Received: 5 January 2024 | Revised: 4 September 2024 | Accepted: 11 October 2024

IET Renewable Power Generation

DOI: 10.1049/rpg2.13151

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Impact of using text classifiers for standardising maintenance data of wind turbines on reliability calculations

Julia Walgern[1,2] | Katharina Beckh[3] | Neele Hannes[1,4] | Martin Horn[1,5] | Marc-Alexander Lutz[6] | Katharina Fischer[1] | Athanasios Kolios[2,7]

[1]Department of Technical Reliability, Fraunhofer Institute for Wind Energy Systems IWES, Hanover, Germany

[2]Department of Naval Architecture, Ocean & Marine Engineering, University of Strathclyde, Glasgow, UK

[3]Department of Knowledge Discovery, Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Schloss Birlinghoven, Sankt Augustin, Germany

[4]Department of Mechanical Engineering, RWTH Aachen University, Aachen, Germany

[5]Department of Wind Energy Systems, Leibniz University Hannover, Hannover, Germany

[6]Department of Energy Informatics, Fraunhofer Institute for Energy Economics and Energy System Technology IEE, Kassel, Germany

[7]Department of Wind & Energy Systems, Risø Campus, Technical University of Denmark, Roskilde, Denmark

Correspondence
Julia Walgern, Department of Technical Reliability, Fraunhofer Institute for Wind Energy Systems IWES, Postkamp 12, 30159 Hanover, Germany.
Email: julia.walgern@iwes.fraunhofer.de

Funding information
Bundesministerium für Wirtschaft und Klimaschutz, Grant/Award Number: 03EE2016A; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/S023801/1

Abstract

This study delves into the challenge of efficiently digitalising wind turbine maintenance data, traditionally hindered by non-standardised formats necessitating manual, expert intervention. Highlighting the discrepancies in past reliability studies based on different key performance indicators (KPIs), the paper underscores the importance of consistent standards, like RDS-PP, for maintenance data categorisation. Leveraging on established digitalisation workflows, we investigate the efficacy of text classifiers in automating the categorisation process against conventional manual labelling. Results indicate that while classifiers exhibit high performance for specific datasets, their general applicability across diverse wind farms is limited at the present stage. Furthermore, differences in failure rate KPIs derived from manual versus classifier-processed data reveal uncertainties in both methods. The study suggests that enhanced clarity in maintenance reporting and refined designation systems can lead to more accurate KPIs.

1 | INTRODUCTION

Maintenance data of wind turbines is essential for analysing operation and maintenance (O&M) activities and for calculating related key performance indicators (KPIs). Corresponding data can facilitate the optimisation of O&M through logistic concept improvements or the implementation of preventive maintenance strategies, reducing the levelised cost of energy (LCoE). However, maintenance data in the wind industry is seldom available in a machine-readable and standardised format. As a result, this data is either overlooked or requires significant manual effort of a domain expert to process the information content.

Numerous reliability studies based on manually labelled maintenance reports have been published to understand which components undergo maintenance. A comprehensive review of

available reliability data is given in [1] and [2]. Some publications provide KPIs for all wind turbine subsystems (see e.g. [3–5]), while others focus on specific subsystem such as the power converter (see e.g. [6, 7]), the pitch system (e.g. [8, 9]) or the main bearing (e.g. [10]). Direct comparisons between such studies can be challenging due to the different categorisation systems and variations in provided KPI definitions. For instance, Carroll et al. reported reliability figures from about 350 offshore wind turbines, identifying the "pitch / hydraulics" subsystem as the most frequently failing one [4]. In contrast, the system performance, availability and reliability trend analysis (SPARTA) initiative noted the blade adjustment system as having the second highest monthly repair rate based on an analysis of 1045 offshore wind turbines located in UK waters [5]. Given that Carroll et al. use annual failure rates while SPARTA uses monthly repair rates, the studies base their findings on different KPI definitions. Moreover, the components included in the defined subsystems likely differ, complicating KPI comparisons. Anderson et al. highlight that even the definition of "failure" can vary in field-data based studies, impacting the KPI values [11]. One solution for uniformly defining subsystems and components of wind turbines is to employ standards and guidelines like the reference designation system RDS-PP for wind turbines [12] or RDS-PS [13]. Regrettably, these standards have not gained wide acceptance in the wind energy industry yet. Instead, many proprietary classification systems are in use which do not easily translate to the mentioned standards.

Most of the reliability studies mentioned above rely on data recorded before 2015. The extensive manual effort required to pre-process maintenance information often results in significant delays between documenting site visits and publishing data-based findings.

On the one hand, [14] developed a digitalisation workflow to standardise wind turbine maintenance information. This process involves optical character recognition, information extraction and text classification. After reviewing various classifier methods, they employed support vector machine (SVM) approach to train and test a text classifier to label service reports with RDS-PP components. As RDS-PP is organised in a hierarchical structure (cf. [14]), classification results for different levels have been presented and compared using F1 scores. While initial classification results displayed promising micro F1 scores, these were not explored further for productive application. Notably, training text classifiers for isolated RDS-PP levels is of limited practical relevance, as only the combined levels provide insights into the affected components and subsystems.

On the other hand, [15] analysed three different methods of labelling service reports, namely expert labelling, text classification and AI-assisted tagging in combination with a rule-based approach, to differentiate between predictive and corrective maintenance work orders. Afterwards, failure rates for the overall wind turbine system derived from the differently pre-processed data sets—making use of the simple categorisation of predictive and corrective activities—were compared. Results show that the AI-assisted tagging approach reduces data preparation time significantly, however, calculated KPIs are not reliable [15]. These findings are based on data of a single wind farm.

In contrast, this paper addresses the challenge of efficiently digitalising wind turbine maintenance data, traditionally hindered by non-standardised formats requiring manual expert intervention. This study investigates the efficacy of various text classifiers in automating the categorisation process of maintenance data against conventional manual labelling. The novelty of this research lies in the comprehensive evaluation of different text classifiers trained on diverse datasets and their impact on reliability KPI calculations. By comparing manual labelling with classifier-processed data, this paper reveals the uncertainties in both methods and suggests improvements in maintenance reporting and designation systems to achieve more accurate KPIs.

Within this paper, different text classifiers which classify the text descriptions of wind turbines' maintenance measures into RDS-PP categories, and thus different wind turbine components and subsystems, are analysed and implications for real-world application are assessed. Uncertainty resulting from using text classifiers based on different training data sets varying in size and homogeneity is evaluated. Additionally, it is analysed how reliability KPIs differ depending on the chosen pre-processing method comparing manual labelling against different text classifier results. Our findings offer recommendations for digitising wind turbine maintenance reports, making this study invaluable for researchers and practitioners processing text-based service reports to derive reliability figures or understand spare parts usage.

The paper is structured as follows: First, the state-of-the-art literature on text classification is discussed (Section 2). Second, an introduction of the used methods for classification but also for comparing classification results is given and the analysed data set is described (Section 3). Afterwards, several text classifiers based on different training data sets are evaluated and compared and industry perspectives are presented based on conducted interviews. Next to the classifier performance itself, the impact on reliability KPI calculation and corresponding uncertainties are analysed and barriers to the adoption of text classifiers in the wind energy sector discussed (Section 4). Last, main conclusions are summarised and an outlook to future work is given (Section 5).

2 | STATE OF THE ART LITERATURE ON TEXT CLASSIFICATION

Text classification, a fundamental task in natural language processing (NLP), involves assigning predefined categories to text data [16, 17]. Over the past few decades, this field has witnessed significant advancements, driven by the evolution of machine learning and deep learning techniques. This literature review highlights the key developments and state-of-the-art approaches in text classification.

The initial methods for text classification relied heavily on traditional machine learning algorithms such as naive Bayes, k-nearest neighbours (k-NN), and support vector machines

(SVM). These algorithms typically used bag-of-words or term frequency-inverse document frequency (TF-IDF) representations of text. For instance, [18] demonstrated the effectiveness of SVMs for text categorisation, showing superior performance compared to other methods at the time due to its ability to handle high-dimensional data, while recently, [19] have shown the impactful application of such methods for fault diagnosis for control of critical infrastructure.

To improve classification performance, extensive feature engineering was employed. Techniques like n-grams, part-of-speech tagging, and named entity recognition were used to extract meaningful features from text [20, 21] Ensemble methods, which combine multiple classifiers, were also explored [22]. The Random Forest algorithm, an ensemble of decision trees, proved effective for various text classification tasks due to its robustness and ability to handle large feature spaces [23].

The advent of deep learning marked a significant shift in text classification. Neural networks, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), demonstrated remarkable capabilities in capturing complex patterns in text data [24, 25]. [26] introduced a CNN model for sentence classification that outperformed traditional methods by leveraging pre-trained word embeddings and convolutional filters to capture local dependencies in text.

RNNs, especially long short-term memory (LSTM) networks [27] were effective in handling sequential data, making them suitable for text classification tasks. LSTMs addressed the vanishing gradient problem, enabling the capture of long-range dependencies [28]. This made them particularly useful for tasks like sentiment analysis and document classification, where context plays a crucial role.

The introduction of attention mechanisms further revolutionised text classification. Attention allows models to focus on relevant parts of the input sequence, enhancing their ability to capture context [29]. The Transformer model, introduced by [30] utilised self-attention mechanisms to process entire sequences simultaneously, rather than sequentially as in RNNs. This innovation led to significant improvements in both training efficiency and classification performance.

BERT (bidirectional encoder representations from transformers), developed by [31], built on the transformer architecture and introduced bidirectional context understanding. BERT achieved state-of-the-art results on various NLP benchmarks by pre-training on a large corpus and fine-tuning on specific tasks. Its ability to understand context in both directions of a text sequence made it particularly powerful for text classification [32].

Recent developments in large language models, such as GPT-3 [33] and GPT-4, have further advanced text classification. These models, with billions of parameters, are pre-trained on diverse datasets and can be fine-tuned for specific tasks with minimal additional training [34]. Their deep contextual understanding and ability to generate coherent text have set new benchmarks in text classification performance.

