euspen's 26th International Conference & Exhibition, Krakow, PL, June 2026
www.euspen.eu

Automatic Text Analysis for Machine Maintenance in Float-Zone Crystal Growth Production

Eugenia Freggia[1], Marta Barbieri[2], Irene Spada[1], Gualtiero Fantoni[1], Matteo Calaon[2]

[1]University of Pisa, 56122 Pisa, Italy [2]Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

mabarb@dtu.dk

Abstract
The continuous evolution of electronics and optoelectronics has led to a growing demand for ultra-high-quality silicon wafers. In this context, the Float-Zone crystal growth technique plays a key role, as it enables the production of silicon with very low impurity levels and high electrical resistivity, making it suitable for high-power electronic devices and high-efficiency solar cells. Float-Zone production is characterized by high manufacturing costs, largely driven by the need for extremely pure and expensive raw material. Yield losses may occur in specific industrial contests due to process instabilities, equipment failures, and operational anomalies. In industrial setting, a substantial amount of knowledge related to machine behavior, failures, and corrective actions is recorded during daily operations in the form of unstructured maintenance logs. While this information is potentially valuable for improving yield and reducing costs, it is often difficult to access and exploit due to its fragmented, inconsistent, and multilingual nature. As a result, maintenance knowledge remains largely difficult to access and to exploit. This study addresses this challenge through a case study conducted in a Danish industrial setting involving Float-Zone silicon crystal growth. More than 18 thousand maintenance logs, exhibiting significant variability in structure and terminology, have been analyzed. Building on existing applications of Text Mining and Natural Language Processing in industrial maintenance, this work focuses on their targeted adaptation to a highly specialized and quality-critical production process. We propose tailored preprocessing, text classification and clustering to identify recurring families of problems and corrective actions, thereby structuring previously inaccessible maintenance knowledge. The analysis improves traceability between failures and actions, enhancing the understanding of issues related to process stability. The findings show the potential to support systematic and data-driven maintenance analysis and further predictive maintenance developments.

Keyword: Entities relation discovery, Data processing, Text information, Maintenance records

1. Introduction

The growing demand for electronic devices is driving the increasing production of high-purity silicon wafers. In this context, the Float-Zone (FZ) crystal growth technique is adopted because extremely high material quality is required although its high production cost [1]. Thus, yield improvements are critical objective both in manufacturing and maintenance. Larger volumes of data from industrial records and logs are increasingly leveraged within smart manufacturing and data-driven maintenance [2]. While most existing approaches focus on structured machine and sensor data, textual maintenance records remain largely underexplored despite containing valuable information on machine behavior, failures, and corrective actions.

This work addresses the posed challenge through the application of Text Mining (TM) and Natural Language Processing (NLP) techniques to maintenance logs from an industrial FZ silicon crystal growth case study. The proposed approach normalizes and structures unstructured textual data to identify recurring families of problems and corrective actions and analyze their relationships, thereby supporting improvements in manufacturing processes.

2. Methodology

The methodology consists of an initial text preprocessing phase and five main steps.

Maintenance logs were first segmented into individual sentences by splitting the text at operator identifiers, which typically mark the completion of an action or the description of a problem. Although some sentences contained both elements, this approach proved effective for structuring heterogeneous and inconsistently reported data. The text was then normalized through standard preprocessing operations, including lowercasing, removal of identifiers, and expansion of abbreviations and domain-specific terminology using a custom vocabulary validated by an industry domain expert.

The logs were translated from Danish into English using the open-source *Helsinki-NLP/opus-mt-da-en* model, as most NLP models are primarily trained on English corpora [2]. Translation was performed locally to ensure data confidentiality.

A supervised classifier was then used to categorize sentences as problem or action. The classifier was initially trained on the *Aircraft Historical Maintenance Dataset* (MaintNet), which provides explicitly labeled maintenance records in an industrial context [3]. To address domain differences, it was then finetuned on an annotated subset of 200 company-specific sentences, previously labeled by a domain expert and selected to reflect recurring issues and actions. Misclassified sentences were manually corrected and incorporated into a retraining phase. To limit bias and overfitting, only this annotated subset was combined with the original dataset during retraining, assigning higher weights to company-specific instances to prioritize domain-relevant patterns while preserving general maintenance knowledge.

Subsequently, sentences classified as problems and actions were clustered separately to identify recurring families. HDBSCAN was adopted to capture clusters of varying size and treat non-representative sentences as noise, making it suitable for heterogeneous text data.

Next, representative keywords were extracted using a term-weighting scheme (c-TF-IDF) designed to highlight words that are distinctive of each cluster. These keywords were then used to generate concise semantic labels with the support of a locally deployed large language model, ensuring interpretability while preserving data confidentiality.

Finally, the identified problem and action clusters enabled the organization of maintenance knowledge originally embedded in unstructured textual records, providing a structured basis for subsequent analysis.

3. Results

The proposed approach was applied to a dataset of 18 107 maintenance logs collected from a Float-Zone silicon crystal growth process. In the classification task, domain finetuning improved precision, i.e., the proportion of correctly classified sentences among all sentences assigned to a given class, from 78% to 80% for problem sentences and from 83% to 87% for action sentences. Clustering of the classified data resulted in 19 problem clusters and 21 action clusters, capturing the diversity of recurring issues and corrective interventions.

