EXTRACTING KNOWLEDGE from maintenance reports

Using Natural Language Processing

Data-driven maintenance services are becoming increasingly diffused in the manufacturing sector. Whilst it is true that the analysis of signals coming from industrial assets makes it possible to identify their health status and make decisions regarding the execution of specific maintenance activities, it is also true that another great source of useful data for maintenance workers can be found in the reports com-

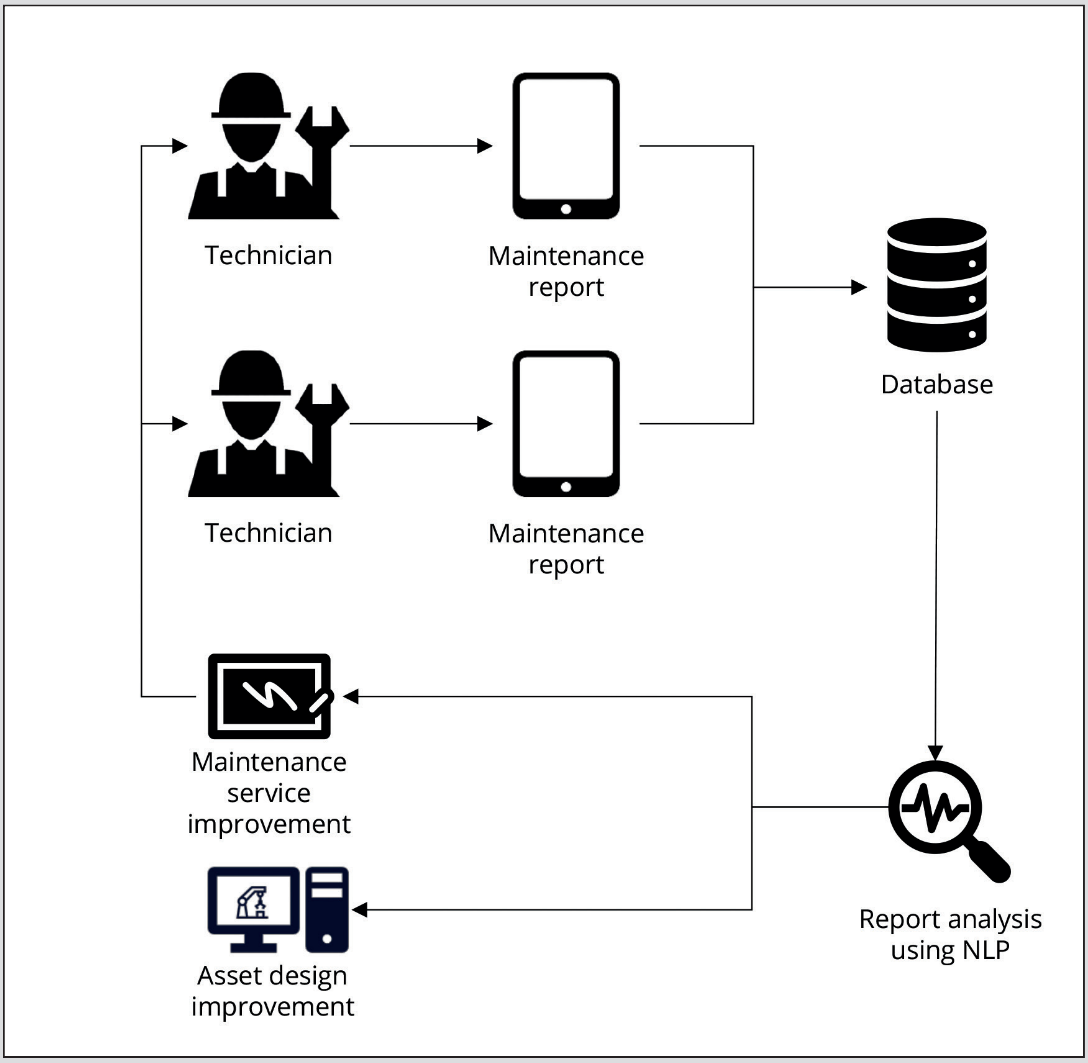

Figure 1. Creation and re-use of knowledge to improve maintenance service and asset design

piled by technicians, which contain information regarding the faults' cause and the resolution method. However, the availability of such data is often underestimated by companies who, due to lack of time or resources, do not analyze reports with lengthy textual descriptions in detail, but often only exploit them for billing or managing disputes with customers.

