1. Introduction

The effective management of maintenance services in the oil and gas industry, particularly for offshore production units, presents substantial logistical and operational challenges. Maintenance routines on platforms involve technically complex operations where consistent failure mode identification and task prioritization are critical for safety and cost efficiency (George et al., 2022). A significant bottleneck in this process is the management of Maintenance Work Orders (MWOs). These documents, often generated by field technicians or operators, contain a mixture of structured data (categorical fields) and unstructured free text describing the demand.

Data quality in MWOs is a well-documented issue in the industry. The unstructured text is often noisy, containing misspellings, nonstandard abbreviations, and industry-specific jargon, which complicates automated analysis (Hoffimann et al., 2018). Technicians often describe the same failure using vastly different vocabularies or syntactic structures, leading to a proliferation of duplicate or semantically similar demands (Stewart et al. 2023). This redundancy forces reliability engineers to manually review thousands of records to identify overlaps, a labor-intensive and subjective task (Stewart et al., 2024).

To address these challenges, recent research has focused on Technical Language Processing (TLP), a subdomain of Artificial Intelligence adapted for industrial texts. While Large Language Models (LLMs) and transformer-based architectures have shown state-of-the-art performance in Natural Language Processing (NLP), they often require domain-specific fine-tuning to effectively handle the idiosyncrasies of maintenance logs (Stewart et al., 2023; Walker et al., 2024; Correa et. al, 2025). Furthermore, relying solely on textual similarity can be insufficient in complex industrial environments. Recent studies suggest that hybrid approaches, which integrate unstructured text with structured data (such as ontologies or metadata), significantly enhance information retrieval performance in industrial contexts (Naqvi et al., 2025).

This study proposes a hybrid approach to identify potential duplicate maintenance demands within Petrobras offshore platforms by measuring the similarity between maintenance work orders, a critical task for reducing redundant maintenance demands and improving planning efficiency. Unlike purely text-based methods, our approach integrates semantic textual similarity with structured categorical features and symbolic representations of equipment (tags). We present the training and evaluation of a CatBoost

model designed to rank potential duplicate MWOs, thereby streamlining the maintenance planning workflow and reducing the cognitive load on engineering teams.

The proposed system supports maintenance planners and reliability engineers by retrieving previously registered MWOs that are similar to a new maintenance note. Its purpose is to reduce duplicate demands, improve the standardization and interpretation of technical documentation, and increase the reuse of historical maintenance knowledge. In offshore maintenance contexts, this can reduce manual review effort and support more consistent planning decisions.

2. Methodology

This section describes the methodological framework used to develop and evaluate the proposed duplicate detection system for maintenance work orders. The approach follows a hybrid pipeline that combines semantic textual representations with structured maintenance metadata to rank potentially duplicate records. First, we describe the CatBoost-based ranking model and its training configuration. Next, we introduce the evaluation metrics used to assess ranking performance and interpret the experimental results. We then describe the dataset and the procedure used to construct positive and negative pairs of maintenance notes. Finally, we present the feature engineering process, including semantic embeddings and structured similarity signals derived from maintenance metadata.

a. Model Architecture and Training

The proposed approach employs CatBoost (Prokhorenkova et al., 2017), a gradient-boosted decision tree classifier, trained as a cost-sensitive binary model using the standard log-loss objective. To address class imbalance and reflect the asymmetric cost of missed duplicate detections in maintenance management workflows, a higher weight was assigned to the positive class. The class weight was fixed at 20, based on empirical analysis and discussions with domain experts, aiming to align model behavior with practical operational requirements.

Model training was monitored using the Area Under the ROC Curve (AUC), a threshold-independent metric suitable for evaluating score-based models. Although the model was trained as a binary classifier, its predicted scores were sorted and subsequently used to rank candidate maintenance records for each incoming demand. This design allows the system to function effectively as a human–AI decision-support tool, where ranking quality is more critical than binary classification accuracy.

b. Evaluation Metrics

While the model can function as a binary classifier, its primary operational utility is to rank existing notes by similarity to a new incoming demand. Therefore, the evaluation focused on ranking metrics rather than solely on classification accuracy.

