Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27823-cd

UNLOCKING MAINTENANCE DARK DATA WITH EMBEDDING-BASED SEMANTIC SEARCH

Melwin Xavier

Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, Sweden. E-mail: melwin.xavier@ltu.se

Ramin Karim

Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, Sweden. E-mail: ramin.karim@ltu.se

Textual maintenance logs from mining operations provide information relevant to predictive maintenance, but Swedish free-text narratives make automated classification difficult. Existing methods are limited by domain-specific terminology and multilingual contexts. A hybrid semantic lexical search system is presented that maps free-text maintenance descriptions to a standardized taxonomy of 72 categories by combining contrastively fine-tuned multilingual embeddings (GTE) with TF-IDF lexical matching. The system uses weighted score fusion (α=0.90) to integrate semantic similarity with exact term evidence. Results on 919 real-world Swedish maintenance logs from mining drill rigs show that the hybrid approach achieves 45.9% Recall@1, 75.4% Recall@5, and 0.585 MRR on real-world data, substantially outperforming semantic-only (73.1% $\rightarrow$ 45.9% validation-to-real drop) and lexical-only baselines (52.7% R@1 on validation). Per-category analysis shows strong performance for fire safety systems (100% R@1) while indicating persistent challenges for motor failures (0% R@1).

Keywords: text embeddings, taxonomy classification, predictive maintenance.

1. Introduction

Large mining operations generate thousands of maintenance work orders each day documenting equipment failures, repairs, and operational issues. These records support predictive maintenance strategies intended to reduce downtime and improve maintenance scheduling Dalzochio et al. (2020); Sharma et al. (2011). In Swedish underground mines, modern drill rigs produce Swedish-language maintenance descriptions authored by operators who use inconsistent terminology and variable levels of technical detail. Mapping such unstructured entries to standardized taxonomies enables systematic failure analysis, trend detection, and data-driven maintenance planning.

Manual taxonomy assignment is labor intensive and cannot scale to the volumes produced in industrial environments. Conventional keyword rules are brittle and fail to capture semantic variation and synonymy common in operator-written text.Multilingual dense retrieval with learned embeddings has improved cross-lingual semantic search Conneau et al. (2020); Reimers and Gurevych (2019), but embedding-only methods may overlook critical exact-match terms (e.g., equipment codes, part numbers). In contrast, lexical methods such as TF-IDF and BM25 Robertson et al. (2009) handle exact term overlap effectively yet often underperform on paraphrases and semantic equivalence.

The present study addresses this gap by developing a hybrid semantic lexical system that combines the strengths of both retrieval paradigms. State-of-the-art multilingual embedding models are fine-tuned on synthetic Swedish maintenance data, and semantic similarity scores are fused with TF-IDF lexical scores using a weighted combination optimized through systematic α tuning. The resulting system is evaluated on real-world maintenance logs from a Swedish mining operation.

Industrial maintenance narratives differ from general-purpose text classification tasks. Operators frequently use domain-specific abbreviations (e.g., “uh” for equipment hours, “hydr” for hy-

draulics), mix Swedish and English terminology, and describe multi-component failures with uneven technical detail. Real-world labeled data are scarce because expert annotation is expensive and maintenance records are proprietary. These constraints motivate a workflow that relies on synthetic data generation for training and systematic validation on authentic operator logs to quantify domain transfer.

The main contributions are: (i) a hybrid retrieval architecture that combines fine-tuned multilingual embeddings with TF-IDF for Swedish industrial text, (ii) a systematic baseline comparison of 8 embedding models followed by contrastive fine-tuning, (iii) comprehensive α parameter tuning (21 values, 2 lexical methods, 3 models) to optimize score fusion, and (iv) real-world evaluation on 919 mining maintenance logs showing 45.9% Recall@1 with hybrid fusion compared to 52.7% for TF-IDF-only on validation.