In this context, it is useful to analyze how Natural Language Processing (NLP) can help solve the limitations associated with the analysis of maintenance reports, allowing companies to transform the unstructured text into a set of structured data that can support the generation of new knowledge in the scope of improving maintenance service and industrial asset design (Figure 1).

One of the most common applications of Natural Language Processing is "topic modelling", which aims to indicate the topics dealt with in a series of documents without the need (for the user) for a detailed reading of each one. Topic modelling can be useful for identifying, within unstructured reports, the most common problems faced by technicians.

Methodologically, topic modelling requires various steps that must be performed in a more or less recursive man-

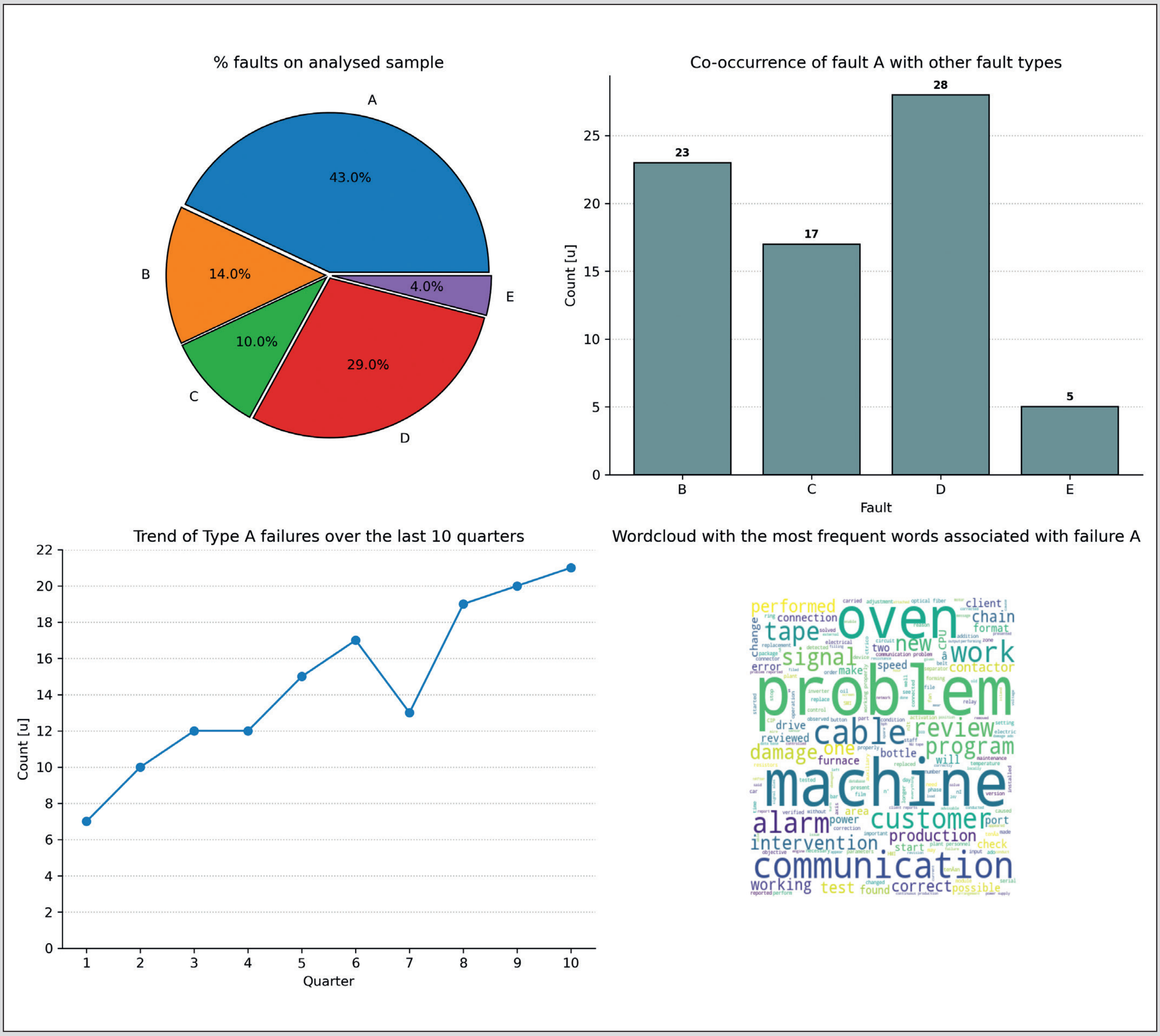

Figure 2. Example of the analyses that can be performed following topic assignment

ner. As in all data analysis activities, a data pre-processing phase must be performed, which in turn consists of several sub-activities. Among the most common, we can mention "tokenization", which consists of subdividing the text into elementary units (individual words) called tokens, preparatory to the activities of "stop word removal", i.e. the removal of all those very common words in a language (e.g. the words "the" or "you") that do not provide useful information for analysis, and the "lemmatization" phase, i.e. the reduction of a word to its basic form (e.g. the word "engines" becomes "engine").

These steps serve to clean up and standardise the text to make it easier for the algorithm to discover recurring words and patterns to identify topics within the text. Another activity consists of the identification of n-grams, i.e., sets of n words that appear consecutively in the text with a certain frequency and that help contextualise its content. For example, the word "claw" gives a certain amount of information since there may be several claws in an asset, the n-gram "claw X" on the other hand makes it immediately clear which claw is being referred to. It is important to emphasise that the pre-processing phase is a step that is performed recursively and may require several steps before a satisfactory result is obtained.

Another activity consists in assigning each token a weight that describes its contribution to contextual understanding, an activity that can be carried out by exploiting functions such as "term frequency-inverse document frequency" (Tf-Idf). For example, a word that appears very often in all analysed texts will be assigned a low weight because it does not help contextualise a document. Conversely, a word that appears only in a small circle of documents will be assigned a higher specific weight because it helps to understand the context of the document. An example would be the word "engine", which will only appear in documents in which the technician addresses an engine-related problem.