Transfer learning, where pre-trained models are fine-tuned on specific tasks, has become a dominant approach in text classification. Models like BERT, RoBERTa [35], and T5 [36] exemplify this trend. Fine-tuning these models on domain-specific data leads to substantial improvements in performance, as they leverage the rich knowledge gained during pre-training.

The integration of text with other modalities, such as images and audio, has opened new avenues for text classification [37]. Multimodal models that combine textual and visual data are being explored for tasks like social media analysis and sentiment classification [38]. Hybrid approaches that combine rule-based systems with machine learning are also gaining traction, offering a balance between interpretability and performance.

Text classification has found applications across various domains, including sentiment analysis, spam detection, topic labelling, and more [39]. The ongoing research focuses on improving model interpretability, handling low-resource languages, and reducing biases in text classification models [40]. Future directions include the development of more efficient models that require less computational power and the exploration of unsupervised and semi-supervised learning techniques to leverage unlabelled data.

In addition to text classifiers, the use of multimodal knowledge graph (KG) databases presents a promising approach for managing maintenance data of wind turbines. Knowledge graphs integrate heterogeneous data sources, including structured data (sensor readings, operational logs) and unstructured data (maintenance reports, technical manuals), enabling a holistic representation of information [41]. KGs can enhance data interoperability, facilitate advanced analytics, and improve decision-making processes by connecting related entities and capturing complex relationships [42]. For instance, in the healthcare domain, KGs have been used to integrate clinical data and literature, aiding in diagnosis and treatment planning [43]. Similarly, in wind turbine maintenance, a KG could unify data from various sources, providing a comprehensive view of turbine health and maintenance needs. A comparative analysis of text classifiers and KGs could reveal synergies, such as using classifiers to populate KGs, ultimately improving data utilisation and operational efficiency in wind energy systems.

This study advances the state of the art by integrating advanced text classification techniques with domain-specific fine-tuning to automate the categorisation of wind turbine maintenance data, a task traditionally requiring extensive manual effort. Unlike previous methods, our approach leverages the hierarchical structure of RDS-PP for precise component-level categorisation, enhancing the reliability and accuracy of maintenance logs. Additionally, by comparing classifier performance across diverse datasets and exploring the integration of large language models, we address scalability and adaptability challenges, providing a robust framework for standardising maintenance data and improving operational efficiency in the wind energy sector.

3 | METHODOLOGY AND DATA SETS

The methodology for this study involves a structured approach with distinct steps to ensure clarity and reproducibility (see

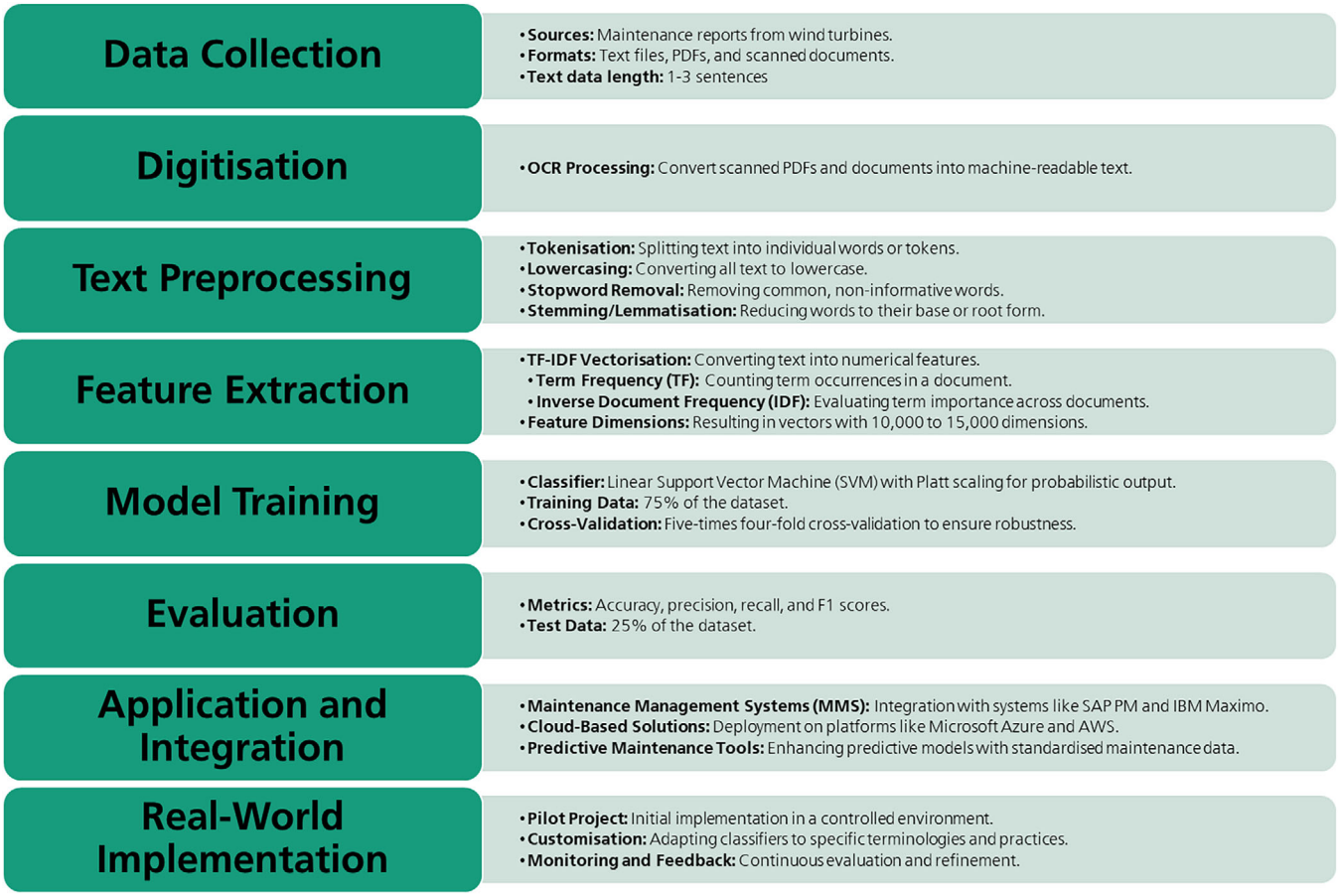

FIGURE 1 Workflow of the methodological framework.

Figure 1). The text classifiers are built using maintenance reports collected from wind turbines, which are initially available in various formats including text files and PDFs. The research employs natural language processing (NLP) to automate the categorisation of wind turbine maintenance logs, enhancing reliability assessments. Initially, maintenance reports are digitised using optical character recognition (OCR), followed by text preprocessing to standardise formats. Term frequency-inverse document frequency (TF-IDF) vectorisation converts text data into numerical features for model training. Support vector machine (SVM) classifiers, optimised with Platt scaling for probabilistic outputs, are trained on manually labelled data adhering to the reference designation system for power plants (RDS-PP). The classifiers undergo rigorous evaluation through five-times four-fold cross-validation, ensuring robust performance metrics, including precision, recall, and F1 scores. The classifiers then categorise maintenance activities, facilitating the calculation of failure rates and other key performance indicators (KPIs). Comparative analyses between classifier and manually derived KPIs reveal the models' efficacy. Industry feedback is incorporated to tailor classifier configurations, aiming for seamless integration into maintenance workflows, thus improving data-driven decision-making in wind turbine operations.

3.1 | Data sets

The data sets employed in this analysis originate from 15,000 maintenance reports spanning 342 wind turbines, both onshore and offshore. These reports represent approximately 800 operational turbine years and are sources from two distinct operators. As respective wind turbines are located in different countries and maintenance is performed by different companies, the reports are generally available in English but in rare examples German and French language is used as well. Additionally, different turbine types are included in the analysis which naturally leads to terminology variance. This is further exacerbated by the documentation from different companies. Despite these variations, the reports generally follow a standard structure that includes crucial sections and details necessary for accurate classification.

- Header information: The header typically contains metadata about the maintenance event, such as the date, time, wind turbine identifier, and the technicians' names. This section provides contextual information but is often not directly used for text classification.
- Summary of maintenance activity: This section briefly overviews the maintenance activity performed. It may include

a high-level description of the issue addressed and the actions taken. For example, a summary might state, "Replaced faulty pitch motor in turbine T123."
- Detailed description: The detailed description is the core of the maintenance report. It includes a step-by-step account of the maintenance process, components, tools and materials, and any observations or measurements taken. This section can vary in length from a few sentences to several paragraphs, depending on the complexity of the task and the organisation. Detailed descriptions are crucial for text classifiers as they contain the technical terms and context needed for accurate categorisation.
- Parts and materials used: This section lists all the parts and materials used during maintenance. It typically includes part numbers, quantities, and sometimes supplier information. This structured data can be cross-referenced with the textual descriptions to enhance classification accuracy.
- Recommendations and next steps: Maintenance reports often conclude with recommendations for future actions or follow-up maintenance tasks. This section may also note any potential issues that need monitoring. Although this information is valuable for ongoing maintenance planning, it is secondary for the initial classification of the report.
- Signatures and approvals: The report may include signatures from the technician and supervisory personnel, indicating that the maintenance activity has been reviewed and approved. This section is typically irrelevant for text classification but ensures the report's validity and compliance.

For every maintenance activity, the relevant components are categorised using the reference designation system RDS-PP for wind turbines [12]. For this, mainly the summary of maintenance activity and detailed description was utilised as other information categories were not available within all maintenance reports. This labelled dataset forms the foundation for training and testing text classifiers.

3.2 | Methodology

3.2.1 | Text classification and corresponding metrics

The chosen classification method is support vector machines (SVM). More specifically, a linear support vector classification [44] is used and a probabilistic output is achieved with Platt scaling [45]. Text data is transformed with TF-IDF (term frequency—inverse document frequency) vectorisation method as outlined in [44, 46, 47]). This straightforward approach is inspired by [48] who contrasted the language model BERT [31] and linear SVM [49] for text classification tasks, emphasising the trade-off between enhanced performance and computational expense.

