention;
iii. Recognize rare event occurrences among the reported maintenance interventions. This is done by a novel use of the DBSCAN method, which is not forced to assign a maintenance report to a cluster if it is dissimilar to the other reports of the same cluster.

The remainder of the work is organized as follows. In Section 1.1, a literature review on NLP applications in maintenance decision-making is presented. In Section 2, the developed procedure is presented. In Sections 3–5 the methods for text representation, clustering the data and analysis of the cluster are described, respectively. In Section 6, the case study is introduced. In Sections 7 the procedure is applied to the case

study and in Section 8 the results are presented. Finally, in Section 9, conclusions and future works are discussed.

1.1. NLP for maintenance

This Section considers a selection of works which apply various NLP methods to unstructured maintenance repositories. They can be classified as supervised and unsupervised. Supervised methods aim at classifying the causes of accidents and abnormal conditions [21–24], the type of accident events [25,26] and the severity of their consequences [27]. They are developed using a database of labeled data. On the contrary, unsupervised methods aim at identifying the factors influencing the occurrence and severity of accidents [28–33] and discovering functional dependencies in the systems [34].

The application of NLP to industrial maintenance and Prognostics and Health Management (PHM) has been systematically presented in [6, 35], respectively. The two main open issues that emerge are:

1. The limited number of maintenance reports typically available in industrial applications;
2. The need to provide explainable outcomes to effectively support maintenance decision-making.

According to both [6,35], integrating expert knowledge can address these two open issues. This Section discusses how expert knowledge has been used to support the application of NLP methods to repositories of maintenance reports. Table 1 classifies the works considering the NLP methods used to process the textual data and the methods used to represent the expert knowledge.

NLP methods are applied to maintenance reports without using expert knowledge in [21–34,36–41]. Specifically, in [21] a NLP method based on text processing and word counts is used to find common failure causes in Swedish railways. In [22], BERT is combined with Relational Graph Convolutional Networks (R-GCNs) to classify the causes of accidents from accident investigation reports. In [23], Latent Semantic Analysis and Convolutional Neural Networks (CNNs) are combined for the classification of textual maintenance records, with the final objective of developing a stochastic multi-stage model of the degradation of excavator components used in the mining industry. In [24], the FastText method, which embeds the textual data into numerical vectors, is combined with Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) neural networks for the classification of maintenance work orders for buildings. In [25], Word2vec is combined with deep neural networks to develop a classifier method for supporting rapid prediction of aviation safety risk. In [26], BERT is combined with CNNs to identify and estimate the frequency of contributing factors of accident in confined spaces. In [27], CamemBERT, a version of BERT fine-tuned on French language, is used to cluster maintenance logs for identifying the criticality and duration of maintenance issues. The obtained clusters are compared to those obtained with TFIDF. The obtained results show that the BERT-based method significantly outperforms the TFIDF method. In [28], Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are applied to railroad accident reports to identify common failure trends in train equipment. In [29], LDA is used to discover latent topics in industrial maintenance reports, which and then used for improving maintenance service and asset design. In [30], LDA is applied to identify the risk factors and their coupling from aviation safety reports. In [31], two different variants of BERT are compared in terms of interpretability of the results with the objective of extracting factors influencing accidents from reports related to manufacturing production plants. In [32], a NLP method based on text chains is developed to extract fault features from accident reports of high-speed trains, with the objective of maintenance improvement. In [33], LDA is used to extract information from reports of severe accidents of trains and quantitatively estimate their severity. In [34], Doc2Vec is combined with two clustering methods (Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) and Fuzzy C-Mean (FCM)) to automatically discover functional dependencies in on-board train control systems. In [36], the Bag-of-Words (BoW) method of the text is combined with three different machine learning classifiers (Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM)) with the objective of automatizing the assignments of the personnel for the maintenance of buildings. A similar analysis is performed in [37], where a procedure based on the combination of NLP with Naïve-Bayes and SVM classifiers is used to identify failure times in maintenance work orders. In [38], TFIDF and LDA are combined with k-nearest neighbor clustering to support maintenance in nuclear power plants by retrieving maintenance cases similar to the candidate one from an historic database. In [39], the factors influencing the occurrence and severity of accidents in coal mines are identified from textual reports by extracting keywords by TFIDF and finding association rules between them. In [40], several pre-trained language models and large language models, including GPT3.5 and Llama2, are compared with respect to the task of classifying the severity of records of hazardous events that have occurred and affected high-speed train braking systems. In [41], BERT, TFIDF and DBSCAN are combined to cluster the reports of aviation accidents and identify the human factors influencing the occurrence of the accidents.

The main limitation of the above works [21–34,36–41] is their reliance on ad-hoc procedures for using expert knowledge in data pre-processing and setting of the methods parameters, with no structured representation of this knowledge.

NLP methods have been already combined with the following representations of expert knowledge:

i. Ontologies, i.e. formal descriptions of the domain knowledge of systems and their processes [58]. In [42], TFIDF is combined with a taxonomy of the factors influencing accidents, to develop a Bayesian network for tracing the evolution of the probabilities of accidents of different severity classes and identifying the factors

Table 1
Literature review.

		NLP method					
		Traditional methods (e.g., Bag of Words)	Topic Modelling	Words embedding techniques (e.g., Word2Vec, bi-LSTM)	Pretrained Language Models (PLMs) (e.g., BERT)	Large Language Models (LLMs) (e.g., ChatGPT)	Other (e.g., descriptive statistics, in-house NLP pipeline)
Representation of expert knowledge	**Not used**	[21,36–39]	[28–30,33,38]	[25,34]	[22,26,27,31,40,41]	[40]	[23,24,32]
	Taxonomy				[42]		
	Ontology		[43]				[44,45]
	Knowledge Graph	[46]		[47,48]	[49–51]		[52]
	Named Entity Recognition	[53]		[54,55]	[50,56]		[3,57]

that influence most their changes. In [43], topic modelling is used to extract information from reports of unsafe occurrences in the aviation industry. The obtained topics are validated considering ontologies about aviation safety and reliability, developed by regulatory bodies. In [44], expert knowledge and accident reports are combined in an ontological approach for the analysis of accidents in chemical plants. In [45], expert knowledge on aero-engine faults is structured and combined with their textual descriptions for developing a case-based system to automatically analyze the faults and supporting maintenance of aero-engines;

ii. Knowledge Graphs (KGs), i.e. explicit conceptualization of the knowledge using a set of entities connected by logic relations. Differently from ontologies, KGs are designed for a specific system and/or process [59]. In [46], an approach based on KGs is developed for modelling railway operational accidents with the objective of supporting railway operators in accident prevention. In [47], bidirectional Long Short Term Memory (bi-LSTM) is combined with a RF classifier to develop a KG to support quantitative risk assessment of railway accidents. In [48], a bi-LSTM model is combined with an expert-based ontology to build a KG for providing maintenance suggestions of aircrafts. In [52], a model of accidents of oil and gas pipelines is built extracting a set of features from accident reports, using an annotation technique. Then, a feature graph is developed to identify the recommendations and procedure adopted in similar accidents;

iii. Named Entity Recognition (NER), which automatically assign words, or a combination of them, to classes defined considering the application and context [60]. In [53], corrosion related incidents of pipelines are analyzed by a combination of a NER method to extract features from incidents reports and a Bayesian Network (BN) to quantify the probabilities of the causes and the severity of consequences. In [54], the Word2Vec model is combined with literature maintenance vocabularies for developing a pipeline for technical language processing with the final objective of automatically annotating maintenance work orders. In [55], CNNs are considered to develop a NER method to support the HAZOP of aircraft systems. In [3], automatic annotation of textual maintenance work orders is performed. The method is built using a repository of textual documents that have been annotated by field experts to address technicalities in the language. In [57], an NLP pipeline is combined to an expert-based causal model to develop a method for annotating health reports of centrifugal pumps in oil and gas plants.

