We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

6,100
Open access books available

167,000
International authors and editors

185M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

BOOK CITATION INDEX
CLARIVATE ANALYTICS
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Chapter

Data Mining in Electrical Machine Maintenance Reports

Karlis Athanasios, Falekas Georgios,
Verginadis Dimosthenis and Jose A. Antonino-Daviu

Abstract

Industrial electrical machine maintenance logs pertinent information, such as fault causality and earlier indications, in the form of a semi-standardized report, previously written and now in digital form. New practices in predictive maintenance, state-of-the-art condition monitoring, include increasing applications of machine learning. Reports contain a large volume of natural text in various languages and semantics, proving costly for feature extraction. This chapter aims to present novel techniques in information extraction to enable literature access to this untapped information reserve. A high level of correlation between text features and fault causality is noted, encouraging research for extended application in the scope of electrical machine maintenance, especially in artificial intelligence indication detection training. Furthermore, these innovative models can be used for decision-making during the repair. Information from well-trained classifiers can be extrapolated to advance fault causality understanding.

Keywords: artificial intelligence, big data applications, computer aided engineering, condition monitoring, deep learning, electrical machines, industry 4.0, knowledge acquisition, predictive maintenance, supervised learning, text mining

1. Introduction

Maintenance of electrical machines (EMs) follows industry-established procedures according to subject type and encompasses measurements, maintenance efforts' results, and technician expertise, all of which are necessarily logged in what is called a maintenance report (MR). Widespread information storage forms include numerical and audiovisual data. However, this data is almost always accompanied by a natural text (NT) in the industry's regional language form, to provide context. Each fault typically follows a similar pattern, observations of which are logged in the sequential flow or keywords in the natural text, containing information about degradation procedures, fault causalities, and any similar relevant comments. These semantics are naturally followed and produced by human cognition, effortlessly granting the reader understanding about the fault type and solution, generating patterns.

Numerical data is the main contribution of an MR. Values of interest are tracked according to well-established sensor hardware configuration standards and expertise. However, a standalone number is not enough to assess equipment condition. For example, low stator winding insulation ohmic resistance could indicate either particle contamination or thermal aging [1]. Additional information is required to collapse this quantitative observation into a qualitative one. The industry tends to be the easiest additional experiment, which in this case is a visual inspection (VI). A logical avenue to log such information is in one of three ways: videos or pictures, evidential objects, and NT. Since insertion of observation in the already existing report form as NT is the easiest, fastest, and cheapest solution, it is almost always preferred. Evidence and media are preserved for more dire situations and only when necessary.

NT is a convenient medium for human cognition. We are naturally trained to extract semantics and information from a high redundancy medium, our language, spoken or written. Furthermore, parallels and conclusions regarding causality and precedence can be drawn seamlessly, and transferred via the critical assessment of the responsible engineer to the interested party. These correlations are often made without conscious effort and therefore vary throughout. Thus, expertise is realized and advanced throughout the years.

The advent of Industry 4.0 has brought novel tools and techniques capable of expanding the understanding, processing, and handling of systems and procedures. Heralding the emergence of Big Data, interconnectivity, and digitization, each observation or case study can aid in establishing accurate correlations and training science's newest and sharpest tool, Artificial Intelligence (AI). AI relies on copious amounts of information in the form of input–output numeric pairs. With this outlook, NT proves costly for statistical processing, resulting in a significant volume of unused data.

Natural Language Processing (NLP) is an interdisciplinary sub-field of computer science, artificial intelligence, and linguistics, with the aim to quantify semantics and information in NT. NLP techniques are increasingly being investigated in broader literature for processing and understanding this data. Applications are rapidly multiplying in the past 5 years, facilitated by combined efforts and new hardware. These now robust methodologies are ready to be investigated and applied in numerous fields, such as EM condition monitoring (CM). In combination with traditional numerical causality analysis, graphs depicting common patterns and decision trees can be composed [2]. Case studies are of paramount importance to not only aid in the optimization of concurrent techniques by means of additional input, but also draw attention and confirm results.

Specifically concerning EM CM, such an attempt has not been extensively researched in literature, to the best of the author's knowledge. Conceptually, this discipline provides two important facilitations: extensive expert knowledge and established procedures. This chapter aims to further the establishment of an innovative concept for NLP in the environment of industrial EM CM. A vision of this work is synergy between experts and machinery, in the form of understanding the context of observations, asking for further information, and then providing a verdict, which is the typical procedure undertaken in the industry with the limitations explained above. AI can automate report cognition and event causality graph production. The authors consider that this endeavor will aid the acceleration of PM in the new industrial paradigm by enabling access to a previously inaccessible vault of information. An overview of similar endeavors in a broader context can be found in [3].
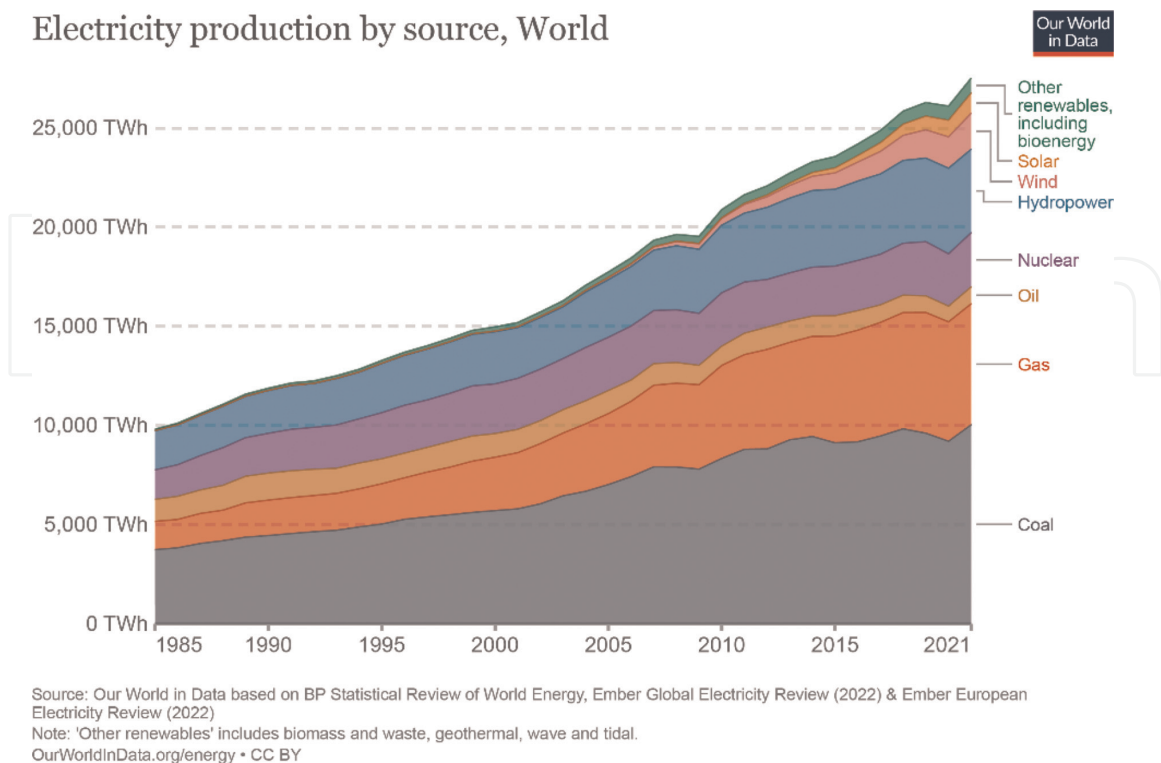
2. Standard diagnostic procedure

EMs are the primary mobilization force of industry and electricity generation especially. **Figure 1** presents global electricity production [5]. Apart from solar and other renewables, the rest of the sources are utilized as kinetic sources for turning the rotor of a synchronous generator (SG), or an induction machine (IM) in the case of newer wind turbines. Therefore, 95% of global energy includes SGs in its production chain.

