Journal of Applied Automation Technologies

Research Paper

# RoBERTa-Based Sentiment Mining for Engineering Technical Reports

**Piotr Aleksander Kamiński[1], Natalia Joanna Dąbrowska[1], Anna Maria Nowicka[2] and Natalia Woźniak[2,*]**

[1] Faculty of Computer Science, Wrocław University of Science and Technology, 50-370, Wrocław, Poland
[2] Faculty of Computer Science, University of Warsaw, 00-927, Warsaw, Poland
*Corresponding author: natalia.w@student.uw.edu.pl

**Abstract.** Engineering technical reports document the operations and changes, providing a basis for decision-making. Due to the highly specialized nature of these documents, sentiment extraction becomes challenging because they contain complex jargon, formal structures, and the implicit nature of evaluative language. This paper uses a customized RoBERTa model to build a sentiment analysis system for engineering technical reports. These three methods are used to extract subtle evaluative expressions in technical papers. These methods include expanding domain-specific vocabularies, adaptive segmentation and hierarchical embeddings, as well as customized attention mechanisms. A multi-source engineering corpus containing over 30,000 real technical documents was evaluated, and detailed annotation and segmentation were performed. The proposed model significantly outperforms traditional baseline models on datasets with higher term density, achieving an average accuracy of 93% and a macro F1 score of over 0.81 in negative sentiment detection. Without domain vocabulary adaptation and structural encoding, the F1 score would drop by more than 10%. Targeted model adaptation can improve the accuracy of sentiment mining in the engineering field and support the implementation of more reliable safety management, compliance assessment, and other measures.

**Keywords:** *Computer Sentiment Analysis, Domain Adaptation, RoBERTa, Engineering Reports, Technical Text Mining, Deep Learning*

Received on 16 November2025, Accepted on 12 February 2026, Published on 28 February 2026

Copyright © 2026 Author, licensed to JAAT. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

## Introduction

Technical reports are the primary tools for documenting, decision-making, and communicating various parts of industrial systems in engineering practice. The goal of disseminating data, insights, and recommendations within the organization is to optimize processes, reduce risks, and ensure compliance at both the organizational and cross-organizational levels [1]. Large construction projects require a system to manage these records in order to create operational trails, verify work outcomes, and impart technical knowledge to new workers [2]. As the complexity and scale of modern engineering projects (such as power and transportation facilities, high-tech manufacturing, and new era construction) increase, quality management, resource utilization, continuous monitoring, and control become increasingly important [3,4]. As the process of digitalizing enterprises progresses, natural language processing (NLP) technology is now used to automatically extract, classify, and analyze important information in technical documents to enhance the organization's responsiveness and adaptability [5].

Methodological issues, such as technical terminology, strict structural norms, and the objective and factual language of these reports, can distinguish emotions and implied meanings from engineering reports. Technical documents usually do not exhibit emotions; instead, they are typically expressed in a conservative manner, such

as cautious evaluations or prudent recommendations. Consumer reviews and public discussions on social media are more direct than these emotions [6]. In technical writing texts, traditional sentiment analysis models, such as dictionary-based methods, shallow neural networks, and classic transformer architectures, usually perform poorly. This is mainly attributed to the lack of lexical adaptability, insensitivity to the context of technical writing, and the inability to handle embedded evaluation differences [7]. Due to the complex relationships between terminology, domain knowledge, and report structure, some important indicators are overlooked in the process of automated opinion mining and subsequent decision support analysis [8]. These issues must be addressed in order to build high-performance, fault-tolerant sentiment intelligence systems, thereby truly enhancing the capabilities of human experts in safety management, defect tracking, and engineering project supervision [9,10].

This paper proposes a domain-adaptive RoBERTa model, which has been optimized to extract sentiment and evaluation signals from engineering technical reports. By improving the attention mechanism to better identify structural and contextual dependencies, and by enhancing input representations through custom tokenization and specialized vocabulary, the proposed model aims to address previous shortcomings and provide a new high-quality reference for engineering-oriented sentiment analysis. Comprehensive experiments used many real-world technical corpora, and the results were compared with traditional state-of-the-art benchmark models. In-depth ablation studies were conducted to determine the innovative effects of each model.