For this study, various text classifiers were implemented to classify maintenance measure descriptors into RDS-PP categories. In this realm of machine learning, such label categorisation is termed as "predictions". These text classifiers do not differ in their classification method but in their training data set and the detail of the RDS-PP predictions. A comprehensive overview of all scenarios, including training and test set specifications, is provided in Table 1.

The scenario description is defined as follows: First, the data set the classifier is trained on is identified, whether it covers turbines of different original equipment manufacturers (OEMs) or solely from one OEM. Second, the type of predictions made by the classifier is specified. Scenarios 1 to 22 pertain to the varied hierarchy levels within RDS-PP, addressing all subsystems and component categories defined by RDS-PP. Higher levels offer more intricate component description. Scenarios 23 to 26 only focus on a subset of the principle subsystems that fail most frequently, with all other subsystems consolidated into an "other" category.

All scenarios utilise single-label classifiers, excluding maintenance reports that have multi label cases. This exclusion ensures a clear assignment of component categories during classifier training. Hence, the filtered data set contained only reports documenting a single component category. In each scenario, the first step is the selection of the respective subset from the whole data set, e.g. in scenario 1 the relevant subset is the maintenance reports from operator 1. For the experiments, four-fold cross-validation was used resulting in splits of 75% training data and 25% test data. To provide a consistent comparison of different classifiers, the training and test set sizes remain constant across respective scenarios, ensuring only one variable is evaluated simultaneously. In scenarios 7–10, which investigate the effect of training size, the data is sampled from the training data according to the indicated fraction. Scenarios 1 to 10 and 23 to 26 can be directly compared, while scenarios 11 to 16 and 17 to 22 should be examined independently. These scenarios aid in contrasting the efficacy of multiple classifiers and assessing the influence of the training set size and homogeneity.

To evaluate the performance of the text classifiers, F1 scores are utilised. These scores are derived from precision and recall [44]:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

The F1 score can range between 0 and 1. An F1 score of 1 indicates perfect precision and recall. Distinctions have been made between micro and macro F1 scores. Micro F1 scores account for label imbalance, while macro F1 scores are determined using unweighted means for each label [44]. For multi-class classification, where each sample has only one valid classification result, micro F1 scores are equivalent to accuracy measures. In order to conduct a statistical analysis, a five-times four-fold cross-validation is employed, repeating the classifier evaluation 20 times with text classifiers trained on randomly sorted training and test data sets. Subsequently, mean values for macro and micro F1 scores are computed and compared for different test scenarios. Boxplot analysis was conducted to measure the variance of the computed metrics due to the dataset size and imbalance.

TABLE 1 Overview of all presented classifier test scenarios.

Scenario	Scenario description	Training set	Test set
1	Operator1_up_to_level2_100%	1787 maintenance reports of operator 1	595 maintenance reports of operator 1
2	Operator1_up_to_level3_100%	1787 maintenance reports of operator 1	595 maintenance reports of operator 1
3	Operator1_up_to_level4_100%	1787 maintenance reports of operator 1	595 maintenance reports of operator 1
4	Operator2_up_to_level2_100%	1787 maintenance reports of operator 2	595 maintenance reports of operator 1
5	Operator2_up_to_level3_100%	1787 maintenance reports of operator 2	595 maintenance reports of operator 1
6	Operator2_up_to_level4_100%	1787 maintenance reports of operator 2	595 maintenance reports of operator 1
7	Operator1_up_to_level2_10%	179 maintenance reports of operator 1	595 maintenance reports of operator 1
8	Operator1_up_to_level2_25%	447 maintenance reports of operator 1	595 maintenance reports of operator 1
9	Operator1_up_to_level2_50%	894 maintenance reports of operator 1	595 maintenance reports of operator 1
10	Operator1_up_to_level2_75%	1340 maintenance reports of operator 1	595 maintenance reports of operator 1
11	OEM1_up_to_level2	1148 maintenance reports of operator 1 only including OEM1 reports	383 maintenance reports of operator 1 only including OEM1 reports
12	OEM1_up_to_level3	1148 maintenance reports of operator 1 only including OEM1 reports	383 maintenance reports of operator 1 only including OEM1 reports
13	OEM1_up_to_level4	1148 maintenance reports of operator 1 only including OEM1 reports	383 maintenance reports of operator 1 only including OEM1 reports
14	Operator1_up_to_level2	1148 maintenance reports of operator 1	383 maintenance reports of operator 1 only including OEM1 reports
15	Operator1_up_to_level3	1148 maintenance reports of operator 1	383 maintenance reports of operator 1 only including OEM1 reports
16	Operator1_up_to_level4	1148 maintenance reports of operator 1	383 maintenance reports of operator 1 only including OEM1 reports
17	OEM2_up_to_level2	638 maintenance reports of operator 1 only including OEM2 reports	213 maintenance reports of operator 1 only including OEM2 reports
18	OEM2_up_to_level3	638 maintenance reports of operator 1 only including OEM2 reports	213 maintenance reports of operator 1 only including OEM2 reports
19	OEM2_up_to_level4	638 maintenance reports of operator 1 only including OEM2 reports	213 maintenance reports of operator 1 only including OEM2 reports
20	Operator1_up_to_level2	638 maintenance reports of operator 1	213 maintenance reports of operator 1 only including OEM2 reports
21	Operator1_up_to_level3	638 maintenance reports of operator 1	213 maintenance reports of operator 1 only including OEM2 reports
22	Operator1_up_to_level4	638 maintenance reports of operator 1	213 maintenance reports of operator 1 only including OEM2 reports
23	Operator1_7categories_100%	1787 maintenance reports of operator 1	595 maintenance reports of operator 1
24	Operator1_5categories_100%	1787 maintenance reports of operator 1	595 maintenance reports of operator 1
25	Operator1_4categories_100%	1787 maintenance reports of operator 1	595 maintenance reports of operator 1
26	Operator1_3categories_100%	1787 maintenance reports of operator 1	595 maintenance reports of operator 1

While F1 score, which harmonises precision and recall, is widely used, additional metrics like precision–recall (PR) curves and Area Under the receiver operating characteristic curve (AUC-ROC) offer nuanced insights.

- Precision–recall curves: PR curves are particularly valuable for imbalanced datasets, highlighting the trade-off between precision and recall across different thresholds. They provide a clearer picture of classifier performance when positive classes are rare [50].
- AUC-ROC: AUC-ROC evaluates the overall ability of the model to discriminate between classes, plotting true positive rate against false positive rate. It is robust to class imbalance and offers a single scalar value summarising performance across all thresholds [51].
- Recent advances: Recent studies advocate combining these metrics for a more holistic evaluation. For instance, [52] emphasize the importance of using multiple metrics to avoid misleading conclusions in model performance evaluation.

The classification into component labels is based on the hierarchical structure of RDS-PP; meaning it organises information at multiple levels, with each subsequent level providing more detailed specifications. The codes represent broad categories of

TABLE 2 Example evaluation of text classifiers' prediction into RDS-PP labels.

Maintenance description	Model prediction	True label	Evaluation
Converter control board exchanged	MSE10 KF001	MSE10 KF001	Fully correct
Converter control board exchanged	MSE10	MSE10 KF001	Soft correct
Converter control board exchanged	MDA11	MSE10 KF001	False

components or systems at the highest levels, while the lower levels refer to specific parts and their functionalities. This hierarchical structure allows for detailed, systematic categorisation that facilitates better management and analysis of maintenance data.

RDS-PP codes are alphanumeric combinations where each code segment conveys specific information about the component's function, type and location within the overall system. For instance, a code like "MSE10 KF001" is a precise identifier: "MSE" denotes the converter system, "10" specifies an overall subsystem within the converter, and "KF001" indicates a particular control system component within that subsystem. This structured naming approach ensures that every part of a wind turbine is uniquely and consistently identified. Each RDS-PP code segment builds upon the previous one, offering increasing detail.

Evaluations are split between "fully correct" labels, where the text classifier's prediction matches the manually pre-processed label, and "soft correct" labels. The latter occurs when a text classifier predicts a label higher up in the RDS-PP hierarchy compared to the manual label, therefore, not being wrong but more generic than possible. Table 2 offers three illustrative examples: "MSE10 KF001" represents "control system converter system overall", whereas "MSE10" is a broader descriptor of the "converter system overall". A false prediction is exemplified where "MDA11" signifies "rotor blade system 1". Given this dual evaluation approach, each F1 score (macro and micro) is calculated for both "fully correct" and "soft correct" evaluation, respectively.

3.2.2 | Failure rate calculation

In the subsequent phase of this study, average failure rates per wind turbine per year are calculated using differently pre-processed data sets, aiming to quantify their impact on reliability KPIs. The average failure rate f of a specific subsystem is expressed as the ratio of the sum of all failures N of that subsystem over a given time frame to the total number of operational wind turbine years observed within this period T:

$$f = \frac{\sum_{i=1}^{I} N_i}{\sum_{i=1}^{I} X_i T_i} = \frac{N}{T} \quad (2)$$

Herein, N_i denotes the number of failures of the analysed subsystem in the time interval i, X_i represents the count of wind turbines examined during this interval and T_i is the span of the time interval.

Contrasting with the initial segment of uncertainty analysis, wherein random maintenance reports were selected to establish the training and test data sets, this phase uses continuous maintenance reports series in chronological order as test sets to ensure the derived KPIs are meaningful.

4 | RESULTS AND DISCUSSION

4.1 | Performance comparison of text classifiers based on different models

At first, preliminary experiments with a SVM, a CNN [26] with pre-trained word embeddings and a fine-tuned Transformer variant XLM-RoBERTa [53] were performed. A requirement for the experiments was to have access to the model which excluded proprietary models such as ChatGPT. In addition, the utility of open-source models for classification tasks when dealing with technical language is so far lacking [54]. The experiments were performed with an 80-20 train-test split and labels on the most precise level. Table 3 reports the performance of each model and shows that the linear SVM outperformed the other models. Moreover, traditional methods like TF-IDF and SVM are less computationally intensive and can be more cost-effective for specific datasets and tasks. For instance, in our scenarios where the volume of text data is manageable and the complexity of the language is not exceedingly high, these methods can provide competitive performance with significantly lower computational overhead. Therefore, all further experiments were conducted following the SVM approach.