The reviewed works [3,44–48,52–55,57] show that combining systematic representations of expert knowledge with NLP methods can facilitate the identification of the factors influencing the occurrence and severity of failures and malfunctions and improve the performances. However, they use NLP techniques, such as Bag of Words and other methods of descriptive statistics, which do not rely on the capabilities of the newest PLMs and LLMs of understanding (and reproducing) the content and form of textual data by numerically embedding the semantic, positional and logic relations among the words in the text. Typically, these methods are trained on large general-purpose datasets and fine-tuned on case specific datasets. Nevertheless, due to the extensive use of technical language and the limited number of safety and maintenance reports in real repositories, fine-tuning may not be sufficient to exhaustively understand technical language. Therefore, combining of PLMs and LLMs with systematic representations of expert knowledge is expected to be effective [35].

Some examples in this direction are reported here. In [49], a KG is used in combination with BERT for automatically annotating reports of accidents involving hazardous chemicals. In [51], BERT and bi-LSTM are used to extract features from bridge inspection reports and build a KG aiming at developing a question-and-answer tool for supporting the safe operation of bridges. In [56], BERT and topic modelling are used to annotate the reports and to extract knowledge on the hazards and accident causes from marine accident reports. In [61], BERT is fine-tuned using an annotated dataset of maintenance work orders and is used to retrieve cases similar to past ones. Other applications combine PLMs and expert knowledge to support maintenance. For example, in [50], PLMs are used in combination with expert knowledge to develop a NER method to automatically annotate maintenance work orders and to build a KG to extract knowledge on the failure modes.

2. Procedure

We consider a repository of D maintenance intervention reports, $R = \{r_i, i = 1,\ldots,D\}$, collected by a company from a fleet of similar systems. Each maintenance intervention report r_i, $i = 1,\ldots,D$, contains:

i. The free-text description d_i, $i = 1, \ldots, D$, of accidents or malfunctions that occurred, and the maintenance intervention performed;
ii. The categorical variable $x_i \in \{1,\ldots,X\}$ indicating the component involved in the event;
iii. The categorical variable $s_i \in \{0,\ldots,S\}$ indicating the event severity, where 0 and S correspond to the least and most impactful consequences, respectively. It is associated to the events by the maintenance operators following company standards.

In this work, a systematic procedure is proposed for achieving three objectives: (1) automatically clustering the maintenance interventions recorded in the reports stored in the repository in groups of similar maintenance interventions, (2) identifying the common characteristics shared by the maintenance interventions belonging to the clusters and (3) recognizing occurrences of rare events of maintenance interventions, which do not belong to any identified cluster.

The procedure combines the three phases of (Fig. 1):

i Text representation, which transforms the free-text descriptions of the maintenance interventions $\{d_i, i = 1,\ldots,D\}$ into numerical vectors $\{\alpha_i, i = 1,\ldots,D\}$. This is done considering the information about the severity of the events $\{s_i, i = 1,\ldots,D\}$;
ii Clustering, which receives in input the pairs $\{(\alpha_i, x_i), i = 1,\ldots,D\}$ of numerical vectors and component labels, and provides in output the M clusters $\{c_1,\ldots,c_M\}$ of similar maintenance intervention reports;
iii Representation of the expert knowledge, using a taxonomy of the words of the reports to identify the common characteristics of the events of a cluster.

Sections 3–5 describe the methods developed in this work for text representation, clustering, and analysis of the obtained clusters, respectively.

3. Text representation

The representation of the free-text descriptions of the maintenance interventions $\{d_i, i = 1,\ldots,D\}$ in numerical vectors $\{\alpha_i, i = 1,\ldots,D\}$ is obtained using the attention matrices of the BERT model (Fig. 2). Section 3.1 describes the report pre-processing, the BERT method and its fine-tuning process, Section 3.2 introduces the attention mechanism of BERT, whereas Section 3.3 presents the final encoding of the reports.

3.1. BERT

Firstly, the reports are pre-processed to properly deal with technical terms such as abbreviations of components and spare parts, which are difficult to automatically handle by BERT, especially when the number of available reports is limited, as in the situation of this work. The specific steps performed to pre-process the maintenance reports are detailed in Section 7.

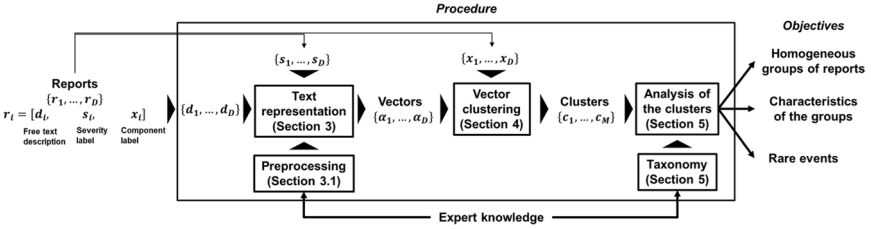


Fig. 1. Systematic procedure proposed.

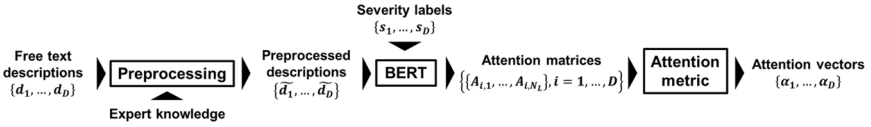


Fig. 2. Developed method of text representation.

BERT is a NLP model based on the transformer architecture [62]. Specifically, a stack of N_L transformer encoders $\{1,\ldots,l,\ldots,N_L\}$ are sequentially connected. Each transformer encoder is composed by a multi-head attention part, which encodes the contextual relationship between the words of the input text, and a feedforward neural network, wherein the numerical matrices processed by the multi-head attention layer are concatenated and propagated to the next encoder layer [63]. BERT is trained on the two tasks of predicting whether given two sentences the second one follows the first, and of predicting a masked word using the surrounding words of the same sentence. Given the large number of parameters, training BERT from scratch is unfeasible with the standard computational resources and limited textual databases. A possible solution is to rely on transfer learning, which allows the reuse of a version of BERT pretrained on very large textual databases to perform other downstream tasks such as classification on smaller domain-specific datasets [63].

Specifically, BERT parameters are fine-tuned considering the task of classifying the severity of the events. This is done by adding a fully connected linear layer at the back-end of the BERT architecture, which receives in input the numerical vectors representing the free-text descriptions of the maintenance interventions $\{d_1,\ldots,d_i,\ldots,d_D\}$ and produces in output the labels of the severity of the events $\{s_1,\ldots,s_i,\ \ldots,s_D\}$ [63].

3.2. Attention mechanism

The core property that BERT inherits from transformers is the attention mechanism [62]. The main idea behind the attention mechanism is to find a weight distribution of the input text which assigns larger values to words and sequences of words that are more relevant to the specific task which BERT is trained for.