Hence, proper operation with minimal losses and downtime is of paramount importance. Even operation under sub-optimal power factor -increased losses- should be avoided, as a considerable amount of energy is wasted. EMs consist of a multitude of electromechanical parts which can be ailed by various faults, with their severity ranging from minimal (power factor reduction) to catastrophic (destructive failure). CM tackles the possibility of these faults by their statistical order of appearance. **Figure 2** depicts fault distributions in EMs [6]. Insulation faults represent the highest share in large industrial SGs and therefore attract research focus, followed by the bearings. Rotor faults present a universal, constant appearance. The cited research agrees with similar surveys done by EPRI and IEEE [7].

Therefore, the EM insulation system and especially that of the stator must face and withstand various faults while remaining reliable for the EM to stay healthy and optimal in its operation. The typical EM stator is similar in both SGs and IMs and consists of [8]:

- the copper conductors, which must have a large enough cross-section in order to carry the required current without overheating;

Source: Our World in Data based on BP Statistical Review of World Energy, Ember Global Electricity Review (2022) & Ember European Electricity Review (2022)
Note: 'Other renewables' includes biomass and waste, geothermal, wave and tidal.
OurWorldInData.org/energy • CC BY

Figure 1.
Global electricity production by source. Based on the primary source, 95% of global production utilizes mostly synchronous generators [4].
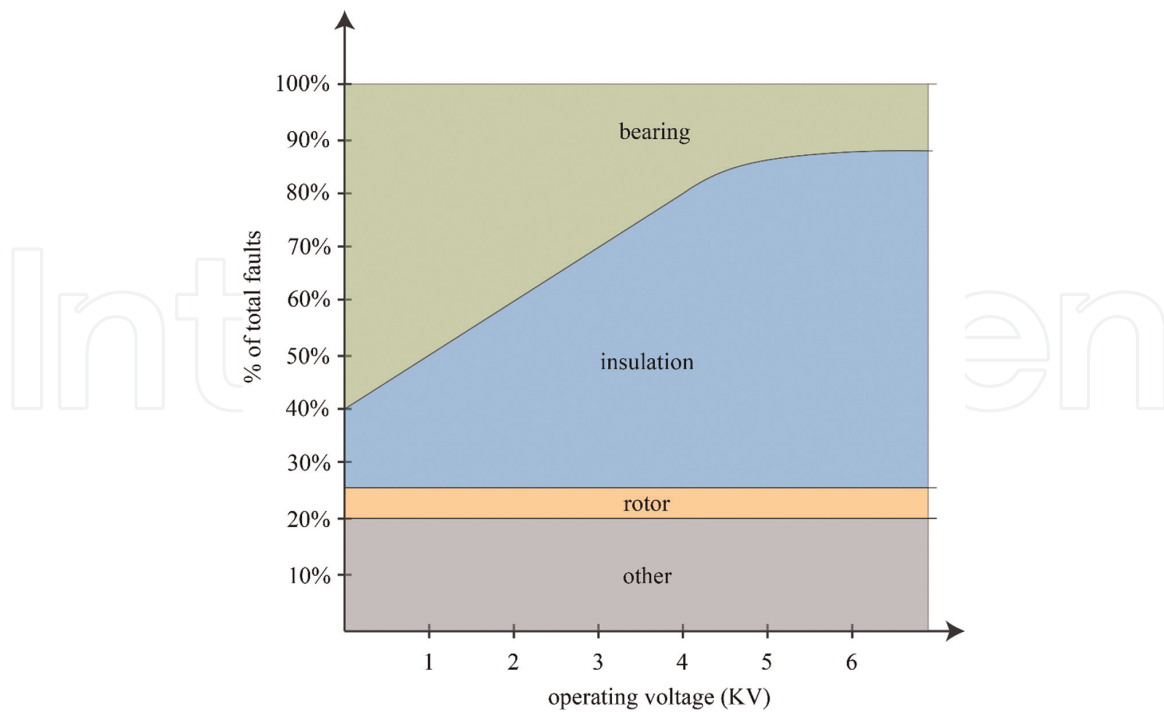
Figure 2.
Fault distribution in IMs according to their operating voltage level. As the voltage level increases, insulation faults become more prominent. The same holds for SGs.

- the stator core, which consists of thin sheets of magnetic steel and prevents the flowing current in adjacent conductive material;

- the electrical insulation, which is passive but necessary for the EMs and consists of:

  - Sub conductor insulation;

  - Turn/strands insulation;

  - Ground wall (main) insulation.

Constant or transient stresses can affect the insulation deterioration of an EM. These stresses are thermal, electrical, mechanical, and ambient, commonly known as TEAM stresses.

Thermal stress [9] is determined by the losses within the conductors and plays the most significant role in the degradation of the insulation system. The operating temperature of the windings, which is a result of I^2R, eddy current, load losses, and heating due to core losses, is the primary source of thermal stress. Increased temperatures lead to more frequent chemical reactions and insulation lifetime is described by using the Arrhenius law:

$$L = Ae^{B/T} \qquad (1)$$

where L is lifetime, T is the absolute temperature, A is the frequency of molecular encounter and $B = -E/R$, where E is the activation energy, which is constant for a given reaction and R is the universal gas constant. Thermal stress results in high-temperature differentials, overload, and hot spots.

Electrical stress [10] is related to the thickness of the electrical insulation of EMs.. It can lead to partial discharges (PD), which are small electric sparks that occur within air pockets of the insulation or on the surface of coils. In this case, the lifetime of the insulation is described by:

$$L = cE^{-n} \qquad (2)$$

where L is lifetime, c is a constant, E is the stress level (kV/mm) and n is the power law constant. Surges, overvoltages, and partial discharges are indicative consequences of electrical stress.

Ambient stress [8] is caused by miscellaneous factors that can lead to stress and typically amplify main stress categories via their mechanisms. Ambient stress sources include moisture on the windings, presence of oil or dust, high humidity, broken particles within the EMs, and aggressive chemicals. The results can be aggressive and reactive chemical reactions degrading the machine parts, as well as contamination.

Mechanical stress [9] is caused by force acting on the parts, stemming from mechanical vibrations or electromagnetic forces, such as end-winding vibrations. The lifetime of the insulation is described using:

$$L = D\sigma^{-m} \qquad (3)$$

where L is the lifetime, D, m are constants related to the insulation material and σ is the applied stress in N/mm^2. There is vibration and oscillation in slot sections and in end winding.

Deterioration of the insulation system is typically caused by two or more stress factors that are responsible for that specific result. Multiple stresses both accelerate and evolve the failure, leading to more significant problems.