4.2 | Performance comparison of text classifiers trained on different data sets

4.2.1 | How well does a classifier perform for different levels of detail?

Figure 2 presents the F1 scores for test scenarios 1, 2, and 3, comparing text classifiers. Although all are trained and tested on data sets of the same size, they predict components with varying degrees of detail. An up-to-level-4 classifier can precisely predict a label when the maintenance description provides ample information. However, if the maintenance report is not detailed, it

TABLE 3 Comparison of three model architectures, SVM, CNN and XLM-RoBERTa.

	Linear SVM	CNN	XLM-RoBERTa
Macro F1	0.41	0.34	0.34
Accuracy	0.71	0.67	0.68

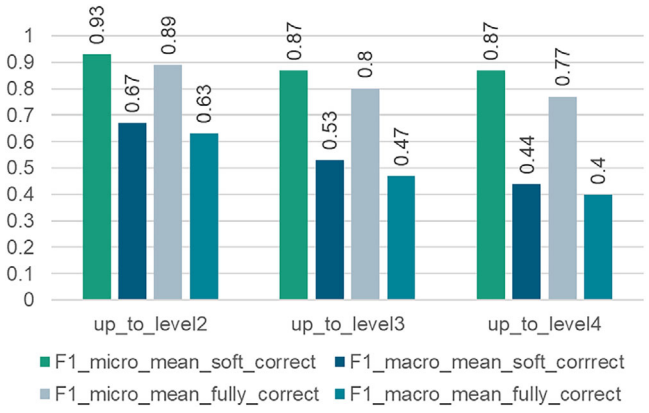

FIGURE 2 Comparison of F1 scores for test scenarios 1 to 3.

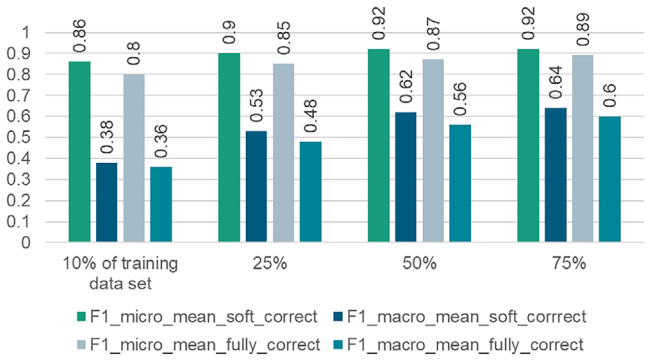

FIGURE 3 Comparison of F1 scores for test scenarios 7 to 10.

might opt for more generalised labels from RDS-PP hierarchy levels two and three. In comparison, an up-to-level-2 classifier always predicts the broader subsystem, corresponding to RDS-PP hierarchy level 2, even when the maintenance reports are more informative.

The findings indicate that as the granularity of the target label increases, classifier performance decreases. This outcome is intuitive, as the more nuanced predictions classifiers can make, they are faced with a greater variety of potential classification categories, intensifying the challenge. A subsequent boxplot analysis revealed a minor fluctuation in in micro F1 scores. In contrast, macro F1 scores experienced more significant variability. This disparity might be attributed to classifications of component categories that are infrequent in a dataset. Dependent on the train-test division, these rare categories might be predicted less accurately which is more reflected in the macro F1 scores than in the micro F1 scores.

4.2.2 | How many data points are required for training to see sufficient classification results?

Manual labelling of service reports is a meticulous and time-intensive task. Given this, the study assessed the efficacy of classifiers trained on smaller data sets. Figure 3 delineates results for test scenarios 7 to 10. These scenarios involve classifiers trained on 75%, 50%, 25%, and 10% of the original training data set size of test scenario 1, respectively.

A general trend is apparent: as the size of the training dataset reduces, the attainable F1 scores deteriorate. Nevertheless, a comparative assessment between scenario 1 (using the original training data set size) with scenario 10 (utilising 75% of original training data set size) reveals only marginal performance declines. Remarkably, even with a substantially truncated training data set, as in scenario 7 (10% of original training data set size), micro F1 scores are still quite competitive. In contrast, macro F1 scores decrease substantially when decreasing the training data set size. Depending on the focus of analysis, the laborious manual effort expended on labelling to devise a training dataset could be considerably reduced if the post-labelling analysis is principally concerned with frequently occurring components or subsystems.

From a machine-learning standpoint, these findings are somewhat unexpected. Conventionally, one would anticipate significantly enhanced classifier performance with more extensive training data sets. One plausible explanation for the reduced impact of training data set size on results might be inconsistencies inherent within the training data set. Such inconsistencies can diminish the advantages offered by larger datasets.

Examining the manual labelling process that employs RDS-PP for component categorisation supports this hypothesis. Classifying based on RDS-PP is not always straightforward for specific wind turbine components. Challenges arise due to ambiguities in distinguishing between different technical setups within RDS-PP. Moreover, RDS-PP guidelines might lack explicit definitions for certain component categories. This ambiguity leaves it to experts to decide the most fitting category, potentially leading to inconsistencies in the labelling process.

4.2.3 | Does the classification result improve when the training set is more specific?

The results of previously presented test scenarios suggest that larger training data sets do not offer significant advantages in terms of text classifier performance. Consequently, scenarios 11 to 22 were established to assess whether the classification results enhance when the training data sets are more tailored. From an engineering standpoint, the available service reports were scrutinised for pronounced differences. A significant observation was the variance in component naming conventions across different OEMs. As such, test scenarios 11 to 13 and 17 to 19 trained specific text classifiers based on data of two distinct OEMs. To draw a comparison with the comprehensive operator classifiers without confounding various factors, scenarios 14 to 16 and 20 to 22 used classifiers trained on randomly selected data points from operator 1, comprising OEM1 and OEM2 data. However, the training data set size was consistent with the OEM-specific scenarios. The performance evaluation of these classifiers was executed on OEM-centric data.

The outcomes of some of these test scenarios are illustrated in Figure 4. When comparing F1 scores for up-to-level-2 (scenario 11) and up-to-level-3 (scenario 12) predictions for OEM1

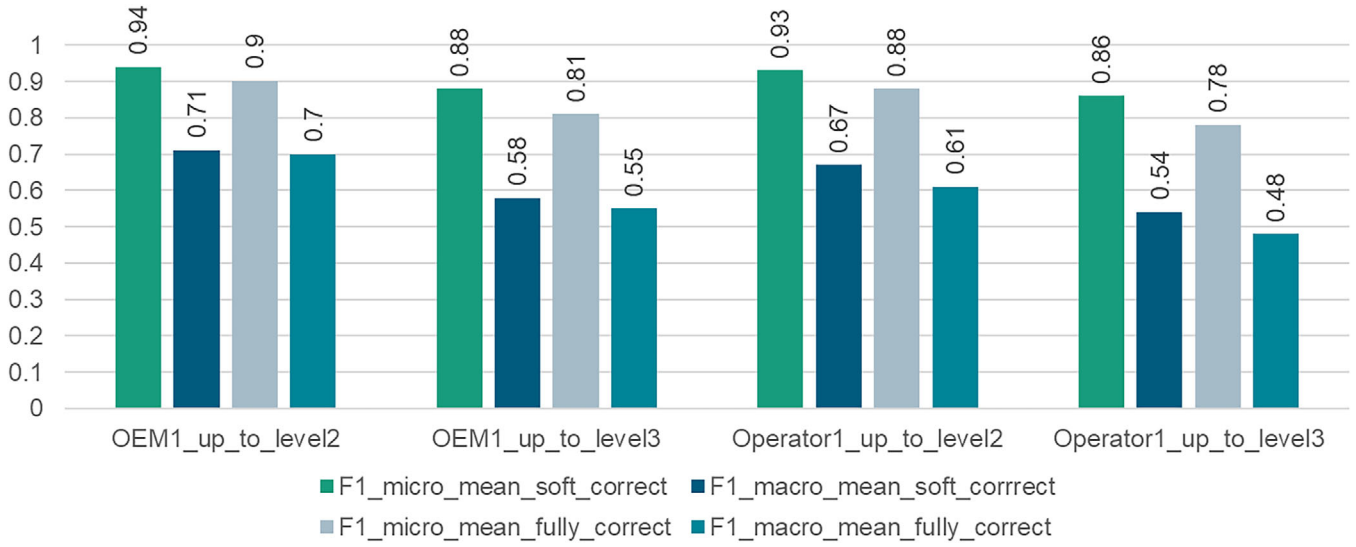

FIGURE 4 Comparison of F1 scores for test scenarios 11, 12, 14, and 15.

with predictions from classifiers trained on the more generic operator 1 data (scenarios 14 and 15, respectively), the classifiers using the more tailored data exhibit marginally better results across all F1 metrics, albeit the differences are not significant. Similar patterns emerged for up-to-level-4 classifiers (scenarios 13 and 16) and classifiers tested on OEM2 data (scenarios 17 to 22). It is noteworthy, known from manual labelling, that variations in terminology exist for identical component categories, contingent on the OEM. These discrepancies can be attributed to factors such as distinct component suppliers, wind turbine manuals and O&M procedure description semantics. Service technicians, influenced by these factors, employ varied terminology in describing and documenting their tasks. As a result, it might be advantageous to utilise smaller, yet more specific data sets for training text classifiers, especially when the goal is categorising components of a specific technology group, like a particular OEM.

4.2.4 | How much does the classification result improve when less label categories need to be predicted?

Depending on the information that needs to be extracted from the labelled maintenance reports and the focal points of the data analysis, varying levels of detail in the labelling process become necessary. For analyses targeting the subsystem level, the RDS-PP level 2 would be sufficient. Often, operators are mainly concerned with the most critical subsystems, which typically correspond to the most frequently failing ones. In such instances, classifiers were trained to distinguish only among these predominant subsystems, grouping all other subsystems in the "other" category. This approach was pursued to investigate the potential enhancement in classification outcomes when fewer label categories are required for predictions.