3. Mean Reciprocal Rank (MRR): Used to evaluate the average rank of the correct duplicate in the list of suggestions. An MRR closer to 1.0 indicates that the true duplicate frequently appears at the very top of the list.
4. Hit@K: This metric measures the proportion of queries where the true duplicate appears within the top K results. We evaluated Hit@1, Hit@3, and Hit@5 to assess the model's utility as a recommendation system for reliability engineers.

c. Data Description and Preparation

The dataset used in this study consists of a large corpus of maintenance notes (MWOs) collected from Petrobras offshore platforms. These notes primarily detail equipment failures requiring exchange or maintenance. From this corpus, we constructed a large-scale dataset of note pairs used for training and validation.

The data preparation phase involved creating positive labeled pairs and negative pairs (dissimilar demands) to train the model to distinguish semantic and operational similarity. Negative pairs are therefore required to provide explicit counterexamples during training, enabling the model to learn a meaningful decision boundary in the feature space. This formulation aligns with pairwise learning and ranking-based approaches, in which discrimination emerges from the relative comparison between positive and negative instances rather than from absolute classification alone. Consistent with challenges identified in the literature, the raw data includes both unstructured text fields (free descriptions of the problem) and structured categorical fields (e.g., work center, planning area), which are essential for contextualizing the demand.

From the original corpus, a subset of duplicate MWOs were identified through an automated weak-labeling process supported by Large Language Models (LLMs). Specifically, LLMs were used to analyze the unstructured textual descriptions and detect explicit references indicating that a given maintenance demand was a duplication of another previously registered note. These references commonly appeared as textual mentions of related work order identifiers or statements signaling repeated or overlapping demands. The

resulting pairs were subsequently reviewed and consolidated to form the set of positive examples used in the supervised learning task.

To construct the training dataset, we adopted a negative sampling strategy constrained at the offshore platform level. For each positive (duplicate) pair, one of the notes was selected as an anchor, and negative pairs were generated by pairing this anchor with randomly sampled MWOs from the same platform. The maximum number of negative samples per positive pair was limited by the total number of available MWOs on each platform.

To investigate the effect of the number of negative samples per positive pair, we conducted a controlled experiment varying the number of randomly sampled negative pairs per anchor note. The evaluated configurations ranged from small sets of negatives to large candidate pools that approximate the operational retrieval scenario.

Each model was trained under identical conditions and evaluated on a fixed test set containing a large pool of randomly sampled negative candidates per query. The train–validation–test split was performed at the query identifier level to avoid data leakage.

Contrary to expectations suggested in ranking literature, increasing the number of negative samples did not lead to substantial improvements in generalization performance. As shown in Figure 1, the Mean Reciprocal Rank (MRR) stabilizes once the model observes a sufficiently large and diverse set of negative examples. Similarly, Hit@1, Hit@3, and Hit@5 metrics show only marginal variation beyond this point.

This behavior suggests that performance saturation occurs once the model has observed sufficient negative variability to establish a stable decision boundary. Additional randomly sampled negatives appear to contribute limited new discriminative information, particularly given that most random pairs represent relatively easy negatives.

Based on these findings, we selected a large but computationally efficient number of negative samples per positive pair for the final model. This configuration balances computational efficiency with performance stability while approximating the scale of candidate work orders typically observed in the operational scenario.