Various diagnostic tests can be used for the evaluation of the condition of an EM. These tests are undertaken after the EM has been manufactured, installed on-site, during periodic maintenance checks, or when fault indications occur. Standard offline diagnostics follow a common sequence, which utilizes the following categories of experiments.

2.1 Visual inspection (VI)

VI is the standard and usually the first offline diagnostic procedure, because it gives information for most possible faults both on the stator and rotor, indicating the necessity of further testing. VI utilizes a borescope, which is an optical device consisting of a rigid or flexible tube with an eyepiece or display on one end and an objective camera on the other, linked together by an optical or electrical system in between. A typical borescope is shown in **Figure 3** [11]:
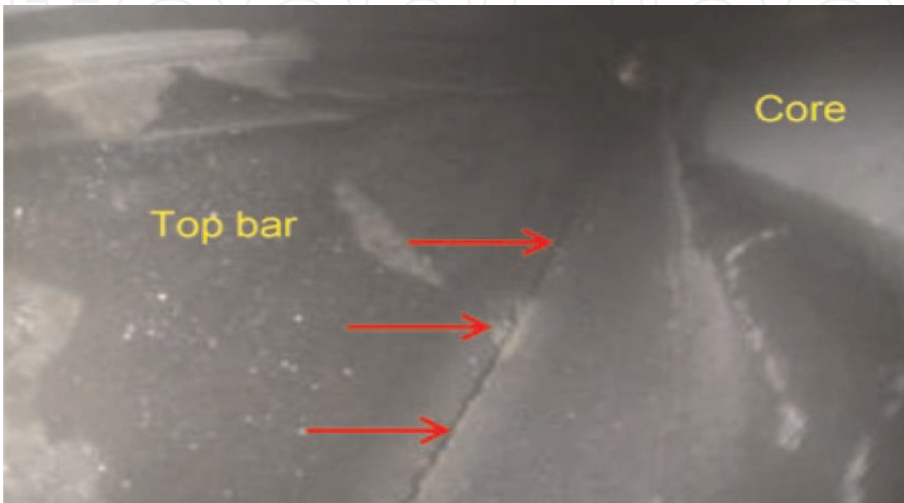
- A flexible or a rigid tube

- An eyepiece

- A light source

- Optical lenses

Figure 3.
Commercial typical Borescope [11].
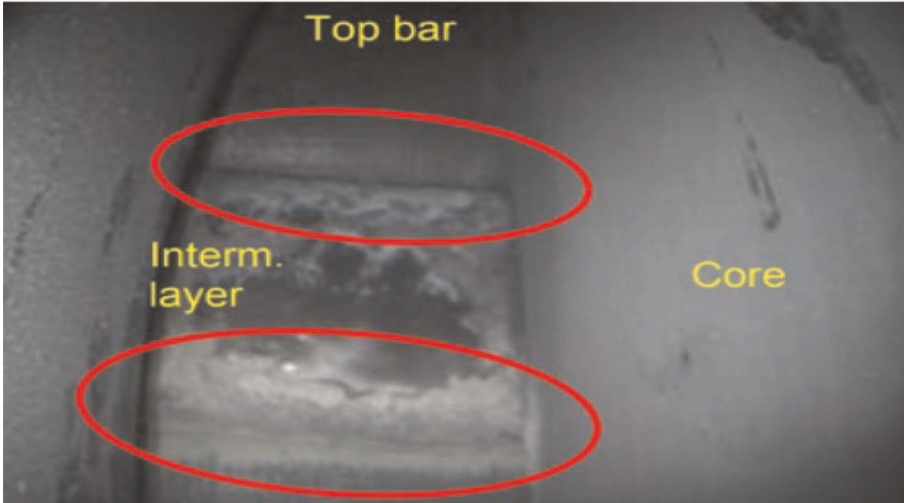
Figure 4.
Mechanical erosion.

Figure 5.
Electrical erosion.

Figure 6.
PD effects.

Figure 7.
Broken Bar.

Common goals are the detection of partial discharge spread on bars and resulting mechanical erosion. It is performed through core ventilation channels from inside the cooler channels. VI is also suitable for detecting humidity, thermal and mechanical deterioration, cracks, ground wall insulation, insulation degradation, and turn-to-turn failures.

The following figures show several problems detected by using a borescope inside the stator of a real SG. Specifically, **Figure 4** shows significant mechanical erosion at the top bar of the SG, **Figure 5** shows the effects of electrical stresses, **Figure 6** shows electrical erosion at the bar due to PDs, and **Figure 7** shows cracks at the bar of the SG. These images may suggest following electrical testing.

2.2 Insulation resistance (IR) and polarization index (PI) test (Std. IEEE 43: 2013)

Indicative quantities are depicted as the IR value denoted by the time of calculation (in minutes) after the application of a voltage source on the ends of the insulation

component. Typical values include R_1 = 100 Ω for most AC windings built after 1970 and R_1 = 5 for most EMs with random-wound stator coils and form-wound coils rated below 1 kV and DC armatures [9, 12].

PI tests follow the same standard with IR. Minimum PI = 1.5 for thermal class A (105) and minimum PI = 2 for class B (130) and above.

The two previous tests are always performed together -due to their dependence, explained below- in what is called the IR/PI test, which is suitable for checking machine windings for insulation deterioration. These tests are the most common for detecting potential problems in windings caused by contamination and pollution. Moreover, humidity, moisture, and cracks can be detected with these tests. PI can also determine if there is thermal deterioration.

The IR/PI test is done right at the machine terminals, one phase at a time, with cables and transformers disconnected. A high-voltage DC supply and a sensitive ammeter are required. The IR test measures the resistance of the electrical insulation between the copper conductor and the core of the stator or rotor. The PI is defined as the ratio between the IR measured after the voltage has been applied for 10 minutes and IR measured after 1 minute. Both IR and PI values decrease as an EM operates over the years because of the inevitably higher pollution penetration into the EM's windings. This means that IR is both initially reduced as time progresses and is also less resilient to the constant thermal stress applied by the current. IR is based on Ohm's law:

$$R = \frac{V}{I} \tag{4}$$

where V is the applied voltage and I is the sum of capacitive current, conduction current, leakage surface current, and absorption current. Lower IR is an indication that a problem exists within the insulation system since resistance has been lowered by contaminants or defects.

PI is a variation of the IR test. PI is the ratio of the IR measured after the voltage has been applied for 10 minutes (R_{10}) to the IR measured after just 1 minute (R_1):

$$PI = \frac{R_{10}}{R_1} \tag{5}$$

2.3 Measurement of power factor (PF) and dissipation factor (DF) (Std. IEEE 286: 2000)

According to the aforementioned standard and common knowledge, PF can be between 0 and 1. The same goes for DF, according to IEC 60034–27.3. Dissipation and power factors provide an indication of the dielectric losses within an insulation system. These measurements are conducted to identify if there are variations in C, DF, and PF over time, which indicate partial discharges or insulation degradation.

The DF is measured with a balanced bridge-type instrument, where a resistive-capacitive network is varied to give the same voltage and phase angle (tan delta) as measured across the stator winding. The DF is calculated from the R and C elements in the bridge that give the null voltage. This test is used for detecting humidity, moisture, PD, dielectric losses, and insulation degradation [13] (**Figure 8**).