TABLE 4 Most frequently failing subsystems within the analysed data set.

	Subsystem	Corresponding RDS-PP code
1	Rotor system (incl. pitch system)	MDA
2	Converter system	MSE
3	Control system	MDY
4	Drive train system (incl. main bearing and gearbox)	MDK
5	Power generation system (incl. generator)	MKA
6	Yaw system	MDL

An overview of the most frequently failing subsystems is provided in Table 4 together with the RDS-PP codes. If we were to consider the subsystems based on their frequency of mention in the maintenance reports, the drive train system swaps places with the control system.

Based on these findings, the text classifier in test scenario 23 is configured to predict the six subsystems highlighted in Table 4, reverting to the label "other" for instances where none of the specific categories are applicable. This configuration yields a total of seven distinct categories. In contrast, text classifiers from test scenarios 24 to 26 increasingly relegate more subsystems under the "other" category. For example, in test scenario 26, the classifier discerns among the three different categories: MDA, MSE or "other".

The corresponding F1 scores of these classifiers are shown in Figure 5. As these test scenarios employ predefined categories rather than a hierarchical system, distinctions between "soft correct" and "fully correct" evaluations become redundant. In these cases, predictions are either entirely accurate or mistaken, which means only two F1 scores per classifier are depicted in Figure 5. The results indicate that narrowing the focus to frequent labels, while reducing the overall number

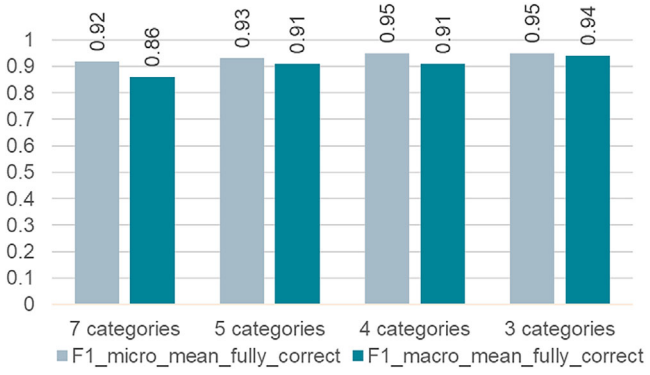

FIGURE 5 Comparison of F1 scores for test scenarios 23 to 26.

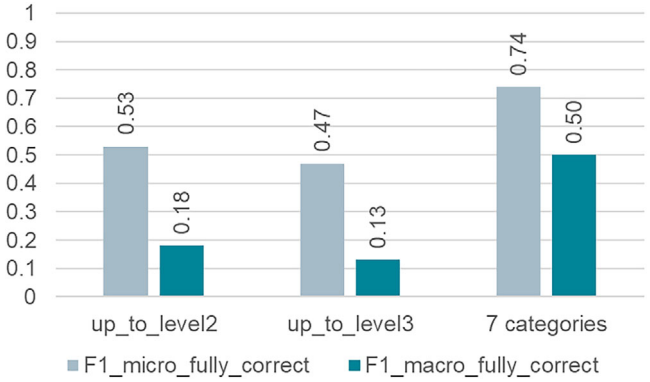

FIGURE 6 Classifier performance for different wind farms belonging to the portfolio of the same operator.

of labels, enhances classification performance. In comparison to the results of the up-to-level-2 classifier of test scenario 1, which registered a micro F1 score of 0.89 and macro F1 score of 0.63, scenarios 23 to 26 exhibit superior evaluation metrics. This convergence becomes more pronounced as the number of prediction categories diminishes, mitigating label imbalance and diminishing the effects of weighting.

4.2.5 | How well do the classifiers perform for different wind farms?

Until now, all discussed results were derived from classifiers trained and tested on data from the same wind farm or collective group of wind farms. Given the encouraging outcomes for practical application, the study sought to understand how these classifiers perform if applied on data sets from different wind farms.

In scenarios 4–6 classifiers were trained with the same level of detail as in scenario 1–3. However, training was undertaken using data from a distinct operator. These scenarios provided a preliminary insight into the adaptability if pre-trained classifiers are transferred to alternative datasets. F1 scores from these scenarios were only about half or two-thirds of those achieved in scenarios 1 to 3, as depicted in Figure 2. A plausible explanation for this could be variations in terminology and report structures across different organisations. These outcomes suggest that tailoring a classifier to each operator seems necessary, albeit being more labour-intensive due to the need for manual data labelling.

Subsequently, classifiers from test scenarios 1, 2, and 23 were assessed on a comprehensive dataset from wind farms not used during classifier training. These wind farms, however, were still under the portfolio of the original operator. Figure 6 presents the results, which clearly indicate a suboptimal performance relative to prior scenarios. Specifically, the up-to-level-2 and up-to-level-3 classifiers correctly predict only around half of the labels. Even though the up-to-level-2 classifier's performance mirrored that in test scenario 4, the up-to-level-3 classifier showed improved accuracy than in test scenario 5. This suggests possible similarities in terminology at the subsystem level, whereas greater disparities exist at the component level. When comparing these outcomes with results from test scenarios 1 and 2, it is evident that there is a significant decline in performance when classifiers are extended to different wind farms—even if they belong to the same operator. Such inconsistencies might arise from divergent documentation standards across service entities operating in varied regions. Notably, the 7-categories classifier managed to correctly predict 74% of the maintenance activities, outperforming the other two classifiers. However, in comparison with test scenario 23 in which 92% of labels were predicted correctly, it is clear that there is room for improvement.

Therefore, it has to be noted that when an existing trained classifier is to be applied to other wind farms the semantic needs to be analysed carefully. If significant differences emerge, investing in retraining the classifier can be beneficial—even if it necessitates additional manual efforts for curating a training dataset.

4.3 | Industry perspective on productive use of classifiers

To understand industry needs and pinpoint the most valuable classifiers, six structured face-to-face interviews were conducted with senior staff from different operators and service providers being active in asset management. These discussions were twofold: Firstly, to determine the requisite level of detail when labelling maintenance reports based on internal processes and secondly, to gauge preferences for the level of detail of classifiers considering the F1 scores achieved in the above test scenarios. The approach involved presenting interviewees with a series of either-or questions to discern their priorities. A visualisation of evaluated answers can be found in Figure 7.

From the data in this figure, it is evident that there are no clear tendencies among interviewees. Responses varied significantly across interviewees and organisations. Some showed a leaning towards a more generic classifier (up to level 2) with a higher performance, while others exhibited a bias for a more specific classifier (up to level 4) even if it came with slightly lower performance scores. Both these preferences were equally popular.

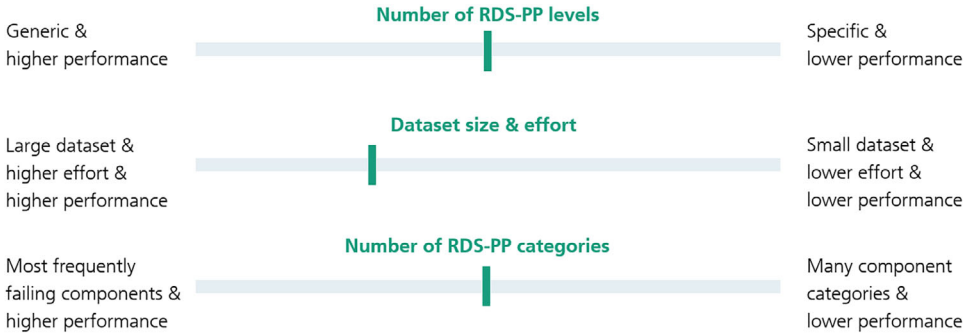

FIGURE 7 Summary of interviewees' preferences for classifier configurations using either-or questions.

Another aspect explored, was whether the interviewees would opt for (a) a high-performing classifier, trained on a comprehensive data set, even if it necessitates more labour-intensive and costly data preparation, or (b) a classifier with optimised performance only for regularly mentioned components within the maintenance reports as it is trained on a smaller data set. Here, there was a discernible tilt towards the former—a preference for larger datasets, even if they required increased effort.

Lastly, interviewees were asked about their interest in the classifiers from test scenarios 1, 23, 24, 25, and 26. These represented classifiers with approximately 25, seven, five, four and three distinct component categories, respectively. Half of the respondents favoured the up-to-level-2 classifier. The other half expressed a preference for the classifier which labels the six most frequently failing subsystems, relegating all other activities under the "other" category. Notably, none of the interviewees expressed an interest in the classifiers of test scenarios 24 to 26, which labelled fewer than six subsystems.

Contrary to the authors' expectations of receiving a consistent set of responses, the interviews revealed a variety of preferences for classifier configurations. This diversity is largely attributed to the distinct requirements and motivations inherent to each interviewed company. Hence, the idea of a "one fits all" solution is not deemed viable.

4.4 | Failure rate comparison of differently pre-processed data sets

To gauge the uncertainty tied in various pre-processed data sets, failure rates of wind turbines' subsystems and components were selected as KPIs, aside from machine learning metrics like F1 scores and accuracy. Industry often relies on failure rates to understand the frequency at which components and subsystems fail. They serve as important KPIs for both benchmarking operational wind farms and planning for future wind farm projects. Consequently, the authors aimed to discern the potential variation in these KPIs based on different pre-processing approaches applied to maintenance reports. Therefore, two distinct analyses were carried out: Firstly, maintenance reports were labelled using selected classifiers. Without having access to these results, the same reports underwent manual labelling. Secondly, maintenance reports were manually classified by two different organisations both using RDS-PP as labelling guidelines. Following these processes, the failure rates derived from each differently pre-processed data set were computed and set for comparison. The results of these investigations are detailed in the subsequent sections.

4.4.1 | Manual labelling versus text classifier

The data set and text classifiers used in this study were the same as the ones employed in the analysis presented in Section 4.2.5, with the results visualised in Figure 6. For the 13 subsystems that fail most frequently, normalised failure rates were deduced from the different pre-processed datasets and are illustrated in Figure 8.