Figure 1 and Figure 2 depict clustering results in a two-dimensional space, referred as *Problem Map* and *Action Map*.

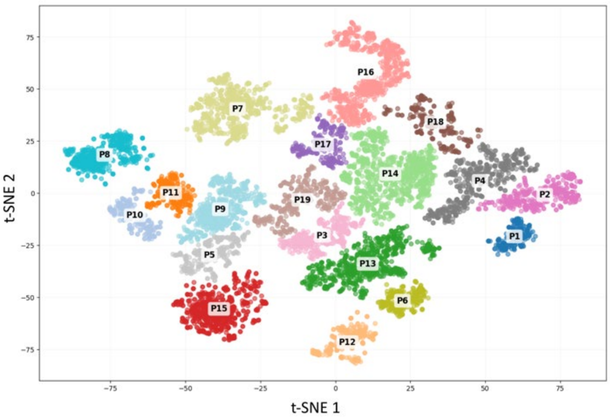

Figure 1. Clusters of problems extracted from logs in the semantic space.

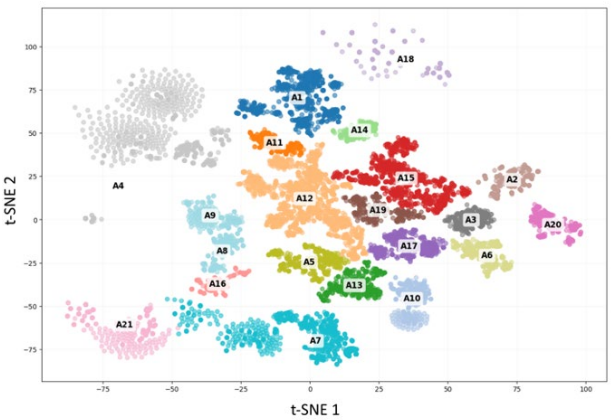

Figure 2. Clusters of actions extracted from logs in the semantic space.

In Figure 1, clusters span horizontally from component-specific to machine-level issues, reflecting fault severity, and vertically from production-related to production-independent problems, indicating production dependence. In Figure 2, clusters range horizontally from generic interventions to highly specific corrective actions, reflecting action specificity, and vertically from software-related activities to hardware-related interventions, indicating the action domain. Cluster quality was assessed through the silhouette score, yielding 0.416 for problem clusters and 0.5066 for action clusters. These values are consistent with moderately separated clusters and reflect a certain degree of overlap, which is expected given the heterogeneous and ambiguous nature of maintenance logs.

Figure 3 presents the correlation between problem and action clusters, illustrating how frequently specific problems are associated with specific corrective actions.

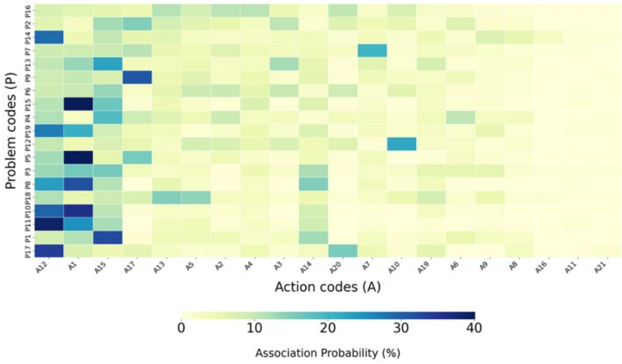

Figure 3. Correlation map linking clusters of problem and action.

The correlation shows that critical problems are often addressed through immediate interventions, suggesting a predominantly reactive maintenance strategy. At the same time, some problem-action pairs exhibit strong one-to-one relationships, suggesting the presence of specialized corrective procedures. In contrast, several problem clusters display weak and dispersed correlations across multiple actions, indicating a lack of standardized corrective responses.

Overall, the findings indicate the effectiveness of TM and NLP techniques in structuring maintenance logs and addressing the identified data-related limitations

4. Conclusion

This work discusses how TM and NLP techniques can be applied to transform unstructured maintenance logs into structured knowledge in an industrial context. The approach enables performance improvements by identifying problem and action instances, assessing their relevance, and providing insights into maintenance practices through the analysis of relationships between problems and corrective actions. We acknowledge several limitations, including the sentence segmentation strategy, the lack of longitudinal analysis of historical data, and the absence of integration between textual maintenance logs and machine-level data. Future work may explore the use of domain-specific large language models to address these limitations and further enhance information extraction in data-driven maintenance applications.

References

[1] W. Von Ammon, 'FZ and CZ crystal growth: Cost driving factors and new perspectives: FZ and CZ crystal growth', *Phys. Status Solidi A*, vol. 211, no. 11, pp. 2461–2470, Nov. 2014, doi: 10.1002/pssa.201400043.

[2] K. Zhong, T. Jackson, A. West, and G. Cosma, 'Natural Language Processing Approaches in Industrial Maintenance: A Systematic Literature Review', *Procedia Computer Science*, vol. 232, pp. 2082–2097, 2024, doi: 10.1016/j.procs.2024.02.029.

[3] P. Wang, J. Karigiannis, and R. X. Gao, 'Ontology-integrated tuning of large language model for intelligent maintenance', *CIRP Annals*, vol. 73, no. 1, pp. 361–364, 2024, doi: 10.1016/j.cirp.2024.04.012.