The PF is measured by accurately measuring the Voltage applied between the copper and the core of a winding and detecting the resulting current. Also, it is necessary to measure the power of the winding with a wattmeter. So, the PF is:
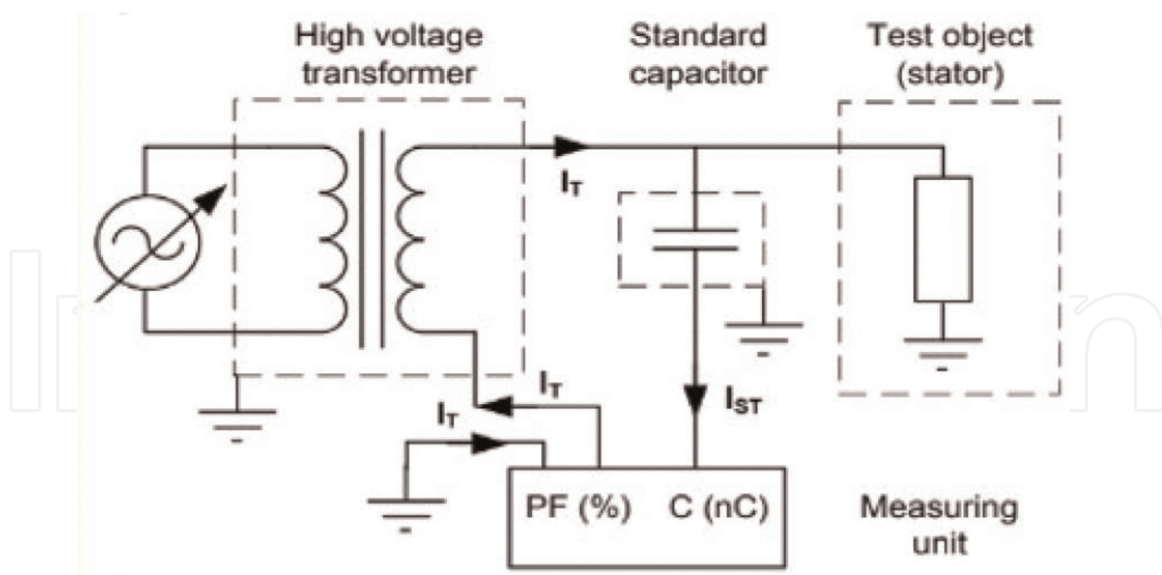
Figure 8.
PF/DF measurement scheme.

$$PF = \frac{W}{VI} \tag{6}$$

Comparing the two methods, the PF test is less accurate but less expensive as there is no need for a bridge-type instrument. The measurement of the DF can give information about PD activity, and contamination, while PF cannot. The DF can be converted to PF using:

$$PF = \frac{DF}{\left(1 + DF^2\right)^{0.5}} \tag{7}$$

The measurements of PF can be used for detecting possible problems and faults, such as humidity, moisture, overheating, dielectric losses, and insulation degradation. A wattmeter is used in order to measure the power to the winding and a voltmeter for measuring the applied voltage between the copper and the core of the winding and detecting the current.

2.4 Impedance test (Std. IEEE 112: 2004)

Humidity, moisture, thermal and mechanical deterioration, insulation degradation, and turn-to-turn failures are the faults, which can be detected by the measurements of the impedance. An AC source is used and different values of current are used in order to investigate the corresponding voltage values. Then, the impedance is calculated by using Ohm's law [14]:

$$Z = \frac{\Delta V}{I} \tag{8}$$

Figure 9 shows the results of an Impedance Test on a real Syncronous Generator. The four different lines are explained above the diagram. SR1 indicates the impedance of slip ring 1, while SR2 indicates the impedance of slip ring 2.
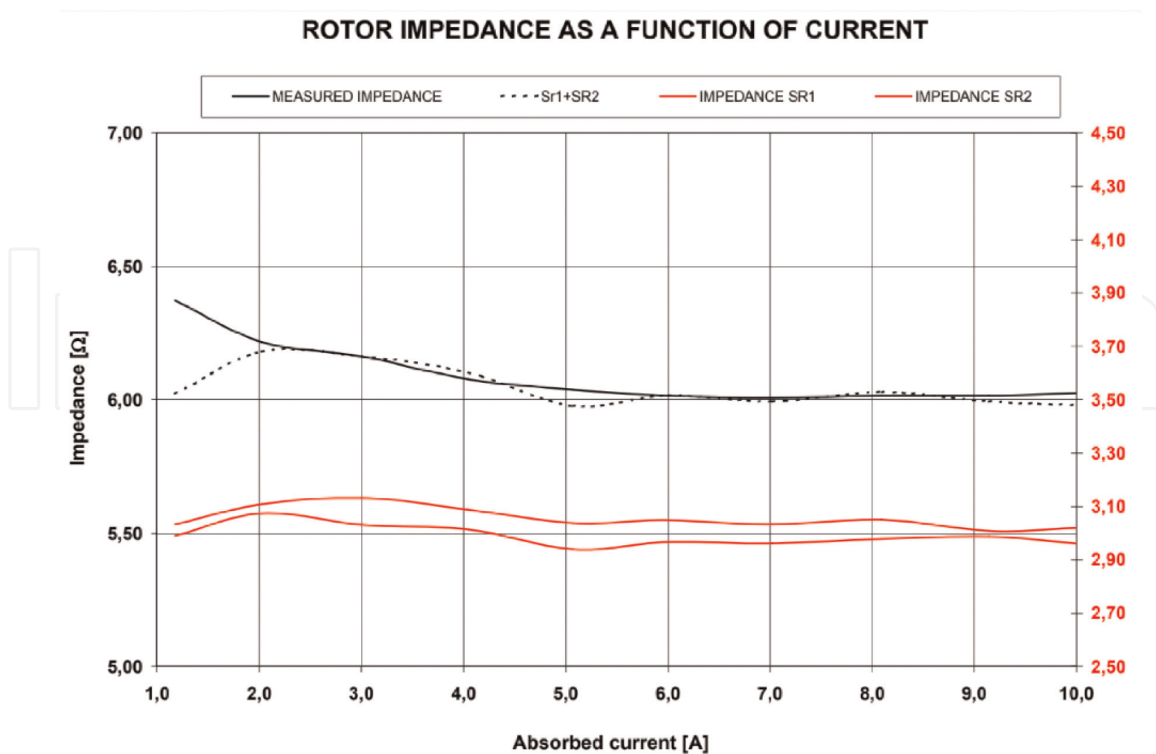
ROTOR IMPEDANCE AS A FUNCTION OF CURRENT