To normalise the data, the failure rate of the most frequently failing subsystem, i.e. the rotor system (MDA) as per the manually labelled data set, was utilised. Although the F1 scores already conveyed that the classifiers' performance for varying wind farms (even from the same operator) might not be fit for productive use (cf. Figure 6), the normalised failure rates provide deeper insights into how these classifiers' function and perform.

Generally, the results reveal that the five most frequently failing subsystems are consistent with the ones of the data set initially used for both training and testing the classifiers (cf. Table 4). Only the order switched between the converter system (MSE) and control system (MDY). For many of the top failing subsystems, the failure rates calculated from classifier-processed data somewhat align with the manually labelled data. In contrast, categories like the environmental measuring system (CKJ) or ancillary systems (XMM, XSD) barely make a mark in the statistics when processed by the classifier. Interestingly, there are two major subsystems, namely the control system (MDY) and the power generation system (MKA), which showcase substantially reduced failure rates when deduced from classifier-labelled data. Therefore, this suggests that these component categories are difficult to predict for the classifiers, despite their relatively frequent occurrences in the training data set, in comparison to the ones of, e.g. ancillary systems. Due to the classifiers' inherent struggle to predict labels with a limited representation in the training data set, the failure rate of the category "G", which

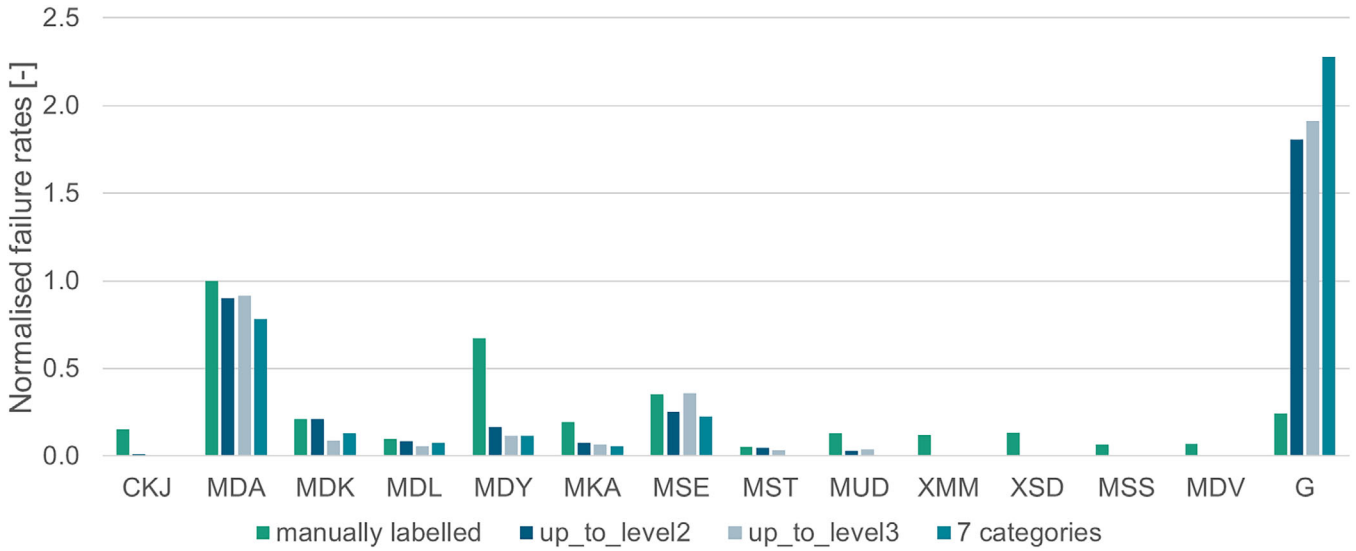

FIGURE 8 Comparison of normalised failure rates of differently pre-processed data sets (for translation of RDS-PP codes see Appendix 1).

stands for the "overall system energy conversion" and is the fallback label for all text descriptions which were not possible to be sorted into one of the more specific categories, is seven to nine times higher when derived from classifier-processed data, in contrast to the manually labelled data.

As concluded in Section 4.2.5, applying a pre-trained classifier to data from different wind farms, even if they are from the same operator, can lead to unsatisfactory categorisation results. However, the failure rate comparison demonstrates that the prediction for some subsystem categories is still reasonably accurate, e.g. when utilising the classifier predicting labels up to RDS-PP level 2, even though it had only half of its predictions correct for the entire dataset. Hence, failure rate deviations will be notably smaller for data sets labelled by a classifier specifically trained for them. Another approach would be to artificially enhance the training data set with text examples from subsystem or component categories that are underrepresented. This could help achieve comparable performance for typically less frequent labels to those that appear regularly. Moreover, given the common assignment to the overarching category "G", it might be advisable to first use a trained classifier for labelling the data set. Subsequently, entries labelled as category "G" can be manually relabelled. This approach would considerably reduce the manual work while maintaining the reliability of the result.

4.4.2 | Uncertainty related to manually labelling maintenance reports

In the previous section, we analysed the differences in failure rates that arose from differently labelled data sets, benchmarking the classifiers against the expertise of wind energy professionals. However, during the data preparation phase for training the classifiers, it became evident that even expertly labelled data can vary. To gauge the significance of this variation, failure rates from the same data set of operator 2, categorised by two distinct organisations, were calculated. The subsequent step involved determining the difference in failure rates as multiples from organisation 1. Results at the level of wind turbine subsystems are showcased in Figure 9. Variances span from 0.71 for category MSC (generator switching system) to a 3.5-fold higher failure rate for category XGM (Fire extinguishing system). Substantial discrepancies in KPIs are also evident in subsystems like the drive train system (MDK) and lifting gears (XMM) which exhibit 1.79- and 1.77-times higher failure rates, respectively.

This can be explained by several aspects:

1. Interpretation of RDS-PP guidelines: The guidelines of RDS-PP are crafted to categorise any wind turbine technology. As a result, component categories are not too specific. Some components could arguably fit into multiple RDS-PP categories, leaving room for interpretation. Even though both organisations engaged in regular discussions about such categorisation ambiguities, it remains challenging to unanimously decide which components "clearly" fall into a particular RDS-PP category. Ultimately, the decision is up to the individual.
2. Uncertainties in ZEUS labelling: The State-Event-Cause-Code "ZEUS" [55] served as a guide to standardise the labelling of a component's state or the corresponding maintenance actions on the wind turbine. Within this study, a failure is defined as a fault necessitating technician intervention and spare parts usage to restore the function of the wind turbine. Therefore, only corrective maintenance measures involving component replacement is deemed a failure event. The ambiguity arises when maintenance reports lack detail, leading experts to potentially judge certain replacements as corrective or planned differently. This can subsequently result in variant KPIs.

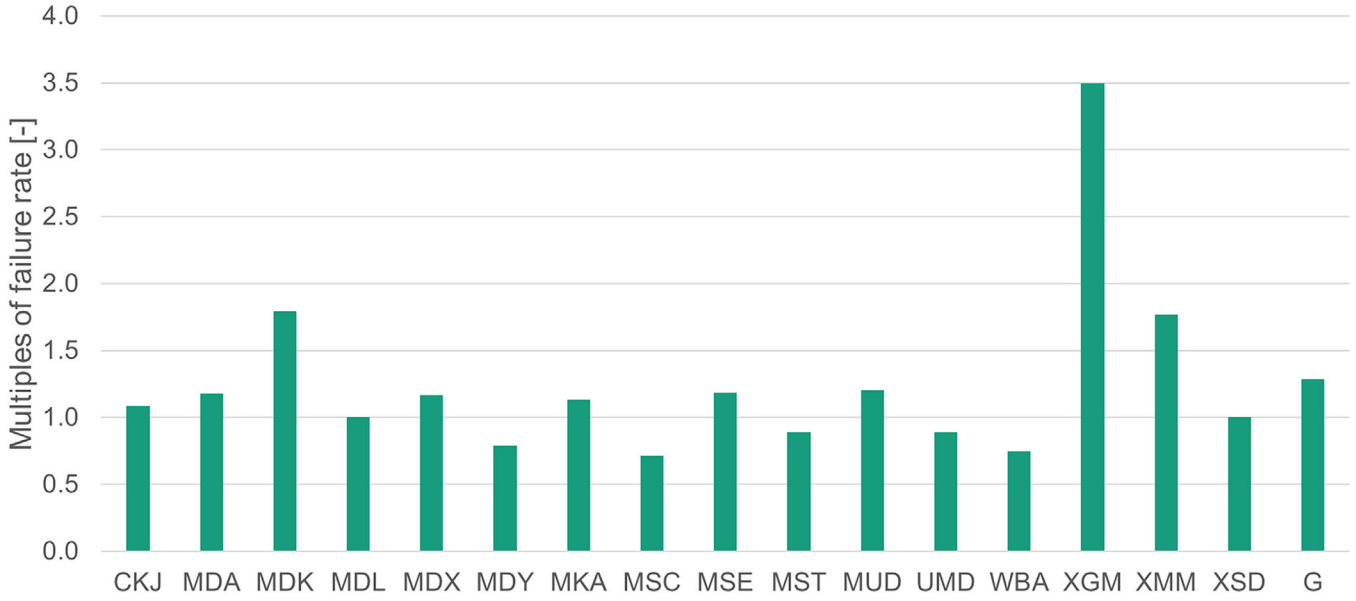

FIGURE 9 Multiples of failure rate for each wind turbine subsystem shown by RDS-PP categories comparing results based on pre-processed data sets by organisation 1 and organisation 2 (for translation of RDS-PP codes see Appendix 1).

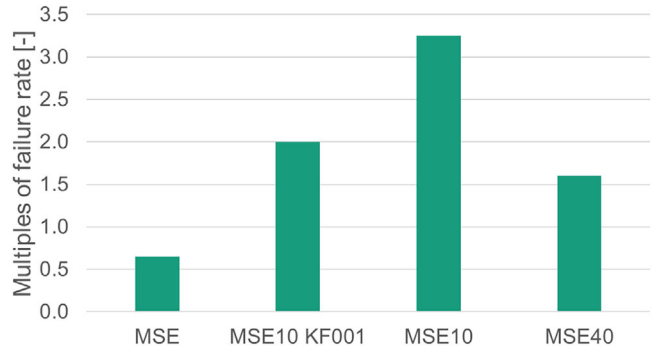

FIGURE 10 Multiples of failure rate exemplarily for components of the converter system shown by RDS-PP categories comparing results based on pre-processed data sets by organisation 1 and organisation 2 (for translation of RDS-PP codes see Appendix 1).