In practice, given a report r composed of the sequence of N words $\left\{p_1,\ldots,p_j,\ldots p_N\right\}$ the attention matrices are extracted by performing the following sequential steps:

i. Each word is searched into the vocabulary that BERT learned during pretraining and, if not present, it is split into tokens already present in the vocabulary. For example, in the case of interest in this work the words "*trains*" and "*fixing*" are split into the tokens "*tr*", "*ain*" and "s" and "*fix*" and "*ing*", respectively. The result is that the report r is converted into the sequence of the T tokens $\{t_1,\ldots,t_k,\ldots,t_T\}$;
ii. Each token t_k is transformed into a numerical vector e of length N_e, which was obtained during the training phase by considering the semantic meaning of the token, the relations with the other tokens and its relative position in the sequence [62]. The report r is, therefore, transformed into a numerical matrix $E=[e_1,\ldots,e_T]$ of dimensions $T\times N_e$;
iii. The embedding matrix E is, then, passed to the BERT architecture. The multi-head attention part of a generic layer l, is composed by three different weight matrices $W_{Q,l}$, $W_{K,l}$ and $W_{V,l}$ of dimensions $N_e\times N_e$, called query, key and value matrices, respectively. The idea behind these matrices is that they attempt to numerically represent the semantic and syntactic characteristics of the input reports [62];
iv. Each one of the weight matrices $W_{Q,l}$, $W_{K,l}$ and $W_{V,l}$ is divided into $N_f=N_e/N_H$ sub-matrices $\left\{W^1_{Q,l},\ldots,W^h_{Q,l},\ldots,W^{N_H}_{Q,l}\right\}$, $\left\{W^1_{K,l},\ldots,W^h_{K,l},\ldots,W^{N_H}_{K,l}\right\}$ and $\left\{W^1_{V,l},\ldots,W^h_{V,l},\ldots,W^{N_H}_{Q,l}\right\}$ of dimensions $N_e\times N_F$, respectively. The pieces $\{1,\ldots,h,\ldots,N_H\}$ are called "heads" in the jargon of BERT;
v. The embedding matrix E is multiplied by the sub-matrices $\left\{W^1_{Q,l},\ldots,W^h_{Q,l},\ldots,W^{N_H}_{Q,l}\right\}$, $\left\{W^1_{K,l},\ldots,W^h_{K,l},\ldots,W^{N_H}_{K,l}\right\}$ and $\left\{W^1_{V,l},\ldots,W^h_{V,l},\ldots,W^{N_H}_{Q,l}\right\}$, and the sets of matrices $\left\{Q^1_l,\ldots,Q^h_l,\ldots,Q^{N_H}_l\right\}$, $\left\{K^1_l,\ldots,K^h_l,\ldots,K^{N_H}_l\right\}$ and $\left\{V^1_l,\ldots,V^h_l,\ldots,V^{N_H}_l\right\}$ of dimensions $T\times N_f$ are obtained. Notice that the estimation of these matrices can be done in parallel for each head to reduce the computational burden;
vi. The attention matrix A_l of dimension $T\times N_e$ related to the l-th layer of BERT is then computed by concatenating the attention scores $\left\{a^1_l,\ \ldots,a^h_l,\ldots,a^{N_h}_l\right\}$ [62]:

$$a^h_l = softmax\left(\frac{Q^h_l K^{h^T}_l}{\sqrt{N_f}}\right)V^h_l \tag{1}$$

where, $K^{h^T}_l$ is the transpose of the matrix K^h_l and the function "softmax" is applied to convert the ratio in a value between 0 and 1. The idea behind (1) is that attention is quantified as a weighted sum, where the weights estimate how similar the query is to the key.

Steps (*i*) to (*vi*) are repeated for each layer of BERT and the collection of attention matrices $\{A_1,\ldots,A_l,\ldots,A_{N_L}\}$ associated to the report r is obtained.

3.3. Attention vectors

Since the objective of explainability of the clusters is obtained based on the words present in the reports, the attention matrices are used to transform the free-text descriptions of the maintenance interventions d_i into a vector α_i, $i=1,\ldots,D$, such that a generic element of α_i represents a

measure of importance of the corresponding word in d_i. Specifically, inspired by the procedure in [64], the attention vectors are obtained by:

i. Computing the N_L attention matrices $\{A_{i,1},\ldots,A_{i,l},\ldots,A_{i,N_L}\}$;
ii. Averaging them and, then, averaging the resulting matrix along the N_e columns to obtain a vector of length T;
iii. Summing the sub-set of elements of the obtained vector corresponding to the tokens in which each word is split, to extract the attention of the words originally present in the report r_i;
iv. Converting the obtained vector into a sparse vector α_i of fixed length, V, equal to the number of words used in the vocabulary of the reports $\{p_1,\ldots,p_v,\ldots,p_V\}$, i.e. the list of all the unique words used in the reports; notice that an element is equal to zero if the corresponding word p_v is not present in the report.

The outcome of the procedure is the collection of the attention vectors $\{\alpha_1,\ldots,\alpha_i,\ldots,\alpha_D\}$.

4. Clustering

Once the free-text descriptions of the maintenance interventions $\{d_i, i=1,\ldots,D\}$ are converted into the attention vectors $\{\alpha_i, i=1,\ldots,D\}$, a clustering approach for mixed data is applied. The objective is to find groups of similar maintenance interventions by clustering the pairs (α_i, x_i) of attention vectors, α_i, and associated component labels, x_i, with $i=1,\ldots,D$. The clustering algorithm uses the Gower distance for mixed data [65]:

$$\delta^{GOW}_{i_1,i_2} = w^{CAT}\delta^{CAT}_{i_1,i_2} + w^{NUM}\delta^{NUM}_{i_1,i_2} \tag{2}$$

where i_1 and i_2 are two generic reports of set R, $\delta^{CAT}_{i_1,i_2}$ is the distance between the component labels x_{i_1} and x_{i_2}:

$$\delta^{CAT}_{i_1,i_2} = \begin{cases} 0 \ if \ x_{i_1} = x_{i2} \\ 1 \ if \ x_{i_1} \neq x_{i2} \end{cases}, \tag{3}$$

and $\delta^{NUM}_{i_1,i_2}$is the cosine distance between the attention vectors:

$$\delta^{NUM}_{i_1,i_2} = \frac{\sum_{k=1}^{N_V}\alpha_{i_1}(k)\alpha_{i_2}(k)}{\sqrt{\sum_{k=1}^{N_V}\alpha^2_{i_1}(k)}\sqrt{\sum_{k=1}^{N_V}\alpha^2_{i_2}(k)}}, \tag{4}$$

with $\alpha_i(k)$, $k=1,\ldots,N_V$, indicating the k-th element of the attention vector α_i. The weights w^{CAT} and w^{NUM} in Eq. (2) are set by the user in the interval $[0,1]$ and are normalized in such a way that $w^{CAT}+w^{NUM}=1$. It has been verified that if $w^{CAT} \gg w^{NUM}$, then the reports are clustered solely based on component label, even if the description of the intervention is significantly different. On the opposite, if $w^{CAT} \ll w^{NUM}$, reports with similar description are grouped together even if they are related to two different components. In this work, the two weights are set equal to 0.5 in order to: (i) group together similar maintenance interventions performed on the same type of component, and (ii) not assign to any cluster those maintenance interventions that are very dissimilar from all others performed on the same component.

The cosine distance is used for computing the distance between numerical vectors since the dimensionality of the vector is very large and we are interested in finding vectors with a similar direction in the space, i.e. reports for which the same set of words is relevant [66].

Based on the Gower distance, the clustering technique DBSCAN is applied to identify compact and tight neighborhoods of vectors in the data [12]. The clusters are recognized by searching for vectors of the dataset around which a sufficiently large number of other vectors are present. The search is controlled by the size, i.e. the radius of the circle centered on the candidate vector, and the cardinality of the neighborhood, i.e. the minimum number of vectors that must fall within the circle to consider the vector as a cluster center. After the search, the vectors that are too far from the cluster centers are identified as outliers (noise).

The hyperparameters of the DBSCAN algorithm, i.e. optimal size and cardinality of the neighborhood, are set by trial-and-error. Specifically, a grid search is performed and the obtained clusters are evaluated considering the silhouette metric, which measures how compact and separated the obtained clusters are [67].

In this work, DBSCAN is applied to mixed numerical and categorical data since the component label provides a clear indication to the maintenance operators and system experts on the occurred accident and malfunctions, thus allowing to obtain results which are more accurate and useful in practice. Notice that other solutions are possible, including considering other labels or adding the component labels as words to the description of the event. This last option has been discarded to allow BERT to capture as much as possible the semantic content of the original report without modifying excessively the attention distribution by adding more words to it.