2. Related Work

2.1. *Retrieval Methods*

Dense retrieval. Queries and documents are encoded as vectors in a shared continuous space and ranked using vector similarity. Contrastive learning has adapted BERT-based encoders for passage retrieval Karpukhin et al. (2020). Sentence-BERT Reimers and Gurevych (2019) introduced siamese architectures that enable efficient sentence embedding. For multilingual retrieval, XLM-R Conneau et al. (2020) provides cross-lingual representations, while E5 Wang et al. (2022) trains general-purpose text embeddings through multi-stage contrastive learning.

Lexical matching. TF-IDF remains a competitive baseline for text retrieval by weighting terms using frequency and inverse document frequency. BM25 Robertson et al. (2009) extends TF-IDF with document length normalization and is often effective for technical terminology. TF-IDF is used in this study due to its strong performance on Swedish industrial text with limited training data.

Hybrid search. Combining dense and sparse methods is motivated by the complementary failure modes of semantic matching and exact term overlap. Toolkits such as Pyserini facilitate reproducible research across sparse and dense representations Lin et al. (2021). In ColBERT Khattab and Zaharia (2020), contextualized embeddings interact late during scoring. Score-fusion approaches integrate semantic and lexical evidence using weighted combinations or learned reranking. The present work uses weighted fusion with systematic α tuning optimized for Swedish maintenance text.

2.2. *Maintenance Text Mining and Swedish NLP*

NLP applications for industrial maintenance logs have grown into an active research area Wang et al. (2025); Dalzochio et al. (2020). Predictive maintenance analytics demonstrate the operational value of extracting structured knowledge from unstructured logs Wang et al. (2025); Dalzochio et al. (2020). Taxonomies can be treated as structured knowledge representations and are often expressed as graphs, for which learning methods on knowledge graphs provide relevant foundations Ye et al. (2022).

Although Swedish is a well-resourced European language, industrial Swedish NLP remains challenging. Swedish BERT models Malmsten et al. (2020) perform well on general text but typically require domain adaptation for technical terminology. Mining maintenance logs exhibit code-switching between Swedish and English equipment terms, non-standard abbreviations, and compound words characteristic of Swedish morphology. Multilingual models trained on diverse corpora Conneau et al. (2020); Wang et al. (2022) offer an alternative to Swedish-only models by enabling cross-lingual transfer from English technical documentation while retaining Swedish language coverage. This study evaluates whether fine-tuning multilingual embeddings on synthetic Swedish maintenance data can bridge this gap for real-world classification tasks.

3. Methodology

The input consists of a Swedish free-text maintenance description q and a taxonomy of $N = 72$ categories; the objective is to retrieve the most relevant category. Each category t_i is represented

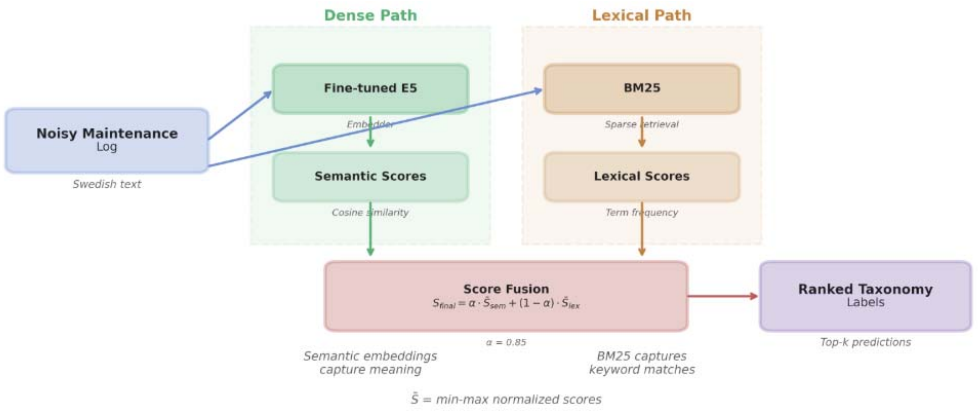

Fig. 1. Hybrid search architecture combining fine-tuned GTE embeddings and TF-IDF lexical matching via weighted score fusion.

by multiple text descriptions $\{d_{i,1}, d_{i,2}, \ldots, d_{i,k_i}\}$ extracted from a structured taxonomy file containing Problem, Cause (object), Cause, and Description fields (207 total descriptions). The retrieval system ranks categories by their estimated relevance to the query.