## Background and Related Work

### Sentiment Analysis for Engineering Documents

Engineering technical documents are widely used across various industries and serve as the basis for various operational data and expert opinions [11]. Compared to other types of writing, engineering reports are characterized by objectivity and factuality, and they use a lot of technical terminology. Since evaluative information is usually not directly stated, automatic sentiment analysis involves the aforementioned factors [12].

Sentiment analysis is used in engineering documents in three aspects [13]: The first is to extract risk assessments from maintenance logs. The second is to identify uncertainties or confidence in project proposals. The third is to mark potential quality issues or non-conformities in the inspection record. Safety risks are also another reason. A decline in well-being or warnings of abnormal behavior may be signals of real-time operational risks [14]. In large engineering companies, regularly monitoring changes in the emotions of different groups can help identify potential issues early on [15]. General natural language processing models cannot recognize such attitudes or emotional signals, their organized and concise style, and the terminology, exceptions, or specific suggestions used by engineers in their opinions [16].

Sentiment extraction has become more difficult due to the complexity of domain-specific vocabulary, the use of abbreviations, and the complex systems within the report sections [17]. Project documents, standards, and specifications form the foundation of many technical reports. Text mining algorithms may not be able to achieve this goal. The industry has a high demand for high-performance sentiment analysis tools, but technical issues also limit their use [18].

### Pretrained Language Models and RoBERTa

The field of sentiment analysis has been continuously advancing with the development of natural language processing technology. Deep learning models based on large-scale language models [19] are being used. Early domain adaptation attempts used static word embeddings or sentiment lexicons, but they could not handle variations in syntactic forms and domain-specific terminology [20]. BERT is an example of context-aware. By creating high-quality semantic representations and applying them in various contexts, it has transformed the field of research and development [21].

RoBERTa (Robustly optimized BERT approach) is an optimized version of BERT. Changed the pre-training process, removed the next sentence prediction objective, and used larger dynamic data batches for training [22]. Bidirectional understanding will be deeper, and subsequent tasks of classification and extraction will be more accurate. RoBERTa is a multi-head self-attention mechanism that can consider the relationships between words in different positions to better understand the context of evaluative expressions and their application in technical papers [23].

RoBERTa is relatively strong in text mining within the engineering field, with enhanced resistance to out-of-vocabulary words, better handling of long-distance dependencies, and finer-grained differentiation of emotional fluctuations caused by conditional recommendations or risk assessments [24]. RoBERTa is a highly scalable pre-training system that performs well in terms of adaptability to specific domains. It can be fine-tuned on relatively small, high-value engineering corpora to achieve practical sentiment detection results. RoBERTa is a strong candidate for addressing the discovery of hidden emotional issues in well-structured and highly technical texts, due to its token-based attention mechanism and advanced input encoding options [25].

Despite the aforementioned advantages, the application of general RoBERTa in technical reports may still encounter issues such as vocabulary mismatch, lack of clear sentiment indicators, and misalignment with the hierarchical logic of report organization. Many studies have already begun to fine-tune pre-trained models for specific domains, focusing on learning the vocabulary and knowledge pertinent to those domains.

**Gaps and Research Motivation**

Transformer-based architectures perform well in sentiment analysis, but there are still some issues with extracting engineering reports. Most models cannot be trained on complex, diverse, and emotionally rich datasets, and are unable to handle emotional data from technical or other specialized fields. The language norms and structured layout of engineering documents are not met, which may lead to false positives (describing emotions as false negatives) or false negatives (failing to identify weak evaluations).

The scalability of fine-tuning, the incorporation of domain-specific vocabulary, and the interpretability of attention and classification decisions in safety-critical environments remain significant issues. New technologies are also being used to expand the vocabulary, create context-aware representations, and perform structural parsing of engineering documents. This paper systematically addresses the aforementioned open issues by using and extending RoBERTa. This is to obtain reliable sentiment and evaluation expressions from actual engineering technical reports.