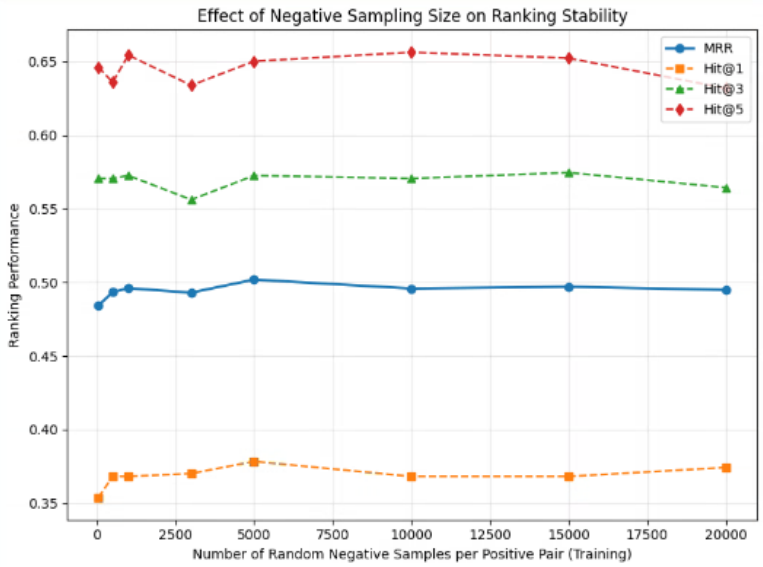

Figure 1: Impact of the number of randomly sampled negative pairs per positive instance (training set) on ranking metrics evaluated on a fixed large test set.

d. Feature Engineering

To capture the multidimensional nature of maintenance similarity, we engineered a hybrid feature set that combines semantic understanding from text and matching of categorical metadata. This aligns with methodologies that advocate for multi-modal learning to improve retrieval in unstructured industrial records (NAQVI et al., 2025). The features input into the model are as follows:

1. Semantic Textual Similarity: We used the *paraphrase-multilingual-minilm-l12-v2* Sentence-Transformers model, based on the BERT architecture, to convert each note's unstructured text into dense vector representations (embeddings), which, in turn, serve as features for the model. While transformer models offer baselines, fine-tuning on domain-specific data is crucial to effectively capture the complex lexical patterns and jargon found in industrial maintenance records (NAQVI et al., 2025). The necessity of this process was evidenced by the performance of the off-the-shelf model, which yielded a Mean Reciprocal Rank at 100 (MRR@100) of only 0.2. Consequently, the model was fine-tuned specifically on the domain's maintenance lexicon using pairs of known duplicate work orders. Given the complexity of the dataset, which included hard positive and hard negative examples, we employed Online Contrastive Loss during training to improve the model's discriminative capability. This fine-tuning process significantly enhanced performance, raising the MRR@100 to 0.37.

2. Tag Overlap (Symbolic/Semantic): A "tag" represents a specific equipment identifier (e.g., PV-23432). We extracted tags from the textual descriptions and created a boolean feature indicating whether the text of the candidate note contains a tag present in the reference note. This serves as an indicator that the demands refer to the same physical asset.
3. Installation Location Overlap: In offshore maintenance systems, each work order is associated with a hierarchical installation location code that represents the physical positioning of the equipment within the platform. This location is structured as a multi-level path, where each level provides increasingly specific spatial information. The levels are concatenated and separated by a dot symbol, for example:

   AAAA.BBBB.CCCC.DD.EE.FF

   where each segment corresponds to a distinct hierarchical level of the installation structure.

   To quantify the spatial proximity between two maintenance demands, we define an installation location overlap score based on the number of shared hierarchical levels between their respective location codes. Only consecutive matching levels starting from the root of the hierarchy are considered (i.e., a *prefix match*), ensuring that the overlap reflects true structural proximity rather than coincidental matches at deeper levels. The final overlap score, denoted as □, is computed as a normalized measure that accounts for potential differences in hierarchy depth between notes. Let □ denote the number of consecutive intersecting levels, and let $□_1$ and $□_2$ represent the total number of installation location levels of each maintenance note. The score is defined as:

   $$□ = \frac{□(□_1 + □_2)}{□_1□_2}$$

   This formulation assigns higher values to note pairs that share deeper hierarchical prefixes, indicating closer physical proximity, while remaining robust to cases where one note contains a more detailed location specification than the other.