Each row in the taxonomy reference is treated as a candidate retrieval passage d. Category labels are encoded as the composite key *Problem; Cause (causing object); Cause*, yielding 72 unique taxonomy labels. To form passage text, the three label fields are concatenated with the human-written Description field. Descriptions are lightly cleaned (whitespace trimming, removal of leading punctuation, and collapsing repeated spaces), and duplicate passages are removed to avoid redundant candidates.

Text preprocessing preserve industrial abbreviations, part numbers, and mixed Swedish/English terminology, no stop-word removal or stemming is applied. For lexical retrieval (BM25 and TF-IDF), tokenization uses a regex that retains alphanumeric sequences and Swedish characters (ÅÄÖåäö), and all tokens are lowercased. Real-world maintenance logs are scarce and often unlabeled.

To enable model development, 3,552 synthetic Swedish maintenance descriptions are generated using two complementary approaches: Python-based generation creates 2,500 samples by programmatically combining taxonomy elements with Swedish maintenance vocabulary templates and introducing realistic variation in formatting, abbreviations, and terminology. LLM-based augmentation produces 1,052 samples by prompting large language models to generate naturalistic operator-written descriptions aligned with taxonomy categories, thereby capturing colloquial expressions and semantic paraphrases. An 80/20 split yields train (2,556) and validation (711) sets for model comparison and fine-tuning.

Evaluation is conducted on 919 real-world Swedish maintenance logs from underground mining drill rigs, manually annotated with ground-truth taxonomy labels. The dataset spans diverse failure types, including hydraulic systems, electrical faults, drilling components, control systems, chassis issues, and operator errors.

3.1. *Model Selection and Fine-Tuning*

Baseline comparison. Eight candidate retrieval models are evaluated on the validation set (Table 1): 5 semantic embedding models (BGE-M3, GTE-Multilingual-Base, E5-Multilingual-Large, LaBSE, Paraphrase-MPNet), 1 Swedish model (KB-BERT), and 2 lexical baselines (BM25, TF-IDF). BGE-M3 provides the best semantic baseline (57.9% Recall@1, 67.9% MRR), with GTE following (54.6% R@1, 65.8% MRR). Despite being Swedish-specific, KB-BERT performs poorly (21.5% R@1), indicating a need for domain adaptation.