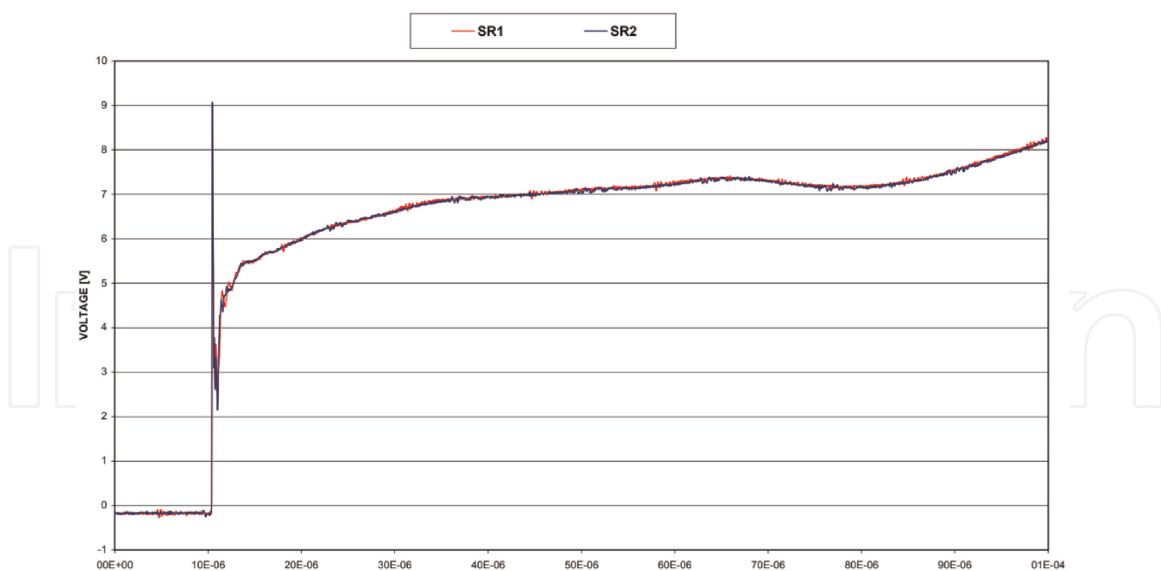**Figure 9.**
Impedance test results.

2.5 Recurrent surge oscilloscope test (RSO) (Std. IEEE 56: 2016)

A low-voltage and high-frequency surge wave is injected at each slip ring. The test is based on an oscillograph inspection of the voltage traveling wave between the slip rings along the symmetrically constructed winding field rotor. A low voltage high frequency surge wave is injected at each one of the slip rings. The two signals are then compared to determine if the same waveform is observed at each slip ring. If the waves are identical then no short circuits are present. Variations in the pattern of the two waveforms would indicate shorts to be present. If the two signals have differences, interturn, ground, and turn-to-turn faults as well as insulation degradation are the possible faults. A power source, a reflectometer, or/and oscilloscope are appropriate for this test [15, 16].

A typical RSO diagram is shown in **Figure 10**. The two waveforms are identical and this means that the rotor is free of the aforementioned possible faults. It must be noted that SR1/OS indicates the voltage of slip ring 1 and SR2/IS indicates the voltage of slip ring 2.

2.6 Structure of a report

Every industrial EM has to be checked periodically by specialized technicians. After that, technical reports, which contain useful information and data for the inspection as well as the history of the inspected EM, are created and given to the industry in order to be informed about the condition of its EM and decide what actions must be taken. Moreover, reports like that can be used and analyzed by researchers in order to create prediction models for the condition of EMs. Specialized experience and real data from measurements on real EMs are elements that are
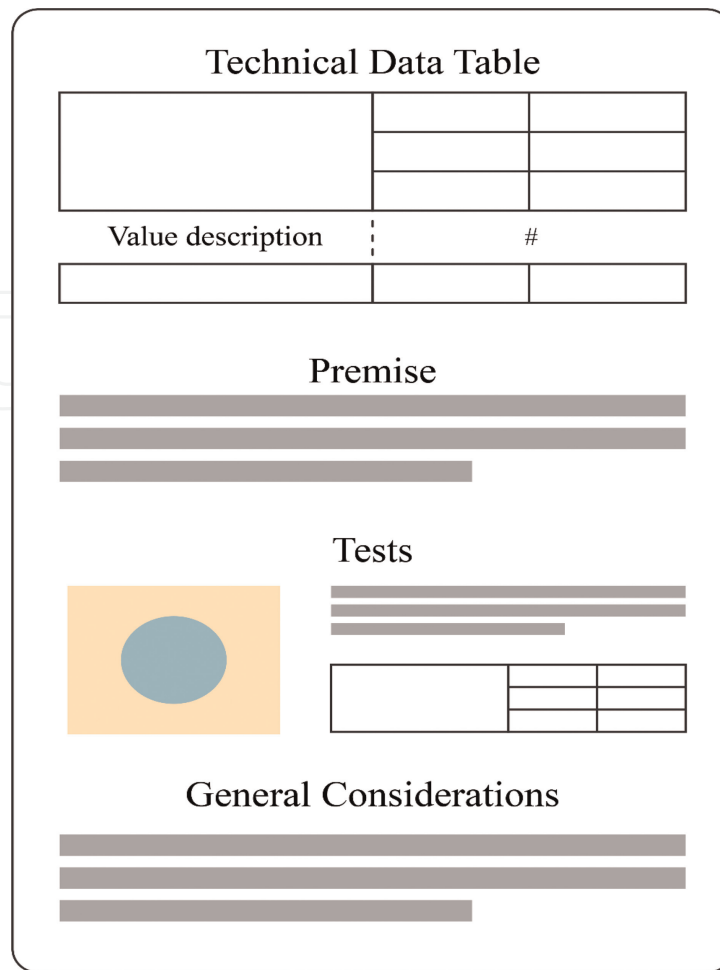
Figure 10.
RSO test results.

missing and they are very useful for creating prediction models with direct connection with the real situations of an industrial EM. Therefore, such reports are very significant for both industrial and research issues.

A typical structure of the reports used for the training of the proposed model of this manuscript is:

- Introduction - EM's Historical Issues - Milestones: The purposes of the Diagnostic Tests as well as information about the EM and significant dates are presented;

- Operation and Technical Data of the EM: rated power, voltage, current, frequency, power factor, dimensions, cooling type, number of poles, and other pertinent measurements;

- Selected Tests and Inspections: Different diagnostic methods are chosen each time according to the condition of the inspected EM, general considerations employed in selection;

- Results of the aforementioned tests and inspection: detailed information, data, diagrams, and pictures about the results of each diagnostic method as well as comparison with the previous years' diagnostics;

- Proposed maintenance actions according to the results of the measurements and information about the next date for diagnosis according to the results of the measurements.