5. Analysis of the clusters

An ad-hoc taxonomy has been developed in collaboration with system experts. It systematically organizes the words of the reports into Q sub-levels τ_q, $q=1,\ldots,Q$. The V_q words of the generic q-th sub-level of the taxonomy will be referred to as $\left\{p^q_1,\ldots,p^q_v,\ldots,p^q_{V_q}\right\}$. Notice that $\cup_{q=1}^{Q}\left\{p^q_1,\ldots,p^q_{V_q}\right\} \neq \{p_1,\ldots,p_v,\ldots,p_V\}$ since not all the V words of the vocabulary of the repository are assigned to the taxonomy (e.g., syntactic words such as conjugations and articles). Table 2 reports the list of the sub-levels of the taxonomy and some examples of the words assigned to them.

Then, the taxonomy is used to identify the characteristics of the clusters. Specifically, for each cluster c_m, $m=1,\ldots,M$, and taxonomy sub-level τ_q, $q=1,\ldots,Q$, we define the matrix $\alpha^q_m = \left[\alpha_{i,v} : r_i \in c_m,\ p_v \in \left\{p^q_1,\ldots,p^q_v,\ldots,p^q_{V_q}\right\}\right] \in \mathbb{R}^{D_m} \times \mathbb{R}^{V_q}$, where D_m is the number of reports r_i assigned to cluster c_m. Its generic element $\alpha_{i,v}$ is the attention given to the word p^q_v of the sub-level τ_q in the report r_i assigned to the cluster c_m. Notice that if a word p^q_v is not present in the report r_i, the corresponding attention $\alpha_{i,v}$ is set to zero. The cluster c_m is, therefore, characterized by the words in $\left\{p^q_1,\ldots,p^q_v,\ldots,p^q_{V_q}\right\}$ belonging to the taxonomy sub-level τ_q, $q=1,\ldots,Q$, with largest associated values in the vectors α^q_m. In this way, it is possible to identify the characteristics of the clusters with respect to each sub-level of the taxonomy, such as the types of components, their working state, the specific failures or malfunctions occurred, the maintenance actions performed and even the personnel that performed the intervention. Also, the vectors $\overline{\alpha^q_m} = \left\{\overline{\alpha^q_{m,i}} : r_i \in c_m\right\} \in \mathbb{R}^{D_m} \times 1$ of the total attention given to the taxonomy sub-level τ_q, $q=1,\ldots,Q$, with $\overline{\alpha^q_{m,i}} = \left[\sum_{v=1}^{V_Q}\alpha_{i,v} : p_v \in \left\{p^q_1,\ldots,p^q_{V_q}\right\}\right]$, can provide additional hints on

Table 2
Developed taxonomy. Some words are related to the "Software" sub-level have been masked for confidentiality.

Sub-level	Description	Words (some examples)
Event	The occurred failure or malfunction	"abnormal", "defective", "leakage"
Component	The involved components, part of components or subsystems	"arrester", "contactor", "pump"
Component state	The health state of the component	"disconnected", "hot", "active"
Maintenance	The performed maintenance actions	"checked", "cleaned", "replaced"
Personnel	The involved personnel when the event occurs or during maintenance	"pilot", "operator", "staff"
Personnel action	The action performed by the personnel	"activate", "followed", "reported"
Software	The software used on the train and at the depot	"screen", "software_1″, "software_2″

the characteristics of the cluster c_m, $m = 1, \ldots, M$. For example, if the total attention to sub-level "Maintenance" in cluster c_{m_1} is larger than in cluster c_{m_2}, it can be expected that the failure and malfunctions that occurred in cluster c_{m_1} required more maintenance actions than the ones occurring in cluster c_{m_2}.

Notice that within the proposed procedure, expert knowledge is used not only for defining the taxonomy but also for the pre-processing of the maintenance reports. Typically, they contain several technical terms such as abbreviations to indicate components and spare parts. Despite the capabilities of BERT to understand text, its fine-tuning on a smaller domain-specific dataset requires to some extent the pre-processing of the data. Expert knowledge plays a key role in this step for addressing the technicalities of the text and allow BERT to adequately understand technical language. The details of the pre-processing actions in the specific case study of this work are given in Section 7.

6. Case study

Railway companies use various systems and methods to collect, store and process data [68]. Specifically, while the train is in operation, failures and malfunctions of electric and mechanical components and systems are recorded and notifications are sent to the train driver and to the maintenance depot. Based on the severity of the event, the train is either allowed to continue the operation, or it is required to reach the depot for inspection and maintenance. Then, a textual report containing the description of the event, its severity, the components involved and the maintenance actions performed is written by the maintenance operators.

In this context, the procedure is applied to a repository of maintenance intervention reports of electrical freight transport trains. The number D of available reports, which is in the order of hundreds, is not provided here for confidentiality reasons. Each report is composed by:

i. A free text written in English by a system operator, which contains the description of the event and of the maintenance actions performed. They contain on average 79 ± 39 words, including many technical words and acronyms. For example, a report of 103 words contains 18 acronyms which refer to the involved components, the personnel and the maintenance activity. The reports are also characterized by several misprints and incomplete sentences;
ii. The severity of the event, which is binary ($S = 1$) and can be "Damage", if one or more components or subsystems have been replaced to solve the problem, or "Anomaly", if the problem has been solved without performing any replacement. Typically, events occur while the train is in service, and these labels are assigned by the maintenance operators after the maintenance activity is performed in the depot. The severities "Damage" will be represented by the label "0″ and "Anomaly" by the label "1″. As expected, the distribution of the reports into the two classes of severity is unbalanced, with a ratio of "Anomaly" to "Damage" of 2 to 1;
iii. The component label, which indicates the component or subsystem that was involved in the event, and which undergoes maintenance.

Table 3 shows an example of a maintenance report of the repository.

7. Procedure development

The D textual reports are first pre-processed performing the following steps:

Table 3
Example of maintenance intervention report of the repository. Some words have been masked using "#" for confidentiality reasons.

Description	Severity	Component label
LP Repor ted that ### got isolated. Helpline g uided ## to normalized from ###, but not normalize, further guided to reset the ###Q. after that normali ze from ### issue not resolved, then restart the loco after that issue not resolve.guided to continue the service is sue will be addressed by ###	1	Circuit Breaker

i. Spell-checking to fix the issues of random deletion of certain characters or random insertion of spaces inside the text, which might have occurred during the digital conversion of the report files (e.g., "p ump" is corrected into "pump");
. Conversion of one-letter acronyms. The reports contain a large number of one-letter technical acronyms to represent components, e.g., "C", "F", "M", "Q" and "R" are used instead of "contactor", "fuse", "motor", "relay" and "resistor", respectively. Since BERT can wrongly model them as part of the adjacent words, the acronyms are converted into their extended form provided by the experts of the company;
. Addition of the acronyms made of more than one letter to the dictionary that BERT learned during pretraining (e.g., "VCB" is "voltage circuit breaker"). The choice of not converting these acronyms in their extended form as in (ii) aims at allowing BERT to learn the technical language used by the operators in the reports;
. Removal of symbols. The symbols "#", "/", "\", "(", ")" are removed since they do not provide useful semantic information for the scope of this work.

Table 4 reports an example of a pre-processed report.

We use the pretrained base form version of BERT taken from [63], characterized by an architecture formed by $N_l = 12$ layers, $N_H = 12$ heads and an embedding length of $N_e = 768$. The total number of internal parameters is equal to 110 million. Typically, the fine-tuning of BERT is performed considering a set of tasks such as masked-word prediction and classification. In this work, the fine-tuning of the pretrained BERT model is performed by adding to the back end of BERT a fully connected linear layer, with the objective of classifying the severity of the accidents using 80 % of the reports of the repository. Although in this case study two classes of severity are considered, the classification model is not restricted to being binary. The performance of the fine-tuned BERT is verified considering the classification of the severity of the remaining 20 % of the reports. Since the final objective of the use of BERT in this work is not the classification of the reports, but their encoding into vectors, the following analysis of the classification performance only aims at verifying the quality of the encoding. In practice, we assume that if the classification accuracy is satisfactory, then the encoding of the reports properly considers their semantic content. To this purpose, the following metrics to assess the classification accuracy

Table 4
Example of a pre-processed report. Some words have been masked using "#" for confidentiality reasons.