## RoBERTa-Based Sentiment Mining Methodology

**Domain-specific Representation Design**

The technical language of engineering reports requires a representation scheme capable of capturing intricate terminology, complex multi-word expressions, and hierarchically encoded domain semantics. To address this, an adaptive tokenization process is constructed in which the input sequence $X = \langle w_1, w_2, \dots, w_N \rangle$ undergoes a hierarchical segmentation that leverages both subword statistics and an engineered vocabulary lexicon $V_{\text{eng}}$ curated from the target corpus. The segmentation function $\Psi$ maps each document to a sequence of composite tokens:

$$Z = \Psi\big(X, V_{eng}\big) = \langle t_1, t_2, \dots, t_M \rangle \quad \text{Eq.(1)}$$

where $t_k$ may consist of single terms, chemical formulae, or standardized engineering abbreviations depending on co-occurrence and relevance, ensuring preservation of domain granularity.

Each composite token is then embedded via a hybridized vector mapping that synthesizes three key semantic facets-contextual intent, technical specificity, and segment position-into a single representation. The token embedding function is formalized as

$$E(t_k) = P_k \odot S_k + Q_k \otimes I_k \quad \text{Eq.(2)}$$

where $P_k$ encodes positional dependencies, $S_k$ captures subdomain alignments via latent clusters, $Q_k$ represents technical term indices, and $I_k$ is an intent signal derived from co-located linguistic cues. The use of tensorized operations between these components enables cross-view integration that surpasses classical embeddings in capturing fine-grained engineering relationships.

Feature composition then aggregates token representations at the sequence, segment, and document level, dynamically adapting the fusion strategy according to report structure. The composite feature function is defined as

https://doi.org/10.64972/jaat.2026v4.115

$$F_{doc} = \sum_{i=1}^{M} \alpha_i E(t_i) + \Gamma\left(\left[E\left(t_{h_1}\right), \ldots, E\left(t_{h_L}\right)\right]\right) \quad \text{Eq.(3)}$$

where $\alpha_i$ are dynamically computed attention weights reflecting token importances, and $\Gamma$ is a structural pooling operator integrating headlines, section headers, and domain-specific delimiters. This encapsulates not only content but meta-structure, empowering robust representation for downstream processing.

**Model Structure and Enhancements**

The backbone of the proposed approach is a multi-layered RoBERTa encoder specifically reengineered for the idiosyncrasies of engineering documentation. Unlike generic transformers, each encoder block processes input not only through standard self-attention, but also through a domain-adaptive attention operator $\mathcal{A}_D$ capable of contextually gating interactions based on technical term dependencies. For an input sequence $H$, attention maps are computed as:

$$\mathcal{A}_D(H) = \text{softmax}\left(\frac{Q_D K_D^{\top} + \Omega_D}{\sqrt{d}}\right) V_D \quad \text{Eq.(4)}$$

where $Q_D, K_D, V_D$ are learned projection matrices for domain-adjusted queries, keys, and values, and $\Omega_D$ models external domain constraints through a knowledge-driven mask.

The core output for each sentence representation is then derived by transforming the contextual token matrix $H^{(L)}$ from the last encoder layer, which undergoes an advanced consensus mechanism. The scoring function takes the form:

$$r_{\text{sent}} = \sigma\left(W_o \cdot \text{ReLU}\left(H^{(L)}\mathbf{1}^{\top}\right) + b_o\right) \quad \text{Eq.(5)}$$

where $W_o$ and $b_o$ are output weights and biases fine-tuned for the sentiment extraction objective, and sequence-level consolidation is achieved by the sum-over activation tracks throughout constituent tokens.