3. Human factors in manual labelling: Manually labelling maintenance reports is exhaustive and time-consuming. Factors like an individual's expertise, their mental state during labelling or their specialisation in either electrical or mechanical components can influence labelling decisions.

To illustrate the last point, a comparison is presented in Figure 10, showcasing failure rates at the component level. As an example, the subsystem frequency converter (MSE) is chosen. Within the subsystem, differentiation is made among three component categories (MSE10 KF001, MSE10, and MSE40), as well as the general category MSE. This general category accumulates all failures that were not described with enough precision in the maintenance reports to label a specific affected component.

In this example, the failure rates for specific component categories from organisation 2 are at least 1.6 times higher than the ones of organisation 1. Meanwhile, the failure rate for the general category is 0.65 times lower in comparison. This discrepancy could indicate the presence of experts in the field of power electronics in organisation 2 who might be more confident in labelling specific component categories over a broader category, given the more detailed information available in the maintenance reports. However, this information might not be as comprehensible to individuals from different engineering backgrounds. Since RDS-PP adopts a hierarchical structure, choosing a more general category for labelling is not incorrect, though the aim should always be to label as specifically as possible. Yet, this approach can also result in variances in failure rates, particularly at the component level. This observation emphasises the challenges inherent to comparing failure rates across different publications, even when they employ the same taxonomies or failure definitions.

This critical finding underlines the necessity of further efforts in standardising the labelling process of maintenance reports also across organisations. While standards and guidelines as RDS-PP and ZEUS give recommendations how to proceed, the forementioned analyses have shown that instructions seem to fail achieving consistency. Therefore, specific examples of maintenance descriptions and how to apply these standards would be beneficial. Most helpful would be a parts list of each turbine type provided by the OEMs with respective RDS-PP translations attached. Making such information publicly available would greatly contribute to consistent data preprocessing allowing for better interpretation and comparability of KPI calculations.

4.4.3 | Barriers to the adoption of text classifiers and potential applications in the wind energy sector

One of the primary technological barriers is the variability and inconsistency in maintenance report formats. Different operators use diverse terminologies and reporting standards,

complicating the training of robust classifiers. The implementation of industry-wide standards, such as RDS-PP, can mitigate this issue by providing a uniform framework for categorising maintenance activities. Furthermore, while text classifiers can achieve high accuracy, their performance can vary significantly based on the quality and representativeness of the training data. Ensuring that classifiers are trained on comprehensive and diverse datasets is crucial to maintain reliability across different wind farms and operators. Additionally, integrating NLP models with existing maintenance management systems (MMS) and enterprise resource planning (ERP) systems poses a challenge. Seamless integration requires APIs and middleware that can handle the specific data structures and workflows of these systems.

Adoption of new technologies necessitates significant change management. Maintenance staff and engineers need to be trained to trust and effectively use these automated systems. There might be resistance due to perceived threats to job security or scepticism about the reliability of automated systems. Moreover, introducing text classifiers into established workflows can initially disrupt operations. Careful planning and phased implementation, starting with pilot projects, can help mitigate disruption and demonstrate the benefits gradually.

Developing, training, and integrating text classifiers involves upfront costs, including technology investments, data labelling efforts, and training programs for staff. For smaller operators, these costs might be prohibitive without clear demonstrations of return on investment (ROI). Furthermore, text classifiers require ongoing maintenance and updates to handle new terminologies, equipment, and failure modes. This ongoing cost needs to be factored into the economic feasibility of adopting such technology.

The integration of text classifiers can be envisioned through several steps. First, an initial data assessment is essential to evaluate the quality and standardisation of existing maintenance logs. Following this, a pilot project can be implemented in a controlled environment, such as a single wind farm or a specific subset of maintenance reports. This pilot phase allows for testing and adjustments before broader deployment. Training and onboarding sessions for maintenance staff and engineers are crucial to familiarise them with the new system and its benefits. Eventually, a full-scale implementation can be pursued, gradually expanding the use of text classifiers across all operations, ensuring continuous monitoring and feedback.

Incorporating text classifiers into current MMS and ERP systems can significantly enhance their functionality. Systems like SAP PM (plant maintenance) or IBM Maximo, which manage extensive maintenance records and data from diverse sources, including sensor readings, operational logs, and manual reports, can benefit significantly. These systems can automate the categorisation and standardisation of maintenance records by embedding text classifiers. This automation facilitates easier tracking and analysis of component failures and maintenance activities, thus improving data accuracy and efficiency. Additionally, platforms such as Microsoft Azure and AWS provide NLP services that can be tailored for specific industry needs, offering scalable and secure deployment options. Demonstrating real-world applicability involves showcasing the efficiency gains and accuracy improvements in maintenance data processing. By leveraging cloud-based solutions, real-time data analytics, and user-friendly interfaces, the adoption of text classifiers can be streamlined, providing tangible benefits in terms of reduced downtime and optimised maintenance schedules.

Furthermore, text classifiers can be seamlessly integrated with predictive maintenance tools that utilise data and machine learning algorithms to foresee equipment failures before they happen. Accurate and standardised maintenance records provided by text classifiers improve the precision of predictive models. This enhancement leads to more effective maintenance strategies, further preventing unexpected breakdowns and extending the life of turbine components. (Figure 11)

5 | CONCLUSIONS AND OUTLOOK

This study assessed the viability of text classifiers for preprocessing wind turbine maintenance reports, highlighting their potential to reduce manual data processing efforts significantly. Main conclusions can be summarised as follows:

- Text classifiers achieved high micro F1 scores when trained on specific datasets, demonstrating their effectiveness. However, their performance decreased when applied to different wind farms, indicating the necessity for context-specific training.
- The research also underscored the importance of cost and resource efficiency, showing that smaller, well-curated training datasets can still produce competitive results. This finding emphasises the need to balance manual labelling efforts with classifier performance for practical application.
- Industry feedback revealed diverse classifier configuration preferences, suggesting that custom solutions are essential to meet varied stakeholder needs.
- While text classifiers tended to over-generalise, leading to skewed KPI calculations, they remain valuable when combined with manual verification for critical categories, enhancing overall reliability.
- A significant insight from the study is the need for standardisation in maintenance reporting. Both automated and manual methods face uncertainties due to inconsistent documentation. Standardised designation systems like RDS-PP can improve data accuracy and reliability, resulting in more meaningful KPIs.

Looking ahead, large language models (LLMs) such as GPT-3 and GPT-4 offer the potential to overcome current limitations. Fine-tuning these models with domain-specific datasets may enhance their applicability in the wind energy sector, improving classification performance. Successful applications of encoder models in other fields, like healthcare (BioBERT) and finance (FinBERT), provide blueprints [56, 57]. Additionally, developing comprehensive datasets that capture the technical jargon and variations in maintenance reports, along with better guidelines for applying standards like RDS-PP, will

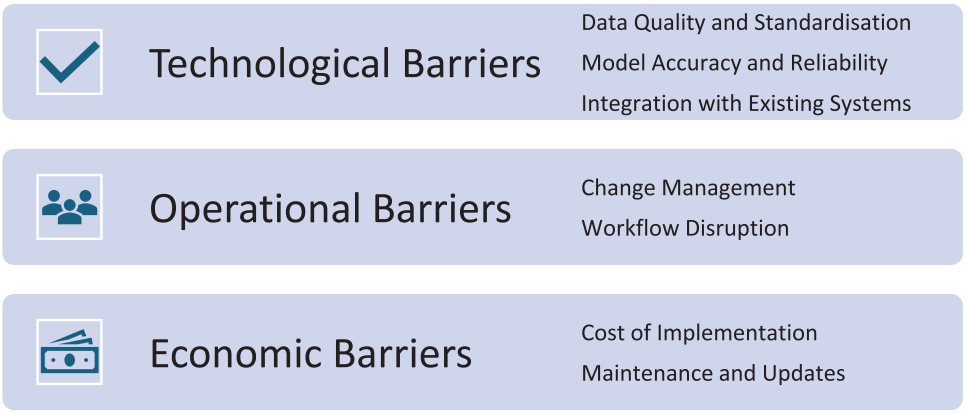

FIGURE 11 Barriers to the adoption of text classifiers.

be crucial. Integrating text classifiers into maintenance management and ERP systems can enhance operational efficiency and decision-making. By focusing on these future directions, this study aims to contribute to the improvement of maintenance data processing in the wind energy industry, ensuring more accurate and reliable analysis and reporting.

AUTHOR CONTRIBUTIONS

Julia Walgern: Conceptualisation; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; visualisation; writing—original draft. **Katharina Beckh**: Methodology; software. **Neele Hannes**: Data curation; formal analysis. **Martin Horn**: Data curation. **Marc-Alexander Lutz**: Data curation. **Katharina Fischer**: Funding acquisition; supervision; writing—review and editing. **Athanasios Kolios**: Funding acquisition; supervision; writing—review and editing.

ACKNOWLEDGEMENTS

The present work was mostly carried out within the research project "Digitalisation of Maintenance Information (DigMa)" funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), grant number 03EE2016A. The authors thank the project partners for providing field data and for sharing their requirements and experience. Further financial support was received by EPSRC through the Wind and Marine Energy Systems Centre for Doctoral Training under the grant number EP/S023801/1.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Research data cannot be shared due to confidentiality.