Figures 11 and **12** highlight the different parts of a commonly found industrial SG report, created as a general template for the MRs studied during this research. The reports are considered as semi-structured, a term used in NLP to describe documents with structured information (tables, figures, lists etc.) interlaced with natural text. The structure within the report can be used to guide the AI processing it. For example, we know that the technical data table lists any number of parameters in any order,
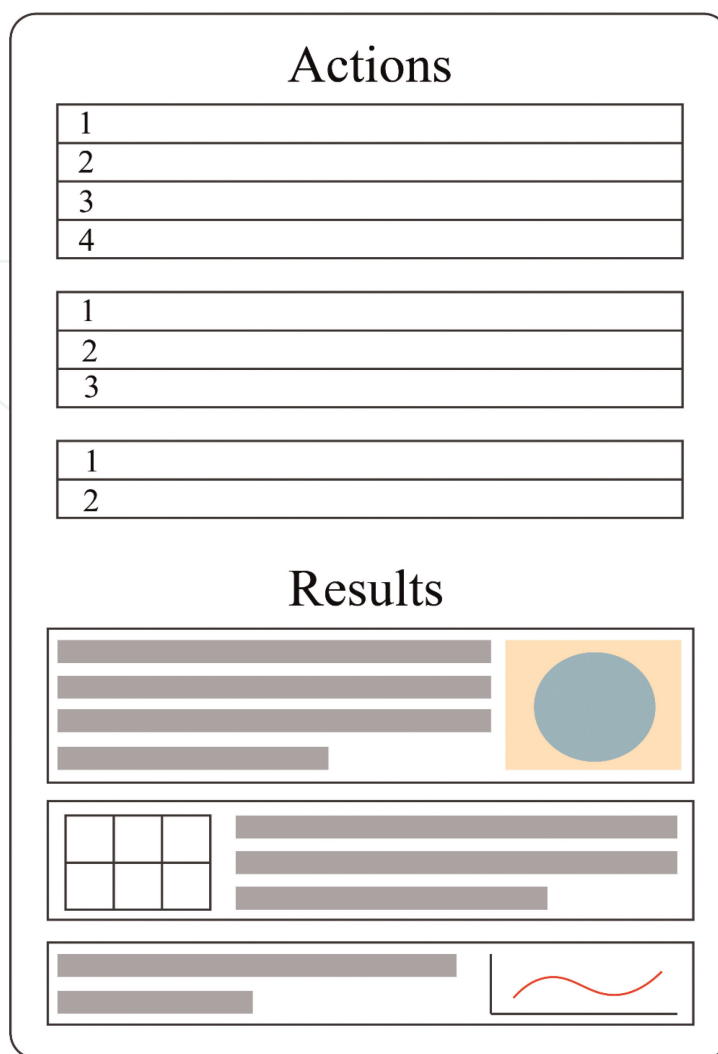
Figure 11.
First, more generic part of a typical SG maintenance report, just after the introduction.

therefore, the AI should expect a variable name partnered with a value. Furthermore, report language terminology is "specialized," meaning the total vocabulary is limited and populated with sector-specific terminology, further assisting by limiting the range of the employed interpretation.

As depicted in the pertinent figures, general information and considerations are commonly stored in just NT. The test section employs the most intuitive form of storage, such as imagery for VI. Actions and results are localized, meaning they are divided in different subsections for each component i.e., stator, rotor, and outside the frame. Actions are almost always depicted in lists (intuitively), while results follow the tests' paradigm of logical choice of medium. This localization can be employed in problem formulation to facilitate a correlation between cause and effect, while the list can provide the order of operations if logged properly.

3. Data mining reports using natural language processing

Processing information with logical processors or computer algorithms requires that its data is in numerical form. Furthermore, this form should also be in the appropriate context and facilitate necessary mathematical transformations. While this procedure is intuitive for numerical data such as measurements due to their

Figure 12.
Second, more specific part of a typical SG maintenance report.

underlying physical meaning and explored mechanisms, text processing requires a sophisticated approach.

The scientific field of NLP concerns itself with constantly improving existent and exploring new ways of information extraction from natural text. In the context of EM CM, it is important to understand the ways state-of-the-art procedures interpret text bodies, as the interpretation is closely tied to fault causality and can improve not only correlation understanding between different faults -since faults always are cascading- but also aid training of Industry 4.0's prodigal child, AI pattern recognition. This chapter introduces general principles of NLP state-of-the-art to aid the investigation of this promising avenue by discussing them in the context of EM CM. The successful application will enable tapping into a previously underutilized information source while paving the way for further storage, necessary under new industrial paradigms.

3.1 General principles

Language interpretation by machines is being explored for several decades, with the most successful principle rising to be word embedding. Understanding the embedding requires learning some preliminary machine learning (ML) concepts. The analysis presented hereafter includes CM examples.

Features are the input variables, here being sentences or words from the text body. Labels are the things to predict. For example, the sentence "the machine was found to have an increased PI value" could have the label "thermal degradation". A straightforward structure is having the words as features. Each selection as an input to the interpreter (could be a sentence or several, or a single word) is referred to as an Example. Examples can be labeled or unlabeled. It is important to understand that, as of now, all examples present in any report corpus are unlabeled. ML algorithms require labeled sentences to be trained. Therefore, an important issue to be solved is the production of labeled examples.

The training result, that is the prediction of a label given the features, is called the model. The model consists of the structure and weights of the classifier. A model can solve either a classification or a regression problem. Regression is continuous value prediction, while classification refers to discrete predictions. The most straightforward way to structure a fault prediction model would apparently be classification, such as the example provided above. However, regressive models could also prove to be useful and should always be considered.

An important metric in evaluating a training result -the weights and bias of the model- is Loss. Loss is a number indicating how bad the prediction of a single example performed. A perfect model would have zero loss, increasing with each failed prediction, and how far from the correct answer it is. Minimization of loss is therefore the function that the training is based upon and should be carefully formulated.

Low loss is not a complete indication that our model performed adequately. A model could have low loss but perform poorly when introduced with new unseen examples. This concept is described as overfitting, which is when our model has a poor generalization capability. Overfitting occurs when the model is more complex than necessary, proven time and time again and manifested in Ockham's razor. But how can we create a model from scratch based on a text corpus and provide it with unseen data for validation? The answer is separating the dataset into the training, validation, and test subsets. Proper separation is of paramount importance to the training, with existing ML paradigms found in the literature. Training on the same dataset, however, still exhausts and overfits the model. It generally is a great idea to keep refreshing the dataset with new reports while continually adjusting the established model.

As previously discussed, in order to enable the model to multiply features with their weights, said features should be numerical values. This process is called Feature Engineering and is a critical step in ML. In the case of text, the most straightforward way to map words is the so-called One-Hot Encoding. This encoding utilizes a vector of dimensionality equal to the vocabulary (total number of different words), where each word corresponds to a specific dimension of the vector, in which place its value is 1, with the rest being 0. Dimensionality can be reduced by aggregating infrequent words into an out-of-vocabulary (OOV) group. While one-hot encoding is the basic understanding example of word encoding, advanced techniques are preferred.

Now on to the gist of NLP. The next step up from proper encoding comes in the form of an Embedding. Shortly, an embedding is a vector in low-dimensional space coded such that the feature it represents is nearby similar ones. Take for example the words "king" and "queen." These words should be relatively close in the embedding space since they refer to a similar quality. Furthermore, an intuitive embedding would also support a mathematical operation, for example, "+", so that the calculation would be: "king" – "male" + "female" = "queen". A possible problem with utilizing the one-hot encoding in this example is that the dimensionality of a sentence would be

arbitrarily large and sentence vectors would be very sparse; a lot of zeroes with very few ones in between. Therefore, proper embeddings are required. Thankfully, NLP research has provided tools for the task.

3.2 BERT: An interesting approach

Concerned with EM CM and NT information extraction specifically, a logical approach would be to use developed and established tools to achieve our goal, rather than making one from scratch. The choice of tools is largely based on their performance in established standards and engineering intuition -whether it suits the pertinent problem-. To that end, Bidirectional Encoder Representations from Transformers (BERT) are the tool of choice to investigate possible correlations in our task [17].

NLP state-of-the-art employs pre-trained language models. That is, these models employ a trained embedding model, with additional semantic ML analysis, quickly defined in the previous subchapter. It is important to remember that with ML tasks, time efficiency is of paramount importance. Due to the complexity and nuance of text mining, model training requires immense hardware capability and a time sink. Thus, pre-trained models have been extensively researched to be employed in final tasks.

Two general separations occur in possible applications: level and approach. The two levels are sentence and token levels. Sentence level includes the entirety of two or more sentences as input and attempts to predict their relationship with a holistic analysis. Token-level tasks provide a more precise output at the word level and are suited for question answering and named entity recognition. CM tasks are approached as sentence-level; we attempt to predict faults via sentence relationships.