4. Categorical Matching: Four boolean features were generated to check for exact matches in structured fields:
   a. Same Work Center: Indicates if the maintenance teams assigned are identical;

b. Same Equipment Class: Indicates if the equipment class is the same in both notes;
c. Same Equipment Number: Indicates if the standardized equipment denominations match;
d. Same Location Classification: Indicates if the equipment location scope (broad or restricted) is the same in both notes.

e. Implementation Timeline and Deployment Maturity

The proposed system was developed through an incremental implementation process. The first operational prototype, based on fine-tuned semantic embeddings and cosine similarity, was completed in approximately two months, including problem definition, data preprocessing, weak labeling, and embedding fine-tuning. A second development cycle of approximately two to three months incorporated structured maintenance metadata, symbolic equipment information, engineered similarity features, and a CatBoost-based supervised ranking model. Finally, approximately one additional month was dedicated to experiments on negative sampling size, ranking evaluation, feature importance analysis, and computational performance assessment. Overall, the first operational version was achieved in about two months, while the current hybrid CatBoost-based version required approximately five to six months from conception to validation.

5. Results

This section presents the results of the proposed hybrid similarity model: (a) feature importance analysis; (b) evaluation of ranking performance using MRR and Hit@K metrics; and (c) assessment of the system's computational efficiency for real-time operational use.

a. Feature Importance

As shown in Figure 2, feature importance analysis shows that **S**emantic Textual Similarity is the dominant predictor (≈ 46%), confirming that the fine-tuned embeddings provide the main discriminative signal for identifying duplicate MWOs.

Installation Location Overlap (≈ 27%) and Tag Overlap (≈ 11%) also contribute substantially, indicating that spatial proximity and shared equipment identifiers are strong complementary

signals. Among categorical features, Work Center (≈ 9%) has moderate relevance, while Location Classification, Equipment Class, and Equipment Number present limited impact.

Overall, the results (Figure 2) validate the hybrid design: textual semantics drives similarity detection, while structured and symbolic features enhance contextual robustness.

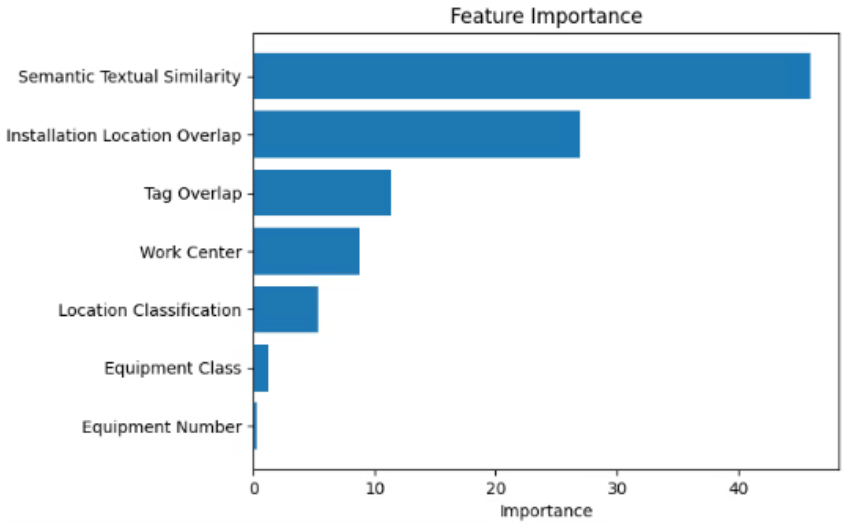

Figure 2: Feature importance scores computed by the CatBoost model based on split contribution.

b. Ranking Capability

The model demonstrates high efficacy in prioritizing duplicates. The Mean Reciprocal Rank (MRR) achieved was 0.73, indicating that, on average, the model positions the correct duplicate among the very first positions in the ranking. Consequently, the user would typically identify the duplicate note immediately upon reviewing the suggestions.

The Hit@K metrics further validate this performance on the test data:

- Hit@1 = 0.63
- Hit@3 = 0.80
- Hit@5 = 0.84

These results show that in 84% of cases, the duplicate maintenance note appears within the top 5 suggestions presented by the model. This high retrieval rate is crucial for reducing

the "unforeseen" losses in operations by ensuring maintenance planners have visibility of existing demands (UTNE et al., 2012).

c. Operational Efficiency

Regarding computational performance, the model is highly efficient for real-time applications. Currently, the system requires approximately 5 seconds to process 10 new notes, allowing for near-instantaneous feedback during the MWO creation or review process.