ORCID

Julia Walgern https://orcid.org/0000-0003-4042-9461
Katharina Beckh https://orcid.org/0000-0002-7824-6647
Marc-Alexander Lutz https://orcid.org/0000-0003-1962-0333
Katharina Fischer https://orcid.org/0000-0001-5737-1572
Athanasios Kolios https://orcid.org/0000-0001-6711-641X

REFERENCES

1. Cevasco, D., Koukoura, S., Kolios, A.: Reliability, availability, maintainability data review for the identification of trends in offshore wind energy applications. Renewable Sustainable Energy Rev. 136, 110414 (2021)
2. Pfaffel, S., Faulstich, S., Rohrig, K.: Performance and reliability of wind turbines: a review. Energies 10(11), 1904 (2017)
3. Gayo, J.B.: Final publishable summary of results of Project ReliaWind (2011). https://scholar.google.com/scholar_lookup?title=Final%20Publishable%20Summary%20of%20Results%20of%20Project%20ReliaWind&author=J.B.%20Gayo&publication_year=2011
4. Carroll, J., McDonald, A., McMillan, D.: Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. Wind Energy 19, 1107–1119 (2016)
5. SPARTA: Portfolio Review 2016: System Performance, Availability and Reliability Trend Analysis. SPARTA Project, Northumberland (2017)
6. Fischer, K., et al.: Reliability of power converters in wind turbines: exploratory analysis of failure and operating data from a worldwide turbine fleet. IEEE Trans. Power Electron. 34(7), 6332–6344 (2019)
7. Fischer, K., Pelka, K., Walgern, J.: Trends and influencing factors in power-converter reliability of wind turbines. In: PCIM Europe 2023; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management, pp. 1–10. IEEE, Piscataway, NJ (2023)
8. Padman, P., Vanni, F., Echavarria, E., Mortstock, K., Wilkinson, M.: Benchmarking Pitch System Reliability and Reducing Cost of Energy through Advanced Design. in EWEA-Workshop "Analysis of Operating Wind Farms", Bilbao, (2016). https://scholar.google.com/scholar_lookup?title=Benchmarking%20pitch%20system%20reliability%20and%20reducing%20cost%20of%20energy%20through%20advanced%20design&author=P.%20Padman&publication_year=2016
9. Walgern, J., Fischer, K., Hentschel, P., Kolios, A.: Reliability of electrical and hydraulic pitch systems in wind turbines based on field-data analysis. Energy Rep. 9, 3273–3281 (2023)
10. Kenworthy, J., et al.: Wind turbine main bearing rating lives as determined by IEC 61400–1 and ISO 281: A critical review and exploratory case study. Wind Energy 27(2), 179–197 (2023)
11. Anderson, F., Dawid, R., McMillan, D., Garcia Cava, D.: On the Sensitivity of Wind Turbine Failure Rate Estimates to Failure Definitions. J. Phys. Conf. Ser. 2626 (2023)
12. VGB PowerTech: VGB-Standard RDS-PP Application Guideline Part 32: Wind Power Plants (2014). https://www.vgb.org/shop/s-823-32.html
13. Danish Standards Foundation: A Guide to RDS—Reference Designation Systems. TAG Numbers for Systems in Accordance with the ISO/IEC

81346 Standard Series" (2020). https://webstore.ansi.org/standards/ds/dshandbook1662017
14. Lutz, M.-A., et al.: Digitalization workflow for automated structuring and standardization of maintenance information of wind turbines into domain standard as a basis for reliability KPI calculation. J. Phys.: Conf. Ser. 2257, 012004 (2022)
15. Lutz, M.-A., et al.: KPI extraction from maintenance work orders—a comparison of expert labeling, text classification and AI-assisted tagging for computing failure rates of wind turbines. Energies 16(24), 7937 (2023)
16. Ghazizadeh, E., Zhu, P.: A systematic literature review of natural language processing: current state, challenges and risks. Adv. Intell. Syst. Comput. 1288, 634–647 (2021)
17. Zhang, Z., Strubell, E., Hovy, E.: A survey of active learning for natural language processing. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, pp. 6166–6190. Association for Computational Linguistics, Stroudsburg, PA (2022)
18. Joachims, T.: Text categorization with support vector features. In: European Conference on Machine Learning, pp. 137–142. Artificial Intelligence Association of Lithuania, Vilnius (1998)
19. Shi, L., Zhu, Y., Zhang, Y., Su, Z.: Fault diagnosis of signal equipment on the Lanzhou-Xinjiang high-speed railway using machine learning for natural language processing. Complexity 2021, 9126745 (2021)
20. Siva Balan, R., Walia, K., Gupta, K.: A systematic review on POS tagging. In: 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022, pp. 1531–1536. IEEE, Piscataway, NJ (2022)
21. Li, D., Luo, S., Zhang, X., Xu, F.: Review of named entity recognition. J. Front. Comput. Sci. Technol. 16(9), 1954–1968 (2022)
22. Abdar, M., et al.: A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf. Fusion 76, 243–297 (2021)
23. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)
24. Amjad, M., Gelbukh, A., Voronkov, I., Saenko, A.: Comparison of text classification methods using deep learning neural networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 438–450. Springer, Cham (2023)
25. Wang, Z., Zhang, Z.: Research convey on text classification method based on deep learning. In: 7th International Conference on Intelligent Computing and Signal Processing, ICSP 2022, pp. 285–288. IEEE, Piscataway, NJ (2022)
26. Kim, Y.: Convolutional neural networks for sentence classification. arXiv:1408.5882 (2014)
27. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997)
28. Zhang, J., et al.: Text classification method based on improved long short term memory network. Proc. SPIE 13105, 1310528 (2024)
29. Yuemei, X., Zuwei, F., Han, C.: A multi-task text classification model based on label embedding of attention mechanism. Data Anal. Knowl. Discov. 6(2–3), 105–116 (2022)
30. Vaswani, A., et al.: Attention is all you need. arXiv:1706.03762 (2017)
31. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, pp. 4171–4186. Association for Computational Linguistics, Stroudsburg, PA (2019)
32. Arabadzhieva-Kalcheva, N., Kovachev, I.: Comparison of BERT and XLNet accuracy with classical methods and algorithms in text classification. In: Proceedings of the International Conference on Biomedical Innovations and Applications, BIA 2021, pp. 74–76. IEEE, Piscataway, NJ (2021)
33. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901 (2020)
34. Elkhatat, A., Elsaid, K., Almeer, S.: Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int. J. Educ. Integr. 19(1), 17 (2023)
35. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019)
36. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn Res. 21(140), 1–67 (2020)
37. Wang, P., et al.: OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. Proc. Mach. Learn. Res. 162, 23318–23340 (2022)
38. Kiela, D., et al.: Supervised multimodal bitransformers for classifying images and text. arXiv:1909.02950 (2019)
39. Bhavani, A., Santhosh Kumar, B.: A review of state art of text classification algorithms. In: Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, pp. 1484–1490. IEEE, Piscataway, NJ (2021)
40. Vinod, S., et al.: Advancements in text classification, a comprehensive review. Humanitarian Technology Conference, R10-HTC, pp. 679–684. IEEE, Piscataway, NJ (2023)
41. Sun, Y., Wang, L., Huang, Y., Zhang, Z.: Multimodal knowledge graph representation learning with entity description. Proc. SPIE 12717, 127171K (2023)
42. Hogan, A., et al.: Knowledge graphs. arXiv:2003.02320 (2021)
43. Wang, Q., Wang, B., Guo, L.: Knowledge graph embedding: a survey of approaches and applications. IEEE Trans. Knowl. Data Eng. 29(12), 2724–2743 (2017)
44. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn Res. (12), 2825–2830 (2011)
45. Platt, J.C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. MIT Press, Cambridge, MA (1999)
46. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2(2), 159–165 (1958)
47. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. 28(1), 11–22 (1972)
48. Lin, Y.-C., Chen, S.-A., Liu, J.-J., Lin, C.-J.: Linear classifier: an often-forgotten baseline for text classification. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1876–1888. Association for Computational Linguistics, Stroudsburg, PA (2023)
49. Boser, B.E., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152. ACM Press, New York, NY (1992)
50. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10(3), e0118432 (2015)
51. Huang, J., Ling, C.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. 17(3), 299–310 (2005)
52. Lipton, Z.C., Elkan, C., Naryanaswamy, B.: Thresholding classifiers to maximize F1 score. arXiv:1402.1892 (2014)
53. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116 (2019)
54. Yang, Z., Batra, S., Stremmel, J., Halperin, E.: Surpassing GPT-4 medical coding with a two-stage approach. arXiv:2311.13735 (2023)
55. Fördergesellschaft Windenergie und andere Erneuerbare Energien: Technical Guidelines for Power Generating Units—State-Event-Cause code for power generating units (ZEUS). Fördergesellschaft Windenergie und andere Erneuerbare Energien, Berlin (2013)
56. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)
57. Yang, X., Yuan, Q., Zhang, C.: FinBERT: a pretrained language model for financial communications. arXiv:2006.08097 (2020)

How to cite this article: Walgern, J., Beckh, K., Hannes, N., Horn, M., Lutz, M.-A., Fischer, K., Kolios, A.: Impact of using text classifiers for standardising maintenance data of wind turbines on reliability calculations. IET Renew. Power Gener. 1–17 (2024). https://doi.org/10.1049/rpg2.13151

APPENDIX A: REFERENCE DESIGNATION SYSTEM RDS-PP

For each maintenance measure, the concerned components are classified using the reference designation system RDS-PP for wind turbines [12]. In Table A1 all mentioned RDS-PP codes mentioned within this paper are summarised and translated.

TABLE A1 Summary and translation of all mentioned RDS-PP codes within this paper.

RDS-PP code	Translation
CKJ	Environmental measuring system
MDA	Rotor system (incl. pitch system)
MDA11	Rotor blade system 1
MDK	Drive train system (incl. main bearing and gearbox)
MDL	Yaw system
MDV	Central lubrication system
MDX	Central hydraulic system
MDY	Control system
MKA	Power generation system (incl. generator)
MSC	Generator switching system
MSE	Converter system
MSE10	Converter system overall, also denoted as "phase module" components including core power electronics (see [7])
MSE10 KF001	Control system converter system overall
MSE40	Heating/cooling converter systems
MSS	Compensation system
MST	Generator transformer system
MUD	Nacelle
UMD	Tower system
WBA	Personnel rescue systems
XGM	Fire extinguishing system
XMM	Lifting gears
XSD	Obstacle warning system
G	Overall system energy conversion