Description	Pre-processed description
LP Repor ted that ### got isolated. Helpline g uided ## to normalized from ###, but not normalize, further guided to reset the ###Q. after that normali ze from ### issue not resolved, then restart the loco after that issue not resolve.guided to continue the service is sue will be addressed by ###	LP Reported that ### got isolated. Helpline guided ## to normalized from ###, but not normalize, further guided to reset the ##relay. after that normalize from ### issue not resolved, then restart the loco after that issue not resolve. guided to continue the service issue will be addressed by ###

are considered:

$$A = \frac{Number\ of\ correctly\ classified\ reports}{Number\ of\ reports} \tag{5}$$

and F1- score:

$$F_1 = \frac{2}{\frac{TP+FN}{TP} + \frac{TP+FP}{TP}} \tag{6}$$

where TP, TN, FP and FN are the numbers of true positive, true negative, false positive and false negative classifications, respectively. The F1-score is considered since the dataset is imbalanced, with almost two-thirds of the D reports belonging to the class "Anomaly" [69]. Table 5 reports the obtained results.

The misclassifications of BERT have been investigated with company experts. The analysis has shown that some reports of the repository were mislabeled by the operators. This is because maintenance reports are written by various teams of maintenance operators located in different maintenance depots, who may follow different logics in assigning the severity classes. Specifically, some reports of the repository describe events that have been incorrectly assigned to the severity class "Anomaly", even if they involve failures and replacements of components, which, according to the company standards, should clearly indicate that the severity class is "Damage". Vice versa, some reports are assigned to the severity class "Damage" even though they do not involve any replacement.

The capability of BERT of identifying mislabeled reports can be useful for verifying the assignment of the severity of the event and, thus, properly inform the decision-making to avoid neglecting critical issues and performing unnecessary maintenance.

Then, the attention vectors are extracted using the procedure of Section 3. Once the attention vectors are extracted, the clustering of the reports is performed using DBSCAN (Section 4). The weights w^{CAT} and w^{NUM} used in (2) to define the Gower distance are set to 0.5 to obtain clusters homogeneous with respect to both components and maintenance interventions. The DBSCAN hyperparameters (size and cardinality of the neighborhood) are optimized by adopting a grid-search approach and evaluating the obtained clusters with the silhouette metric [67]. Table 6 reports the results of the hyperparameters optimization.

8. Results

The proposed procedure is validated with respect to the objectives of:

i. Analyzing the obtained clusters considering the words and the sub-levels of the taxonomy with largest associated attentions. This objective aims to identify the common physical/engineering characteristics of the clusters, such as the types of components, their working state, the specific failures or malfunctions occurred, the maintenance actions performed and even the personnel that performed the intervention;
ii. Analyzing the outliers. This objective aims to identify rare events of maintenance intervention that do not belong to any cluster, and, as such, requiring a case-by-case investigation.

8.1. Analysis of the obtained clusters

The proposed clustering method has assigned 92 % of the reports to $M = 20$ clusters. The remaining 8 % of the reports are considered outliers and are not assigned to any cluster. The clusters are analyzed by considering the metrics of homogeneity, $H_{m,x}$, and coverage, $C_{m,x}$, which evaluate the correspondence between the component labels and the clusters. The former computes the fraction of reports assigned to cluster c_m that involve the component of label x:

$$H_{m,x} = \frac{D_{m,x}}{D_m} \tag{7}$$

where $D_{m,x}$ is the number of reports assigned to the cluster c_m with component label x and D_m is the number of reports assigned to the clusters c_m. The latter is the fraction of reports involving the component of label x that are assigned to the cluster c_m,

$$C_{m,x} = \frac{D_{m,x}}{D_x} \tag{8}$$

where D_x is the number of reports with component label x in the repository, $c_m \in \{c_1, \ldots, c_M\}$ and $x \in \{1, \ldots, X\}$.

Tables 7 and 8 report the obtained results. It can be noticed that all the M clusters are homogeneous with respect to a specific component label, whereas the three clusters c_2, c_3 and c_{15} do not contain all the reports assigned to their corresponding component labels. It can also be noticed that the reports that have not been assigned to any cluster are associated with many different components, i.e. their homogeneity is not equal to one, and that their coverage for some of the component labels is equal to one. This indicates that these are either reports associated to components which rarely fail, or reports associated to rare issues to more common components. In Section 8.2 these behaviors are further investigated.

Then, the clusters are analyzed following the procedure of Section 5. Figs. 3 and 4 show, as example, the obtained distributions for clusters c_7 and c_{12}, respectively.

From the analysis of Figs. 3 and 4, it can be noticed that cluster c_7 is related to events ("*issue*", "*fault*") involving electric components ("*diodes*", "*cb*", i.e. contactors) of the traction control subsystem ("*tcu*", i.e. traction control unit) noticed by the train driver ("*lp*", i.e. locomotive pilot) and which were solved by electrically isolating the faulty parts and continuing the train operation using redundant systems ("*normalize*"). Cluster c_{12} mainly involves issues to the oil pump ("*oil*", "*pump*") which were detected during inspections ("*check*", "*find*") by the maintenance operators ("*ts*", i.e. depot technician) by observing an abnormal behavior ("*vibration*") of the component.

8.2. Analysis of the outliers

With respect to the outliers, they can be originated by:

i Events that involve components that typically do not fail or experience malfunctions;
ii Rare events or unusual maintenance interventions, which involve components that typically experience different types of problems.

To distinguish between these two different types of outliers, it is possible to consider the labels of the components. The procedure adopted in this work considers the coverage metric described in Section 8.1. First, the clusters whose coverage metric is smaller than 1, which

Table 5
Classification performances.

Measure	Obtained values
Accuracy	0.904
F1-score	0.881

Table 6
Hyperparameters of DBSCAN.

Technique	Hyperparameter	Considered values	Selected value
DBSCAN	Size of the neighborhood	[0.001, 0.011,…, 0.5]	0.49
	Cardinality of the neighborhood	[2, 3,…, 50]	5

Table 7
Homogeneity, $H_{m,x}$, of the obtained clusters c_m, $c_m \in \{c_1, \ldots, c_M\}$ and $x \in \{1, \ldots, X\}$. The missing rows refer to clusters that have homogeneity 1 of a component label.

		Component labels							
		1	2	3	…	20	21	…	X
Clusters	c_1	1	0	0	…	0	0	…	0
	c_2	0	1	0	…	0	0	…	0
	c_3	0	0	1	…	0	0	…	0
	…	…	…	…	…	…	…	…	…0
	c_{20}	0	0	0	…	1	0	…	0
	Not assigned	0	0.013	0.013	…	0.013	0.013	…	0.027

Table 8
Coverage, $C_{m,x}$, of the obtained clusters c_m, $c_m \in \{c_1, \ldots, c_M\}$ and $x \in \{1, \ldots, X\}$. The missing rows refer to clusters that have coverage 1 of a component label.