6. Discussion

The presented model can be deployed to rank candidate duplicates for posterior human review. Therefore, the model's ranking capability represents its central contribution. MRR result indicates that the correct duplicate is typically placed among the top-ranked suggestions, while a Hit@5 shows that, in 84% of cases, the relevant duplicate appears within the top five candidates. In comparison to the first-phase duplicate identifier based solely on cosine similarity between notes, which achieved an MRR of 0.38 and a Hit@5 of 0.43, this represents a substantial improvement. This behavior supports rapid human verification and immediate action, which is especially valuable when duplicate failure notifications can degrade reliability analysis and distort planning decisions. In maintenance management, where the practical objective is often to surface the right historical record quickly, ranking quality is a more appropriate operational criterion because humans are oriented to look for patterns and standards. This view aligns with recent work on maintenance work order retrieval, showing that noisy, domain-specific ticket text can challenge standard retrieval pipelines and that improving maintenance results can materially reduce time-to-resolution by ranking the correct record higher for technicians and decision-makers (Siouffi et al., 2025).

The model design follows the principle that decision-support solutions should replicate core stages of human reasoning in maintenance work, strengthening human–AI integration and ensuring alignment with user needs. This framing is consistent with the building maintenance and operations management literature, where effective collaboration mechanisms rely on humans contributing domain knowledge and specialized tools to guide the system while the model performs tasks to meet human needs (Chen et al., 2025).

Although duplicate detection and semantic retrieval are established tasks, this work contributes by integrating them into a real offshore maintenance workflow through a hybrid ranking system that combines fine-tuned embeddings, structured metadata, tag overlap,

and installation-location similarity. Its innovation lies mainly in the operational adaptation of these techniques to noisy maintenance documentation and human–AI decision support.

Finally, computational efficiency strengthens the feasibility of deployment. Processing approximately ten new notes in around five seconds supports near-real-time feedback during work order creation or review, which is critical for embedding decision support into operational workflows. Continuous deployment and interaction with users are also expected to improve system performance over time as more operational examples become available. Prior studies in maintenance and IT operations similarly highlight the benefits of iterative knowledge accumulation and system refinement in AI-assisted operational environments (Chen et al., 2025; Wang et al., 2024). Therefore, these findings suggest that the next frontier is not only improved ranking or recommendation quality, but the development of operationally grounded, governance-aware human–AI systems capable of generating actionable insights while preserving human control in complex maintenance contexts.

7. Conclusion

This study addressed a critical challenge in offshore maintenance management by proposing a hybrid, data-driven approach to identify and prioritize duplicate Maintenance Work Orders in a highly complex and noisy industrial environment. By integrating semantic textual similarity with structured categorical features and symbolic equipment information, the proposed model effectively supports maintenance planners through ranked recommendations rather than automated decisions. The ranking performance demonstrates the model's practical value in surfacing relevant historical demands quickly, thereby reducing cognitive load and mitigating the risk of redundant or inconsistent planning actions. Furthermore, the results reinforce the suitability of a human–AI collaboration paradigm in which AI augments human reasoning by organizing information and prioritizing alternatives, while final decisions remain under human control. From an operational perspective, the model's computational efficiency enables near-real-time integration into maintenance workflows, supporting scalability in large offshore environments. In particular, this work provides a foundation for improving demand visibility and planning consistency in Petrobras' offshore production units.

Moreover, even in cases where the highest-ranked records are not exact duplicates, the retrieved notes often correspond to operationally related maintenance demands. Such similarities may reveal opportunities for coordinated interventions, shared diagnostics, or combined maintenance activities on the same equipment or subsystem. Therefore, beyond

duplicate detection, the proposed system may also support broader knowledge reuse and operational planning by highlighting historically similar maintenance situations.