The approach could be either feature-based or fine-tuning. A feature-based approach would be largely dependent on the task at hand; additional model architecture is designed specifically for the problem. Fine-tuning is a novel technique that utilizes the pre-training and keeps the same architecture with parameter training on the task at hand. Both approaches would be suited to CM and are up for debate. BERT utilizes fine-tuning.

One novelty provided by the BERT approach is its bi-directional representation. Previous state-of-the-art models would approach a sentence unidirectionally or at best aggregate the left-to-right and right-to-left representations. Consider the human interpretation of a sentence; we both speak and process information serially, or left-to-right but intuitively also consider the entirety of the sentence both in formulation and processing. Therefore, a bidirectional representation would theoretically paint a more complete image.

One more parameter to consider is whether the training is done supervised or unsupervised. BERT employs an unsupervised training approach due to the nature of semantics extraction from NT, that is to infer new possible correlations from the information contained within rather than the already established knowledge, which would render the research point moot. Pre-training approach also enables transfer learning, which transfers knowledge from larger datasets and/or supervised tasks. The natural language contains base semantics that apply to multiple different problems in different iterations. Consider the following example: a learned differentiation between "positive" or "negative" would also apply contextually to "healthy" or "faulty". This knowledge is contained within a broader dataset and can be fine-tuned on the task at hand.

As a tool, BERT focuses on the fine-tuning approach, making it essentially plug-and-play for our operation. Architecture appears to fit the CM context, while its

performance based on important metrics is state-of-the-art. Therefore, it is promising for an initial attempt.

3.3 A deeper dive

Transformer [18] utilization presents a unified architecture across different tasks; both at the pre-training and fine-tuning steps, as well as the capability to be employed on multiple input forms such as images and video, allowing for a possible expansion of MR information extraction apart from its text corpus. The architecture allows for training on unlabeled data, which is paramount for the task at hand. While supervised, labeled fine-tuning is required and results in significantly improved performance, processing the entirety of the report corpus would prove immensely difficult. Deep consideration of experts is necessary for this endeavor in labeling examples to initialize training; processing unlabeled data is equally important in time efficiency.

The relationship between two sentences is not directly captured by language modeling, which is where the second stage of our training comes in, further realized at the task level by the fine-tuning mechanism. BERT's attention basis additionally allows for correlation between distant sentences, helping the endeavor. However, this transformer-based architecture may not be easily able to represent the entirety of the CM problem and purpose; while additional recurrent neural network (RNN) with long-short term memory (LSTM) neurons -extensively used in NLP- reduces typical BERT performance, additions may be required to fully encase the problem and have to be researched. Initial case studies with only BERT are being performed, pending judgment by experts. Overall, the underlying novel mechanism's benefits over older approaches can be summarized as:

- Faster and less complex representation;

- More interpretable models;

- Diversity in tasks;

- Behavior related to semantic and syntactic structure of sentences;

- Application to audio, video, and images.

3.3.1 Tokenization

Another interesting BERT component is the employed WordPiece embedding [19]. Due to its namesake, this embedding utilizes a limited vocabulary of sub-word units, further reducing dimensionality i.e., aggregating different forms of the same word. This procedure naturally handles the processing of rare words and is especially useful in semi-structured language corpora, such as our specialized engineering language.

The embedding includes a balance between character and word delimiters, enabling handling of newly seen (OOV) words with a completely data-driven approach and guaranteeing the generation of a deterministic segmentation for any possible sequence of characters. Additionally, an included length-normalization procedure and coverage penalty encourages complete coverage of source sentences in output. This essentially infinite vocabulary allows for open-ended optimization. However, this approach does not allow for observation of training errors with the

built-in fitness function and requires a task-specific procedure to evaluate reward. Therefore, the proper definition of this function is important.

3.4 General approach

Having selected the tools at our disposal, briefly presented above, NLP research paired with ML paradigms and put in the context of the specific problem produces a general algorithm for this endeavor. However, before delving further in the discussion, a couple of disclaimers should be made. Firstly, the above tools have been evaluated and show promise, but researchers should remember that no approach is perfect. Interested parties are encouraged to test a diverse selection of NLP procedures in tackling the problem. Secondly, while generalized approaches do not provide specific solutions, the purpose of this chapter is the introduction of the proposed idea. Furthermore, the problem is open-ended and needs further elaboration in breaking down its complexity while setting standards. Thus, the proposed general representation is deemed appropriate. Finally, the explanation of BERT is kept simple in this chapter. Interested readers may refer to the open-source code and cited work for an in-depth, complete analysis. **Figure 13** presents an overview.
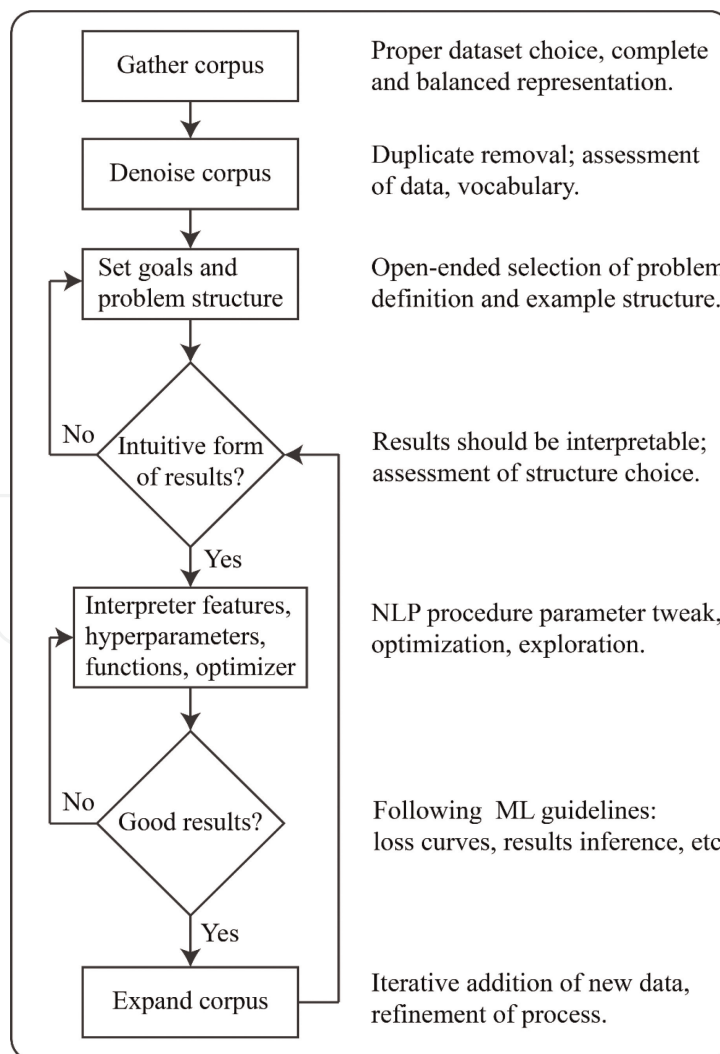
Figure 13.
General iterative process of machine learning problem establishment.