Information is further aggregated layer-wise to remedy information dispersion across depth. The document-level representation $D$ is derived via non-linear hierarchical stacking:

$$D = \phi\left(\sum_{l=1}^{L} \lambda_l\left(H^{(l)}\right)\right) \quad \text{Eq.(6)}$$

where $\lambda_l$ is a learned significance coefficient for each transformer depth and $\phi$ denotes a nonlinear transformation network tuned for engineering evaluation structure.

Figure 1 shows the entire pipeline based on the RoBERTa model. The pipeline processes engineering texts through embeddings, tokenization, multi-layer encoders, cross-attention with a curated domain dictionary, and task-specific decision heads to explicitly leverage prior knowledge in the technical domain.

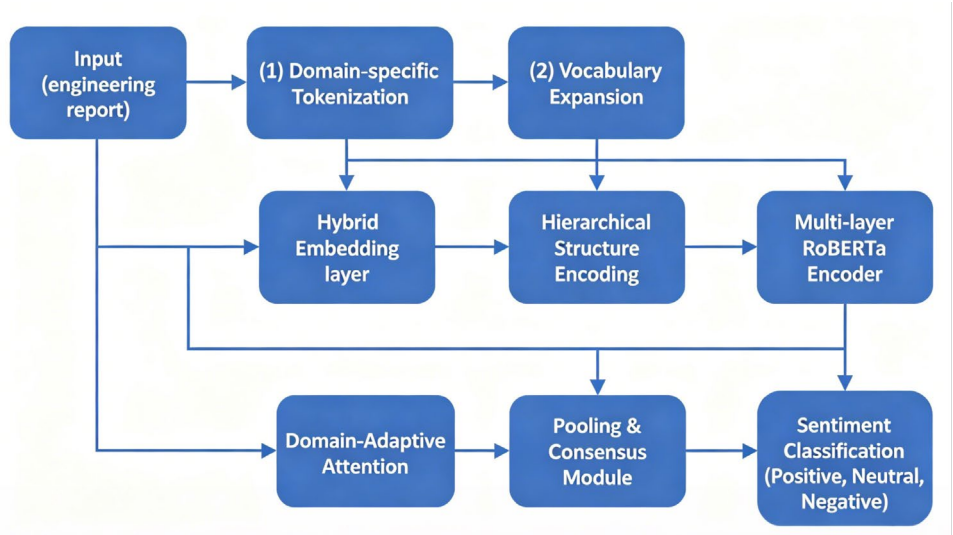


**Figure 1.** RoBERTa-Based Model Structure Overview

https://doi.org/10.64972/jaat.2026v4.115

### Training and Optimization

In order to ensure the accuracy of the model, the engineering report will be cleaned by removing private title information, standardizing unit expressions, and synchronizing chapter numbers before model construction. In the training loop, the adaptive weighted loss function addresses the class imbalance issue in critical event reports. The formal statement of the loss is as follows:

$$\mathcal{L}_{\text{sent}} = -\frac{1}{N}\sum_{i=1}^{N} \omega_{c_i}\left(y_i \log\, \hat{y}_i + (1-y_i)\log\,(1-\hat{y}_i)\right) + \mu\|W\|_* \qquad \text{Eq.(7)}$$

where $\omega_{c_i}$ are category-aware weights tuned per sample $i, y_i$ and $\hat{y}_i$ denote ground truth and predicted sentiment, $\mu$ is a regularizing constant, and $\|W\|_*$ is the nuclear norm penalizing overcomplex decision boundaries.

Simultaneously using curriculum-based sampling, early stopping, and optimizer selection to improve the reliability of convergence and prevent over-specialization of reported artifacts. As shown in Figure 2, the above process includes data collection and feature construction, providing feedback based on performance statistics iterations, applying domain-adaptive RoBERTa encoding techniques, and offering a stable platform for sentiment mining projects.