Although the proposed approach was developed in a challenging operational context, characterized by noisy maintenance notes, inconsistent terminology, spelling errors, nonstandard abbreviations, and limited standardization, The results indicate that the hybrid model achieved a MRR of 0.73. These characteristics are inherent to real offshore maintenance environments and reinforce the relevance of methods capable of combining semantic representations with structured and symbolic information. Rather than limiting the applicability of the proposed approach, these challenges highlight the importance of robust human–AI decision-support systems that can operate under imperfect data conditions. Future work should focus on further improving the system through systematic labeling practices, continued human-in-the-loop validation, and the expansion of curated examples across different failure types, equipment classes, and operational contexts. In addition, future versions may explore new modeling techniques and incorporate richer contextual information, such as summarized maintenance-note descriptions, equipment type, maintenance action, failure mode, subsystem information, and other operational attributes. These enhancements may improve ranking robustness, support broader knowledge reuse, and further strengthen the system's value for maintenance planning and reliability analysis.

Acknowledges

The authors gratefully acknowledge the Alexander von Humboldt Foundation for the fellowship awarded, as well as PETROBRAS, EVCOMX, and the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) for their support of this work, carried out under the EVCOMX–PETROBRAS Cooperation Agreement and funded through resources from ANP's Research, Development and Innovation (RD&I) Clause.

References

Chen, S., Liang, X., Liu, Y., Li, X., Jin, X., & Du, Z. (2025). Customized large-scale model for human-AI collaborative operation and maintenance management of building energy systems. *Applied Energy, 393*, 126169. https://doi.org/10.1016/j.apenergy.2025.126169

Correa, O. C. et al. Automated Classification of Offshore Production Units Maintenance Notes Using Machine Learning and Large Language Models. In: OFFSHORE TECHNOLOGY CONFERENCE BRASIL, 2025, Rio de Janeiro. Proceedings [...] Richardson, TX: Offshore Technology Conference, 2025. https://doi.org/10.4043/36196-MS

George, B., Loo, J., & Jie, W. (2022). Recent advances and future trends on maintenance strategies and optimisation solution techniques for offshore sector. *Ocean Engineering, 250*, 110986. https://doi.org/10.1016/j.oceaneng.2022.110986

Hoffimann, J., Mao, Y., Avinash, W., & Taylor, A. (2018, September). *Sequence mining and pattern analysis in drilling reports with deep natural language processing*. Paper presented at the SPE Annual Technical Conference and Exhibition, Dallas, TX, United States. https://doi.org/10.2118/191505-MS

Naqvi, S. M. R., et al. (2025). Enhancing semantic search using ontologies: A hybrid information retrieval approach for industrial text. *Journal of Industrial Information Integration, 45*, 100835. https://doi.org/10.1016/j.jii.2025.100835

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). *CatBoost: Unbiased boosting with categorical features*. arXiv. https://doi.org/10.48550/arXiv.1706.09516

Siouffi, C., Glandon, P., Kubler, S., & Soumianarayanan, V. (2025). LLM-based contrastive representation learning for enhanced maintenance work order retrieval. *Procedia CIRP, 134*, 981–986. https://doi.org/10.1016/j.procir.2025.02.230

Stewart, M., Hodkiewicz, M., & Li, S. (2023). *Large language models for failure mode classification: An investigation*. arXiv. https://arxiv.org/abs/2309.08181

Stewart, M., et al. (2024). MWO2KG and Echidna: Constructing and exploring knowledge graphs from maintenance data. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 238*(5), 920–932. https://doi.org/10.1177/1748006X221131128

Utne, I., et al. (2012). Shutdown preparedness in oil and gas production. *Journal of Quality in Maintenance Engineering*, *18*(2), 154–170. https://doi.org/10.1108/13552511211244193

Walker, C., et al. (2024). Using large language models to recommend repair actions for offshore wind maintenance. *Journal of Physics: Conference Series*, *2875*, 012025. https://doi.org/10.1088/1742-6596/2875/1/012025

Wang, W.-J., Lain, W.-D., & Wang, Y.-S. (2024). Empowering IT maintenance with human-AI collaboration for cloud native. In *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*. IEEE. https://doi.org/10.1109/ICEIB61477.2024.10602593

.