		Component labels									
		1	2	3	…	15	…	20	21	…	X
Clusters	c_1	1	0	0	…	0	…	0	0	…	0
	c_2	0	0.923	0	…	0	…	0	0	…	0
	c_3	0	0	0.9	…	0	…	0	0	…	0
	…	…	…	…	…	…	…	…	…	…	0
	c_{15}	0	0	0	…	0.933	…	0	0	…	0
	…	…	…	…	…	…	…	…	…	…	0
	c_{20}	0	0	0	…	0	…	1	0	…	0
	Not assigned	0	0.077	0.1	…	0.067	…	0	1	…	1

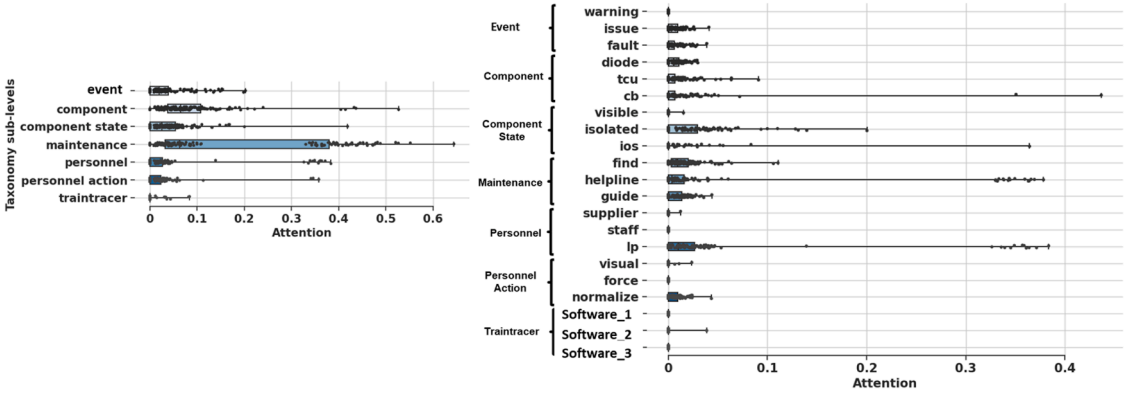


Fig. 3. (Left) Distribution of total attention given to the taxonomy sub-levels for reports assigned to the cluster c_7. (Right) Distribution of the attentions given to the three words with the largest average attention of each sub-level. The distributions are displayed by using boxplots, which show the minimum, the first quartile, the median, the third quartile and the maximum of the distributions.

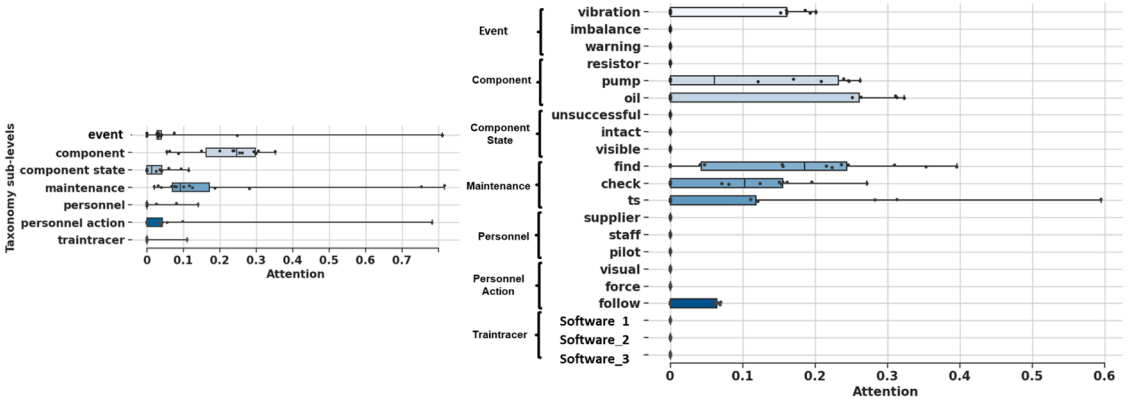


Fig. 4. (Left) Distribution of total attention given to the taxonomy sub-levels for reports assigned to the cluster c_{12}. (Right) Distribution of the attentions given to three words with the largest average attention of each sub-level. The distributions are displayed by using boxplots, which show the minimum, the first quartile, the median, the third quartile and the maximum of the distributions.

indicates that there are some reports of events involving the cluster-specific component label which have not been assigned to the cluster, are identified (Table 8). Then, these reports are analyzed by the experts of the company to confirm the type of outlier. For example, all reports associated to the converter of an auxiliary power module are assigned to cluster c_2 except one labelled as outlier. According to the experts of the company, the maintenance intervention described in the report of the outlier is completely different from the maintenance interventions described in the reports of c_2, in which the maintenance technicians were unable to solve the issues following the standard maintenance procedures. Similarly, all the reports associated to another power module are assigned to cluster c_{15}, except one labelled as outlier. Also in this case, the experts of the company have confirmed that the report of the outlier describes a rare issue, which the maintenance technician was not able to properly address and required the complete replacement of the component.

9. Conclusions

In this work, we have presented a systematic procedure for analyzing maintenance reports to support maintenance decision-making for a fleet of similar systems. The procedure combines the attention mechanism of BERT and the clustering method of DBSCAN to find homogeneous groups of maintenance interventions performed on a fleet of similar systems. A taxonomy of the words is used to identify the characteristics of the clusters, such as the involved components, their working and failure states, their failure and malfunctions, the maintenance actions and the personnel involved in the intervention. An application is worked out with regards to a repository of reports of maintenance interventions of traction systems of a fleet of trains. The obtained results demonstrate that the proposed procedure is able to:

i. Find clusters homogeneous with respect to the type of components involved in the maintenance interventions;
ii. Effectively characterize the obtained clusters using the words and the sub-levels of the taxonomy with largest associated attention;
iii. Identify rare failure modes, malfunctions and maintenance interventions.

From a practical perspective, the outcomes of the analysis of the maintenance reports by the proposed procedure can be used to support maintenance decision-making in terms of:

i. Discovering possible mislabelling of event severity, which can lead to improper or unnecessary maintenance interventions or overlooking of critical issues;
ii. Extracting statistical information (e.g., estimates of the values of the rates of specific failure modes) useful for reliability analysis and risk assessment to inform maintenance planning;
iii. Developing or revising maintenance procedures for cases of rare events of maintenance intervention;
iv. Automatically labelling signal data with the event type as identified by the cluster membership, for practical feasibility of training data-driven fault diagnostic and prognostic models.

Some limitations of the proposed procedure are:

1. The taxonomy used to incorporate expert knowledge does not allow representing causal and logic relations between its sub-levels;
2. The subjectivity of the taxonomy, which is built by field experts according to their knowledge;
3. The procedure for the analysis of the clusters, which requires the intervention of the expert for the analysis of the words with the largest attention for each sublevel of the taxonomy;
4. The difficulty of BERT to model technical language; the extensive use of technical terms in the reports (e.g., acronyms) requires a phase of text pre-processing.

Despite these limitations, this work offers a valuable procedure whose outcomes can help improve maintenance decision-making. There is, undoubtedly, potential for future refinement and broader application. Specifically, industry-standard taxonomies will be considered to reduce the subjectivity and enrich the description of the clusters. For example, the sub-level of the taxonomy proposed in this work related to the component can be further expanded using existing industry-standard taxonomies, which contain more detailed sub-levels, e.g., location, use and type of equipment. Also, the possibility of replacing taxonomies with ontologies and KGs will be investigated. This is expected to provide a more robust and nuanced representation of domain knowledge. Future work will explore using also LLMs. This is expected to allow reducing the efforts of extensive text pre-processing and to enhance the capability of understanding technical language.

CRediT authorship contribution statement

Dario Valcamonico: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Piero Baraldi:** Writing – review & editing, Supervision, Methodology, Conceptualization. **July Bias Macêdo:** Software, Methodology, Formal analysis. **Márcio Das Chagas Moura:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Jonathan Brown:** Validation, Data curation, Conceptualization. **Stéphane Gauthier:** Writing – review & editing, Validation, Project administration, Data curation, Conceptualization. **Enrico Zio:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Piero Baraldi reports financial support was provided by FAIR (Future Artificial Intelligence Research) project. Enrico Zio reports financial support was provided by European Union Horizon 2020 research and innovation program.

Acknowledgments

The participation of Enrico Zio to the project is supported by European project RECET4RAIL which has received funding from the European Union Horizon 2020 Research and Innovation Program under grant agreement No: 101015423. The participation of Piero Baraldi to the project is supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence).

Data availability

The data that has been used is confidential.