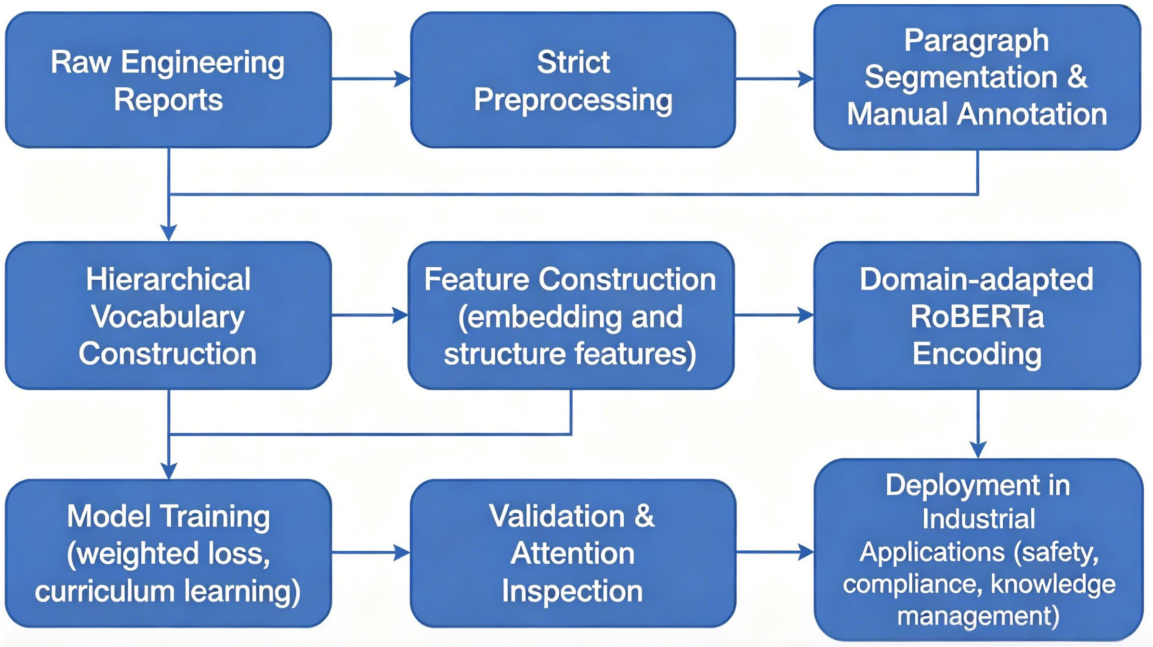


**Figure 2.** Sentiment Mining Workflow

## Empirical Evaluation

### Datasets and Experimental Setup

The empirical foundation of this study comes from multiple sources, including operational compliance assessments, process certification reports, safety audit reports, and maintenance records. The data includes over 30,000 documents; all non-standard records and partially edited submissions have been removed, and this section has been categorized in other ways. According to the consensus of experts and technical editors in the field, they are categorized as negative, neutral, or positive. The distribution of actual industry reports is similar, and the final corpus distribution is naturally uneven.

Before the model is introduced, the original text will be standardized to reduce differences in unit measurements, date and time formats, and formula expressions. This normalization process is defined as a parametrized transformation for document $d$ as follows:

https://doi.org/10.64972/jaat.2026v4.115

$$\Omega(d) = \Theta\left(\epsilon_{\text{unit}} \cdot \varphi_{\text{num}}(d) + \zeta \cdot \chi_{\text{sym}}(d)\right) \quad \text{Eq.(8)}$$

where $\epsilon_{\text{unit}}$ adjusts all measurable units to a canonical reference, $\varphi_{\text{num}}$ encodes decimal and scientific notations, $\zeta$ scales the syntactic harmonization, and $\chi_{\text{sym}}$ standardizes engineering symbols via symbol-to-token rewrite logic, both to ensure lexical compatibility and semantic retention.

When partitioning the dataset, each labeled document is deterministically assigned to the training set, validation set, or test set. This depends on the hash value of the unique document identifier. These groups are different in terms of time and topic, making it difficult to overlap with other document sets. By determining the modulus and threshold before random allocation, and aligning the split ratio with the true temporal and thematic distribution of the corpus, stable and non-overlapping datasets can be obtained. These datasets preserve the heterogeneity of real-world tasks and prevent the leakage of contextual or author characteristics between training and evaluation. The aforementioned regulations also allow for the checking of file duplicates, order dependency errors, or old report versions.