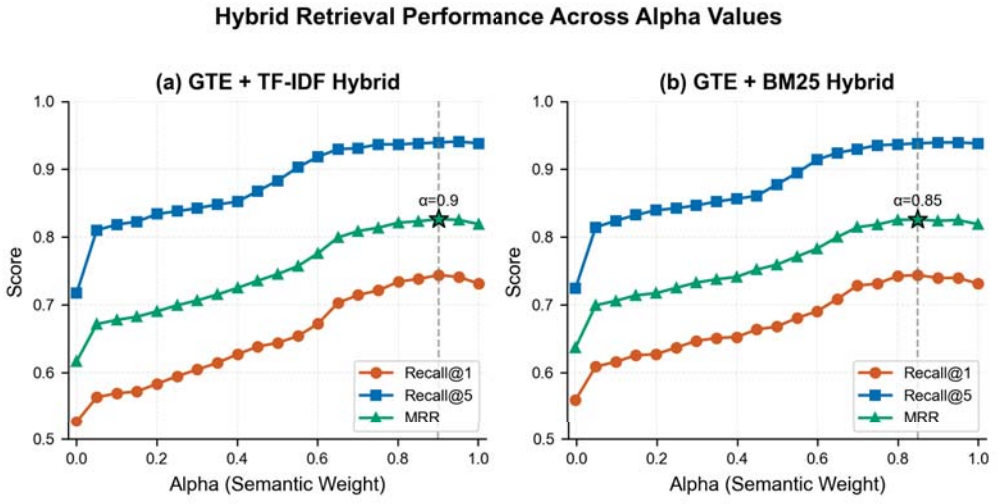

Fig. 2. Alpha parameter tuning curves for GTE-V1.

Table 1. Baseline model comparison on the validation set (n=711); top 6 results shown.

Model	R@1	R@5	MRR
BGE-M3	57.9	79.0	67.9
BM25	56.0	72.4	63.7
GTE	54.6	78.8	65.8
E5-Large	54.0	76.1	64.4
TF-IDF	52.7	72.2	61.6
LaBSE	45.0	73.6	57.7

Fine-tuning. The top 3 semantic models (BGE-M3, GTE, E5-Large) are fine-tuned using MultipleNegativesRankingLoss on the synthetic training data (2,556 query-positive pairs). Fine-tuning uses 2 epochs, batch size 4, learning rate 2×10^{-5}, the AdaFactor optimizer, max sequence length 64, and 4 unfrozen layers. E5 fine-tuning includes the required "query:" and "passage:" prefixes. **Pair construction.** Each synthetic query is paired with a positive taxonomy passage sampled from the set of passages associated with its ground-truth label. This sampling exposes the model to multiple realizations of each category and reduces overfitting to a single canonical description. **Contrastive objective.** With a batch of B query passage pairs, MultipleNegativesRankingLoss treats the other $B-1$ passages in the batch as negatives and optimizes a softmax over cosine similarities:

Table 2. Fine-tuning results on the validation set (n=711).

Model	R@1	R@3	R@5	MRR
BGE-V1	69.7	88.4	95.8	80.3
E5-V1	68.3	87.7	94.4	79.4
GTE-V1	66.9	88.0	94.4	78.8

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(s(q_i, p_i)\right)}{\sum_{j=1}^{B} \exp\left(s(q_i, p_j)\right)} \qquad (1)$$

where $s(\cdot, \cdot)$ denotes cosine similarity between L2-normalized embeddings. As shown in Table 2, performance improves substantially: BGE-M3 57.9% $\rightarrow$ 69.7% (+11.8%), E5-Large 54.0% $\rightarrow$ 68.3% (+14.3%), and GTE 54.6% $\rightarrow$ 66.9% (+12.3%). Each fine-tuned model attains >93% Recall@5 on the validation split.

3.2. *Hybrid Retrieval and Score Fusion*

Figure 1 depicts the hybrid system and its three components:

Semantic component: Semantic similarity between a query q and a taxonomy document d is computed using fine-tuned GTE embeddings:

$$s_{\text{sem}}(q, d) = \cos(\mathbf{e}_q, \mathbf{e}_d) \qquad (2)$$

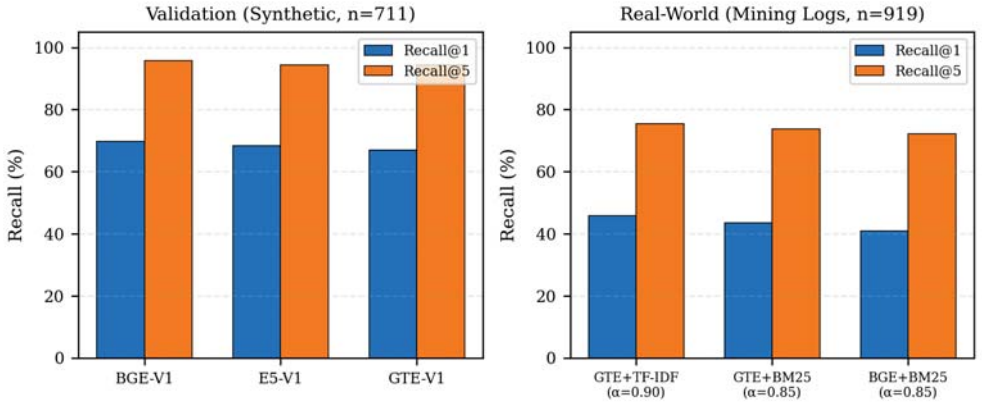

Fig. 3. Comparison of validation (synthetic) and real-world performance, illustrating domain shift.

where $\mathbf{e}_q$ and $\mathbf{e}_d$ denote the L2-normalized embeddings of q and d. Since embeddings are normalized, cosine similarity is computed efficiently as a dot product between the query vector and the precomputed taxonomy embedding matrix.

Lexical component: Lexical relevance is computed on the same taxonomy passages using two alternatives. BM25 uses standard Okapi scoring (k_1=1.5, b=0.75) on regex-tokenized lowercased terms. TF-IDF uses unigram features with L2 normalization (scikit-learn `TfidfVectorizer` with a custom tokenizer); relevance is computed as the query document dot product in TF-IDF space.

Score fusion: For each query, semantic and lexical scores are min max normalized and interpolated:

$$s_{\text{hybrid}}(q,d) = \alpha \cdot \hat{s}_{\text{sem}}(q,d) + (1-\alpha) \cdot \hat{s}_{\text{lex}}(q,d) \tag{3}$$

The parameter $\alpha \in [0,1]$ controls the semantic lexical balance, and category scores are taken as the maximum document score within each category. Min max normalization is performed per query across all taxonomy passages; if score ranges collapse (all passages receive the same score), normalized values are set to zero to avoid degenerate scaling.

Systematic α tuning is performed on the validation set across 21 values ($\alpha \in \{0.0, 0.05, 0.10, \ldots, 1.0\}$), 2 lexical methods (BM25, TF-IDF), and 3 fine-tuned models (BGE-V1, GTE-V1, E5-V1), totaling 126 configurations. Tuning curves for GTE-V1 are reported in Figure 2. The best configuration is GTE-V1 + TF-IDF at α=0.90 (74.4% R@1, 82.6% MRR on validation), indicating a strong emphasis on semantic matching while retaining lexical signal. At α=1.0 (semantic only), Recall@1 is 73.1%; at α=0.0 (lexical only), Recall@1 is 52.7%.

4. Experiments and Results

4.1. *Experimental Setup*

Metrics: Recall@K ($K \in \{1,3,5,10\}$), Mean Reciprocal Rank (MRR), and Accuracy (equivalent to Recall@1).

4.2. *Results*

Table 3 reports results on the real-world maintenance logs. The best-performing configuration (GTE-V1 + TF-IDF, α=0.90) achieves 45.9% Recall@1, 75.4% Recall@5, and 0.585 MRR. GTE-V1 + BM25 (α=0.85) produces comparable performance (43.7% R@1, 73.7% R@5), while BGE-V1 + BM25 reaches 41.0% R@1. The gap between validation (74.4% hybrid, 66.9% semantic-only) and real-world (45.9%) performance indicates domain shift from synthetic to authentic operator text.

Figure 3 contrasts validation and real-world performance. Fine-tuned models achieve 66.9 69.7% Recall@1 on validation but drop to 41.0