Once the dataset has been prepared, it is possible to start the actual topic modelling activity. As mentioned, the objective is to identify patterns of recurring words to cluster them and identify topics. One of the most widespread algorithms to perform this activity is the Latent Dirichlet Allocation (LDA) algorithm which, based on certain user inputs (e.g., the number of topics to be searched for), assigns to each document a percentage expressing the probability that a certain topic is discussed in that document. First, it is necessary to define the list of topics, an activity that is done by indicating to the algorithm the number of topics to be searched in the text and analysing their content to verify their correctness. As with the pre-processing phase, it is necessary to perform this activity recursively until a satisfactory result is obtained (e.g., by varying the number of topics to search for or other parameters).

Once the topics have been identified, the previously defined LDA model can be used to perform the analysis of the dataset. As mentioned, the LDA algorithm will assign to each document a percentage corresponding to the probability that this document deals with one or more of the identified topics. This value can then be exploited to group documents based on the assigned topics by performing (for example) frequency analysis, co-occurrence, or studying the temporal distribution of topics, resolution strategies, geographical areas where certain problems occur more often, and other aspects of interest to the company. Some of the possible analyses are shown in Figure 2.

As shown in Figure 1, the creation of such knowledge can be useful for making improvements in the way the company handles service requests, plans interventions, or designs assets and their components. Based on the output of the analysis, the company will then be able to make decisions about modifying certain maintenance policies (e.g., performing checks or executing preventive activities more frequently) or varying the approach used to handle certain problems. Similarly, some components may be redesigned to prevent too frequent failures due to specific operating conditions. □

Roberto Sala,
Research fellow, Department of Management, Information and Production Engineering, University of Bergamo

Fabiana Pirola,
Assistan professor, Department of Management, Information and Production Engineering, University of Bergamo

Giuditta Pezzotta,
Associate professor, Department of Management, Information and Production Engineering, University of Bergamo