3.4.1 Discussion

While the proper selection of AI network architecture and hyperparameters is an important and difficult task on its own, the real challenge in this endeavor is structuring the problem. Were it only raw text processing, one would only look to combining NLP paradigms with EM CM expertise to formulate feature selection, perform labeling and optimize the procedure. However, one grand challenge in this endeavor is the robust processing of the reports due to their specific-but-varying structure, which is a double-edged sword; on one hand, this structure can be employed to better infer correlations and aid the interpreter with limitations; on the other hand, it should carefully be considered since the improper setting of the problem structure renders NLP impossible.

ML consensus agrees on an iterative approach. At first, a human manageable corpus is to be selected and denoised. After being deemed proper and balanced in its representation, the problem is set up with feature selection, labeling, and architecture choice, followed by optimization. When results are satisfying and intuitive, outside expertise should assess and offer an outside perspective; then, the process is expanded with new data, a classic procedure in ML tasks.

A human overview is the key to this research. In translation tasks, it has been reported to improve performance by up to 60% (according to pertinent test scores) [19]. It is important to understand that a complete AI CM is far off; this research aims to provide a tool for engineers to automatically and efficiently extract information from untapped datasets. EM MRs represent the intuitive knowledge of experts, which it attempts to quantify.

3.4.2 Dataset standards and default parameters

Finally, two initial steps in ML approaches are reaching milestones pertinent to dataset size, and finding initial parameters. This paragraph attempts to provide a few important values.

- BERT Model: largest allowed by hardware and time constraints;

- Mini-batch size: 16–48;

- Epochs: 2–3;

- Learning Rate: 2e-5 to 5e-5

- Sentences: >4.5 M;

- Vocabulary: 8-40 K;

- Optimizer: default ADAM [20].

Due to being open-source, there are numerous guidelines and examples of BERT applications. The above values have proven sufficient in the considered EM CM context via testing with real MRs, and should serve as a good starting point.

4. Conclusion

Data-driven approaches are substantially beneficial for new industrial and research paradigms such as Industry 4.0 and the emergence of Big Data. New methodologies, such as the Digital Twin [21], can greatly benefit from a large and structured database, especially in the context of EM CM, since faults are deeply correlated and their mechanisms are still partially obscured. This work presents a novel approach for structured data extraction from an untapped source of information, namely the knowledge stored in EM MRs in the form of NT. NLP is increasingly gaining traction due to the aforementioned circumstances and has not yet been employed in this field, to the best of the author's knowledge. We surmise that a breakthrough in this endeavor can greatly benefit the industry and attempt to initialize it with research in this department. This chapter serves as the overview of this attempt, providing extended knowledge acquired during related research.

Author details

Karlis Athanasios[1]*, Falekas Georgios[1], Verginadis Dimosthenis[1] and
Jose A. Antonino-Daviu[2]

1 Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece

2 Department of Electrical Engineering, Universitat Politècnica de València, Valencia, Spain

*Address all correspondence to: akarlis@ee.duth.gr

IntechOpen

References

[1] Verginadis D, Antonino-Daviu J, Karlis A, Danikas MG. Diagnosis of stator faults in synchronous generators: Short review and practical case. In: Proceedings – 2020 International Conference on Electrical Machines. New York City, United States: ICEM; 2020. pp. 1328-1334. DOI: 10.1109/ICEM49940.2020.9270936

[2] Nakata T. Text-mining on incident reports to find knowledge on industrial safety. Annual Reliability and Maintainability Symposium (RAMS). 2017. pp. 1-5. DOI: 10.1109/RAM.2017.7889795

[3] Brundage MP, Sexton T, Hodkiewicz M, Dima A, Lukens S. Technical language processing: Unlocking maintenance knowledge. Manufacturing Letters. 2021;**27**:42-46

[4] Ritchie H, Roser M, Rosado P. Energy. Oxford, England. 2020. Published online: OurWorldInData.org. Retrieved from: https://ourworldindata.org/energy [Online Resource]

[5] Afrandideh S, Milasi ME, Haghjoo F, Cruz SMA. Turn to turn fault detection, discrimination, and faulty region identification in the stator and rotor windings of synchronous machines based on the rotational magnetic field distortion. IEEE Transactions on Energy Conversion. 2019;**35**(1):292-301

[6] Kammermann J, Bolvashenkov I, Schwimmbeck S, Herzog HG. Reliability of induction machines: Statistics, tendencies, and perspectives. In: The 31st IEEE International Symposium on Industrial Electronics (ISIE). New York City, United States: ISIE; 2017. pp. 1843-1847

[7] Swana EF, Doorsamy W. Investigation of combined electrical

modalities for fault diagnosis on a wound-rotor induction generator. IEEE Access. 2019;**7**:32333-32342

[8] Stone GC, Boulter EA, Culbert I, Dhirani H. Electrical Insulation for Rotating Machines: Design, Evaluation, Aging, Testing, and Repair. Vol. 21. Hoboken, New Jersey: John Wiley & Sons; 2004

[9] Cimino A, Jenau F, Staubach C. Causes of cyclic mechanical aging and its detection in stator winding insulation systems. IEEE Electrical Insulation Magazine. 2019;**35**(3):32-40

[10] Saxén C, Gamstedt EK, Afshar R, Paulsson G, Sahlén F. A micro-computed tomography investigation of the breakdown paths in mica/epoxy machine insulation. IEEE Transactions on Dielectrics and Electrical Insulation. 2018;**25**(4):1553-1559

[11] Svensen M, Hardwick D, Powrie H. Deep neural networks analysis of borescope images. In: Proceedings of the European Conference of the PHM Society. Utrecht, The Netherlands; 2018. pp. 3-6

[12] Zoeller C, Vogelsberger MA, Fasching R, Grubelnik W, Wolbank TM. Evaluation and current-response-based identification of insulation degradation for high utilized electrical machines in railway application. IEEE Transactions on Industry Applications. 2017;**53**(3): 2679-2689

[13] Bin Lee S, Younsi K, Kliman GB. An online technique for monitoring the insulation condition of AC machine stator windings. IEEE Transactions on Energy Conversion. 2005;**20**(4):737-745

[14] Verginadis D, Antonino-Daviu JA, Karlis A, Danikas MG. Determination of

the insulation condition in synchronous generators: Industrial methods and a case study. IEEE Industry Applications Magazine. 2022;**28**(2):67-77

[15] Kerszenbaum I, Maughan C. Utilization of repetitive surge oscillograph (RSO) in the detection of rotor shorted-turns in large turbine-driven generators. In: 2011 Electrical Insulation Conference (EIC). Vol. 2011. New York City, United States: EIC; pp. 398-401

[16] Imoru O, Mokate L, Jimoh AA, Hamam Y. Diagnosis of rotor inter-turn fault of electrical machine at speed using stray flux test method. In: AFRICON 2015. New York City, United States. 2015

[17] Devlin J, Chang MW, Lee K, Toutanova K. "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Proceedings of the 2019 Conference. Seattle, Washington, USA. 2014. pp. 4171-4186

[18] Vaswani A et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;**2017**(Nips): 5999-6009

[19] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016. pp. 1-23. ArXiv: abs/1609.08144

[20] Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations. San Diego, California, USA: ICLR; 2015. pp. 1-15

[21] Falekas G, Karlis A. Digital twin in electrical machine control and predictive maintenance: state-of-the-art and future prospects. Energies. 2021;**14**(18)