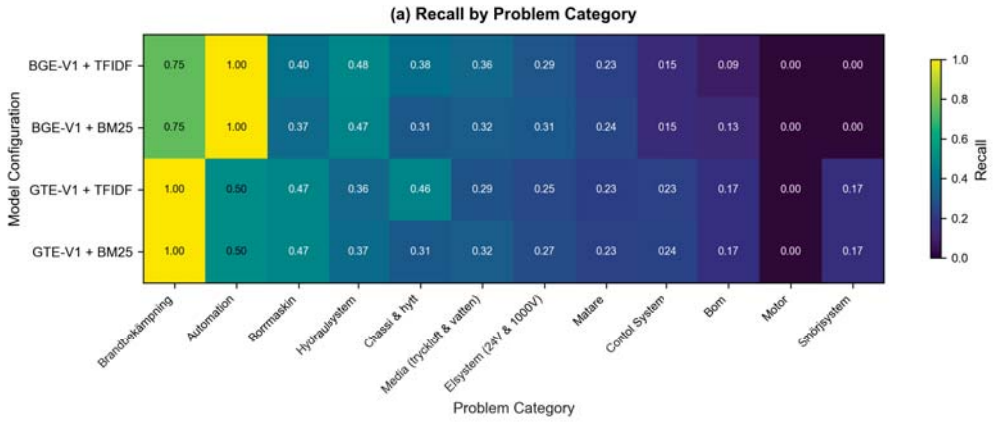

Fig. 4. Per-category Recall@1 on real-world data (GTE-V1+TF-IDF, α=0.90), highlighting variation across failure types.

Table 3. Real-world results on the annotated set (n=919 queries).

Model	Lex.	α	R@1	R@3	R@5	MRR
GTE-V1	**TF-IDF**	**0.90**	**45.9**	**66.6**	**75.4**	**0.585**
GTE-V1	BM25	0.85	43.7	63.0	73.7	0.568
BGE-V1	BM25	0.85	41.0	62.6	72.1	0.545

45.9% on real-world data, highlighting the difficulty of synthetic-to-real transfer. Hybrid fusion nonetheless outperforms single-modality approaches on both datasets.

Table 4 and Figure 4 summarize per-category performance for the best model (GTE-V1 + TF-IDF, α=0.90).

Fire safety systems (Brandbekämpning) achieve perfect classification (100% R@1), plausibly due to distinctive terminology. Motor-related issues yield 0% R@1, suggesting sparse training coverage or ambiguous descriptions. Despite high frequency, drilling machines (Borrmaskin: 47.2%) and chassis systems (46.2%) exhibit only moderate performance. Control system and operator handling errors remain difficult because terminology is inconsistent and symptoms overlap across categories.

5. Discussion

The hybrid approach shows clear advantages over single-modality baselines. On validation data, hybrid fusion (74.4% R@1) outperforms semantic-only (66.9%) and lexical-only (52.7%), indicating complementary strengths. Semantic embeddings capture paraphrases (e.g., “läcker olja” vs “oljeläckage”), while TF-IDF reinforces explicit technical terms in the scoring.

Table 4. Per-category Recall@1 on real-world data for GTE-V1+TF-IDF (α=0.90).

Category	R@1 (%)
Brandbekämpning (Fire safety)	100.0
Automation	50.0
Borrmaskin (Drill machine)	47.2
Chassi & hytt (Chassis)	46.2
Hydraulsystem (Hydraulic)	35.9
Media (Air & water)	29.0
Elsystem (Electrical)	25.3
Control System	22.8
Matare (Feeder)	22.5
Bom (Boom)	17.4
Smörjsystem (Lubrication)	16.7
Motor (Engine)	0.0

The optimal α=0.90 indicates that semantic matching dominates, while lexical matching provides corrections for technical terms and equipment codes. The degradation from validation (66.9 74.4%) to real-world (45.9%) performance underscores domain shift. Operator-written descriptions contain typos, abbreviations, and context-dependent expressions that are underrepresented in synthetic data.

Zero or low Recall@1 for Motor, Smörjsystem, and Bom suggests data imbalance, as these categories represent $<5\%$ of real-world examples. The perfect score for fire safety reflects distinctive terminology as well as a small sample size (n=4). Hydraulic and electrical systems, despite high frequency, show moderate performance (35.9%, 25.3%), indicating semantic variation that likely requires additional training data.