References

[1] Chen L, Bai Q. Optimization in decision making in infrastructure asset management: a review. Appl Sci 2019;9. https://doi.org/10.3390/app9071380.
[2] Ittoo A, Nguyen LM, Van Den Bosch A. Text analytics in industry: challenges, desiderata and trends. Comput Ind 2016;78:96–107. https://doi.org/10.1016/j.compind.2015.12.001.
[3] Brundage MP, Sexton T, Hodkiewicz M, Dima A, Lukens S. Technical language processing: unlocking maintenance knowledge. Manuf Lett 2021;27:42–6. https://doi.org/10.1016/j.mfglet.2020.11.001.

[4] Adnan K, Akbar R. Limitations of information extraction methods and techniques for heterogeneous unstructured big data. Int J Eng Bus Manag 2019;11:1–23. https://doi.org/10.1177/1847979019890771.

[5] Campos J, Sharma P, Gabiria UG, Jantunen E, Baglee D. A big data analytical architecture for the asset management. In: Proceedings of the 9th CIRP IPSS conference: circular perspective on product/service-systems. 64; 2017. p. 369–74. https://doi.org/10.1016/j.procir.2017.03.019.

[6] Zhong K, Jackson T, West A, Cosma G. Natural Language processing approaches in industrial maitenance: a systematic literature review. Procedia Comput Sci 2024; 232:2082–97. https://doi.org/10.1016/j.procs.2024.02.029.

[7] Zhao W.X., et al., "A survey of large language models," arXiv, 2024, doi:10.48550/arXiv.2303.18223.

[8] J. Zeng et al., "Large language models for social networks: applications, challenges and solutions," arXiv, 2024, doi:10.48550/arXiv.2401.02575.

[9] Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature 2023;624:570–8. https://doi.org/10.1038/s41586-023-06792.

[10] Macêdo JB, Moura C, Aichele D, Lins ID. Identification of risk features using text mining and BERT-based models_ application to an oil refinery. Process Saf Environ Prot 2022;158:382–99. https://doi.org/10.1016/j.psep.2021.12.025.

[11] Ramos PMS, Macedo JB, Maior CBS, das C. Moura M, Lins ID. Combining BERT with numerical features to classify injury leave based on accident description. Proc Inst Mech Eng Part O J Risk Reliab 2022. https://doi.org/10.1177/1748006X221140194.

[12] Khan K, Rehman SU, Aziz K, Fong S, Sarasvady S. DBSCAN : past, present and future. In: Proceedings of the international conference on the applications of digital information and web technologies (ICADIWT); 2014. p. 232–8. https://doi.org/10.1109/ICADIWT.2014.6814687.

[13] Nickerson RC, Varshney U, Muntermann J. A method for taxonomy development and its application in information systems. Eur J Inf Syst 2013;22(3):336–59. https://doi.org/10.1057/ejis.2012.26.

[14] Gonzalez-Carvajal S, Garrido-Merchan EC. Comparing BERT against traditional machine learning text classification. J Comput Cogn Eng 2023;2(4). https://doi.org/10.47852/bonviewJCCE3202838.

[15] Galassi A, Lippi M, Torroni P. Attention in natural language processing. IEEE Trans Neural Netw Learn Syst 2021;32(10):4291–308. https://doi.org/10.1109/TNNLS.2020.3019893.

[16] R. Saha, "Influence of various text embeddings on clustering performance in NLP," arXiv, 2023, doi:10.48550/arXiv.2305.03144.

[17] Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl Based Syst 2023;263. https://doi.org/10.1016/j.knosys.2023.110273.

[18] Danilevsky M, Qian K, Aharonov R, Katsis Y, Kawas B, Sen P. A survey of the state of explainable AI for natural language processing. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. 2020. p. 447–59. https://doi.org/10.18653/v1/2020.aacl-main.46.

[19] International Standards Organization (ISO), Petroleum and natural gas industries - Collection and exchange of reliability and maintenance data for equipment. ISO 14224:2016. https://www.iso.org/standard/64076.html.

[20] Bhardwaj U, Teixeria AP, Soares CGuedes. Casualty analysis methodology and taxonomy for FPSO accident analysis. Reliab Eng Syst Saf 2022;218. https://doi.org/10.1016/j.ress.2021.108169.

[21] Stenström C, Aljumaili M, Parida A. Natural language processing of maintenance records data. Int J COMADEM 2015;18(2):33–7.

[22] Chen Z, Huang K, Wu L, Zhong Z, Jiao Z. Relational graph convolutional network for text-mining based accident causal classification. Appl Sci 2022;12(5). https://doi.org/10.3390/app12052482.

[23] Yang Z, Baraldi P, Zio E. A novel method for maintenance record clustering and its application to a case study of maintenance optimization. Reliab Eng Syst Saf 2020; 203:107103. https://doi.org/10.1016/j.ress.2020.107103. April.

[24] Bouabdallaoui Y, Lafhaj Z, Yim P, Ducoulombier L, Bennadji B. Natural language processing model for managing maintenance requests in buildings. Buildings 2020; 10(9):1–12. https://doi.org/10.3390/BUILDINGS10090160.

[25] Zhou D, Zhuang X, Cai J, Zuo H, Zhao X, Xiang J. An ensemble model using temporal convolution and dual attention gated recurrent unit to analyze risk of civil aircraft. Expert Syst Appl 2023;236. https://doi.org/10.1016/j.eswa.2023.121423.

[26] Wang B, Zhao J. Automatic frequency estimation of contributory factors for confined space accidents. Process Saf Environ Prot 2022;157:193–207. https://doi.org/10.1016/j.psep.2021.11.004.

[27] Usuga Cadavid JP, Grabot B, Lamouri S, Pellerin R, Fortin A. Valuing free-form text data from maintenance logs through transfer learning with CamemBERT. Enterp Inf Syst 2022;16(6). https://doi.org/10.1080/17517575.2020.1790043.

[28] Williams T, Betak J. A comparison of LSA and LDA for the analysis of railroad accident text. Procedia Comput Sci 2018;130:98–102. https://doi.org/10.1016/j.procs.2018.04.017.

[29] Sala R, Pirola F, Pezzotta G, Cavalieri S. NLP-based insights discovery for industrial asset and service improvement: an analysis of maintenance reports. IFAC PapersOnLine 2022;55(2):522–7. https://doi.org/10.1016/j.ifacol.2022.04.247.

[30] Xiong M, Wang H, Che C, Sun M. Application of text mining and coupling theory to depth cognition of aviation safety risk. Reliab Eng Syst Saf 2024;245:110032. https://doi.org/10.1016/j.ress.2024.110032.

[31] Usuga-Cadavid JP, Lamouri S, Grabot B, Fortin A. Using deep learning to value free-form text data for predictive maintenance. Int J Prod Res 2022;60(14): 4548–75. https://doi.org/10.1080/00207543.2021.1951868.

[32] Bin C, Baigen C, Wei S. Text mining in fault analysis for on-board equipment of high-speed train control system. In: Proceedings of the Chinese automation congress (CAC); 2017. p. 6907–11. https://doi.org/10.1109/CAC.2017.8244022.

[33] Song B, Zhang Z, Qin Y, Liu X, Hu H. Quantitative analysis of freight train derailment severity with structured and unstructured data. Reliab Eng Syst Saf 2022;224. https://doi.org/10.1016/j.ress.2022.108563.

[34] Tahvili S, Hatvani L, Felderer M, Afzal W, Bohlin M. Automated functional dependency detection between test cases using Doc2Vec and clustering. In: Proceedings of the IEEE international conference on artificial intelligence testing (AITest); 2019. p. 19–26. https://doi.org/10.1109/AITest.2019.00-13.

[35] Li Y, Wang H, Sun M. ChatGPT-like large-scale foundation models for prognostics and health management: a survey and roadmaps. Reliab Eng Syst Saf 2024;243. https://doi.org/10.1016/j.ress.2023.109850.

[36] Mo Y, Zhao D, Du J, Syal M, Aziz A, Li H. Automated staff assignment for building maintenance using natural language processing. Autom Constr 2020;113:103150. https://doi.org/10.1016/j.autcon.2020.103150. no. November 2019.

[37] Arif-Uz-Zaman K, Cholette ME, Karim A. Extracting failure time data from industrial maintenance records using text mining. Adv Eng Inform 2017;33: 388–96. https://doi.org/10.1016/j.aei.2016.11.004.

[38] Peshave A, Aggour KS, Ali A, Mulwad V, Dixit S, Saxena A. Evaluating vector representations of short text data for automating recommendations of maintenance cases. In: Proceedings of the annual conference of the PHM society; 2022. https://doi.org/10.36001/phmconf.2022.v14i1.3196.

[39] He Y, Li J. Analysis of coal mining accident risk factors based on text mining. Proc Inst Mech Eng Part O J Risk Reliab 2024. https://doi.org/10.1177/1748006X241245579.

[40] Zheng S, Pan K, Liu J, Chen Y. Empirical study on fine-tuning pre-trained large language models for fault diagnosis of complex systems. Reliab Eng Syst Saf 2024; 252. https://doi.org/10.1016/j.ress.2024.110382.

[41] Macedo JB, Ramos PMS, Maior CBS, Moura MJC, Lins ID. Human factor identification in aviation accidents using contextual word embeddings. In: Proceedings of the 33rd European safety and reliability conference (ESREL 2023); 2023. https://doi.org/10.3850/978-981-18-8071-1_P233-cd.

[42] Valcamonico D, Baraldi P, Zio E, Decarli L, Crivellari A, Rosa LL. Combining Natural Language Processing and Bayesian networks for the probabilistic estimation of the severity of process safety events in hydrocarbon production assets. Reliab Eng Syst Saf 2024;241. https://doi.org/10.1016/j.ress.2023.109638.

[43] Rose RL, Puranik TG, Mavris DN, Rao AH. Application of structural topic modeling to aviation safety data. Reliab Eng Syst Saf 2022;224. https://doi.org/10.1016/j.ress.2022.108522.

[44] Single JI, Schmidt J, Denecke J. Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. Saf Sci 2020;129. https://doi.org/10.1016/j.ssci.2020.104747.

[45] Chen M, Qu R, Fang W. Case-based reasoning system for fault diagnosis of aero-engines. Expert Syst Appl 2022;202. https://doi.org/10.1016/j.eswa.2022.117350.

[46] Liu J, Schmid F, Li K, Zheng W. A knowledge graph-based approach for exploring railway operational accidents. Reliab Eng Syst Saf 2021;207. https://doi.org/10.1016/j.ress.2020.107352.

[47] Liu C, Yang S. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. Expert Syst Appl 2022;207. https://doi.org/10.1016/j.eswa.2022.117991.

[48] Li C, Yang X, Luo S, Song M, Li W. Towards domain-specific knowledge graph construction for flight control aided maintenance. Appl Sci 2022;12(24). https://doi.org/10.3390/app122412736.

[49] Zheng X, Wang B, Zhao Y, Mao S, Tang Y. A knowledge graph method for hazardous chemical management: ontology design and entity identification. Neurocomputing 2021;430:104–11. https://doi.org/10.1016/j.neucom.2020.10.095.

[50] Stewart M, Hodkiewicz M, Liu W, French T. MWO2KG and Echidna: constructing and exploring knowledge graohs from maintenance data. Proc Inst Mech Eng Part O J Risk Reliab 2022. https://doi.org/10.1177/1748006X221131128.

[51] Yang J, et al. BERT and hierarchical cross attention-based question answering over bridge inspection knowledge graph. Expert Syst Appl 2023;233. https://doi.org/10.1016/j.eswa.2023.120896.

[52] Chen Y, Zhang L, Hu J, Liu Z, Xu K. Emergency response recommendation for long-distance oil and gas pipeline based on an accident case representation model. J Loss Prev Process Ind 2022;77. https://doi.org/10.1016/j.jlp.2022.104779.

[53] Kamil MZ, Taleb-Berrouane M, Khan F, Amyotte P, Ahmed S. Textual data transformations using natural language processing for risk assessment. Risk Anal 2023:1–20. https://doi.org/10.1111/risa.14100.

[54] Gao Y, Woods C, Liu W, French T, Hodkiewicz M. Pipeline for machine reading of unstructured maintenance work order records. In: Proceedings of the 30th European safety and reliability conference, ESREL 2020 and 15th probabilistic safety assessment and management conference, PSAM 2020; 2020. p. 1401–8. https://doi.org/10.3850/978-981-14-8593-0.

[55] Ricketts J, Pelham J, Barry D, Guo W. An NLP framework for extracting causes, consequences, and hazards from occurrence reports to validate a HAZOP study. In: Proceedings of the AIAA/IEEE digital avionics systems conference; 2022. https://doi.org/10.1109/DASC55683.2022.9925822.

[56] Yan K, Wang Y, Jia L, Wang W, Lui S, Geng Y. A content-aware corpus-based model for analysis of marine accidents. Accid Anal Prev 2023;184. https://doi.org/10.1016/j.aap.2023.106991.

[57] Mandelli D, Wang C. A model-based approach to extract health information from textual data. In: Proceedings of the annual conference of the PHM society. 14; 2022. https://doi.org/10.36001/phmconf.2022.v14i1.3249.

[58] Batres R, Fujihara S, Shimada Y, Fuchino T. The use of ontologies for enhancing the use of accident information. Process Saf Environ Prot 2014;92(2):119–30. https://doi.org/10.1016/j.psep.2012.11.002.
[59] Abu-Salih B. Domain-specific knowledge graphs: a survey. J Netw Comput Appl 2021;185. https://doi.org/10.1016/j.jnca.2021.103076.
[60] Ehrmann M, Hamdi A, Pontes EL, Romanello M, Doucet A. Named entity recognition and classification on historical documents: a survey. ACM Comput. Surv. 2021;56. https://doi.org/10.1145/3604931.
[61] Naqvi SMR, Ghufran M, Meraghni S, Varnier C, Nicod JM, Zerhouni N. Generating semantic matches between maintenance work orders for diagnostic decision support. In: Proceedings of the annual conference of the PHM society; 2022. https://doi.org/10.36001/phmconf.2022.v14i1.3241.
[62] Vaswani A, et al. Attention is all you need. In: Proceedings of the 31st conference on neural information processing systems (NIPS 2017); 2017. https://doi.org/10.48550/arXiv.1706.03762.
[63] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL HLT 2019 - 2019 conference of the North American chapter of the association for computational linguistics: human language technologies - proceedings of the conference. 1; 2019. p. 4171–86. https://doi.org/10.18653/v1/N19-1423.
[64] Vig J, Belinkov Y. Analyzing the structure of attention in a transformer language model. In: Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP; 2019. p. 63–76. https://doi.org/10.18653/v1/W19-4808.
[65] van de Velden M, Iodice D'Enza A, Markos A. Distance-based clustering of mixed data. Wiley Interdiscip Rev Comput Stat 2019;11(3):1–12. https://doi.org/10.1002/wics.1456.
[66] Walkowiak T, Gniewkowski M. Evaluation of vector embedding models in clustering of text documents. In: Proceedings of the international conference of Recent advances in natural language processing, RANLP; 2019. p. 1304–11. https://doi.org/10.26615/978-954-452-056-4_149.
[67] Shahapure KR, Nicholas C. Cluster quality analysis using silhouette score. In: Proceedings of the 2020 IEEE 7th international conference on data science and advanced analytics, DSAA 2020; 2020. p. 747–8. https://doi.org/10.1109/DSAA49011.2020.00096.
[68] Dong K, Romanov I, McLellan C, Esen AF. Recent text-based research and applications in railways: a critical review and future trends. Eng Appl Artif Intell 2022;116:105435. https://doi.org/10.1016/j.engappai.2022.105435.
[69] Pereira J, Saraiva F. A comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection. In: Proceedings of the 2020 IEEE congress on evolutionary computation (CEC); 2020. p. 1–8. https://doi.org/10.1109/CEC48606.2020.9185822.