Key limitations include: (i) sensitivity to the quality of synthetic training data, (ii) limited support for complex faults with multiple interacting causes, (iii) omission of temporal context such as recurring failures and maintenance history, and (iv) absence of active learning to incorporate operator feedback. Future work can consider iterative refinement using real logs, contextual embeddings that incorporate equipment history, and hierarchical taxonomies for multi-level classification.

6. Future Work

The main bottleneck remains synthetic-to-real transfer. A practical next step is to add a modest set of additional real-world annotations, sampled to cover the long tail of categories (e.g., Motor, Smörjsystem, Bom) and to capture operator shorthand and mixed Swedish English phrasing. These examples can seed an active-learning loop and also calibrate the augmentation pipeline by injecting realistic typos, abbreviations, and equipment codes rather than relying on templated variation. Unlabeled logs are another underused resource; continued contrastive adaptation or self-training on in-domain text may reduce the gap while keeping annotation costs manageable.

On the retrieval side, a single global α is effective but blunt: some queries hinge on exact subsystem codes, while others are best resolved through paraphrase matching. Learning a query-dependent fusion weight, or training a small ranker on semantic and lexical scores, would allow the system to shift emphasis when needed. Investing in the taxonomy passages is similarly low-friction: adding known synonyms and abbreviations, and optionally reranking the top-K candidates with a compact cross-encoder, may improve Recall@1 without changing the overall workflow.

Finally, many work orders contain multiple interacting issues. Supporting multi-label outputs and hierarchical taxonomy levels would better match how faults are reported and used in maintenance analysis. Combining retrieval with contextual signals (equipment ID, recent work orders, recurring failures) and evaluating the tool as decision support time-to-label, consistency, and user trust would help translate retrieval metrics into operational impact.

7. Conclusion

This work introduces a hybrid semantic lexical search system for automated taxonomy classification of Swedish mining maintenance logs. Through systematic model comparison, contrastive fine-tuning, and α parameter tuning, the approach achieves 45.9% Recall@1 and 75.4% Recall@5 on real-world data, substantially outperforming single-modality baselines. The hybrid architecture with GTE-V1 embeddings and TF-IDF fusion (α=0.90) demonstrates the value of combining semantic matching with term-based retrieval for industrial text classification in low-resource multilingual settings. The methodology is applicable to maintenance text mining across industries and languages and can be extended to hierarchical taxonomies and temporal failure patterns.

References

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451.

Dalzochio, J., R. Kunst, E. Pignaton, A. Binotto, S. Sanyal, J. Favilla, and J. Barbosa (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in industry 123*, 103298.

Karpukhin, V., B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih (2020). Dense passage retrieval for open-

domain question answering. In *EMNLP (1)*, pp. 6769–6781.

Khattab, O. and M. Zaharia (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48.

Lin, J., X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira (2021). Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2356–2362.

Malmsten, M., L. Börjeson, and C. Haffenden (2020). Playing with words at the national library of sweden–making a swedish bert. *arXiv preprint arXiv:2007.01658*.

Reimers, N. and I. Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robertson, S., H. Zaragoza, et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval 3*(4), 333–389.

Sharma, A., G. Yadava, and S. Deshmukh (2011). A literature review and future perspectives on maintenance optimization. *Journal of Quality in Maintenance Engineering 17*(1), 5–25.

Wang, L., N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wang, Z., K. Ezukwoke, A. Hoayek, M. Batton-Hubert, and X. Boucher (2025). Natural language processing (nlp) and association rules (ar)-based knowledge extraction for intelligent fault analysis: a case study in semiconductor industry. *Journal of Intelligent Manufacturing 36*(1), 357–372.

Ye, Z., Y. J. Kumar, G. O. Sing, F. Song, and J. Wang (2022). A comprehensive survey of graph neural networks for knowledge graphs. *IEEE Access 10*, 75729–75741.