Supervised evaluation leverages a strict accuracy and precision audit. Let $y_j$ and $\hat{y}_j$ represent true and predicted class for the $j$-th sample. The aggregate metric for class-wise evaluation is:

$$M_c = \frac{1}{|S_c|} \sum_{j \in S_c} \mathbb{I}\left(y_j = \hat{y}_j\right) \quad \text{Eq.(9)}$$

where $S_c$ indexes the subset of samples in category $c$,supporting detailed classwise and macro-level validation of predictive performance.

**Baseline Methods and Metrics**

The benchmark framework is a comprehensive study of modeling methods using TF-IDF vectorized linear Support Vector Machines (SVM), Long Short-Term Memory (LSTM) recurrent networks, dynamic input gating, and linear Support Vector Machines (SVM) with TF-IDF vectorization. The benchmark framework also fine-tuned the standard BERT on the original engineering text. For the sake of comparison, the hyperparameters of all benchmark models have been optimized within this domain. To ensure reproducibility, the random seed generation has been standardized.

Given class imbalance and the subtlety of technical sentiment, classical accuracy is insufficient for operational assessment. Instead, the macro F1-score is adopted:

$$F_1 = 2 \cdot \frac{\bar{P} \cdot \bar{R}}{\bar{P} + \bar{R}} \quad \text{Eq.(10)}$$

where $\bar{P}$ and $\bar{R}$ denote average precision and recall over all classes.

Recall for class $c$ is defined in terms of ordered pairs:

$$R_c = \frac{TP_c}{TP_c + FN_c} \quad \text{Eq.(11)}$$

with $TP_c$ and $FN_c$ denoting true positives and false negatives for category $c$, ensuring model sensitivity to minority classes.

To dissect the technical merit of each approach, a composite component score for ablation diagnosis is proposed:

$$S_{\text{comp}} = \eta_o \cdot F_1 + \lambda_r \cdot R_c + \mu_a \cdot M_c \quad \text{Eq.(12)}$$

where $\eta_o, \lambda_r, \mu_a$ are empirically determined weights reflecting practical priorities in engineering context sentiment assessment.

**Component Ablation and Sensitivity Analysis**

The ablation analysis delves into the contribution of tokenization schema, customized embedding augmentation, and the domain-informed attention stratagem. Each component is selectively masked while monitoring drops in comprehensive sentiment recognition and the impact on minority label precision. The impact coefficient for a given module $Q$ is expressed as:

https://doi.org/10.64972/jaat.2026v4.115

$$\beta_Q = \frac{\Delta F_1(Q)}{\sum_k |\Delta F_1(Q_k)|} \quad \text{Eq.(13)}$$

where $\Delta F_1(Q)$ quantifies F 1 -score degradation upon ablation versus the full model, normalized by aggregate changes from all ablated modules.

Sensitivity to hyperparameter adjustment is quantified via normalized variance analysis across controlled runs. Let $\theta$ index a vector of tunable parameters over $R$ runs, then the sensitivity variance is defined as:

$$\sigma^2_{\text{sens}} = \frac{1}{R}\sum_{r=1}^{R}\left(S^{(r)}_{\text{comp}} - \frac{1}{R}\sum_{k=1}^{R} S^{(k)}_{\text{comp}}\right)^2 \quad \text{Eq.(14)}$$

This captures model robustness, confirming that high utility components consistently yield stable performance under realistic tuning, while superfluous modules exhibit sensitivity spikes, validating design minimalism.

The practical value, adaptability, and depth of sentiment analysis based on the RoBERTa framework in complex engineering documents have been demonstrated. The evaluation axs include dataset rigor, baseline clarity, and detailed ablation methods.

## Results and Discussion

### Overall Performance Comparison

Table 3 shows the performance of all proposed models on multiple real-world engineering datasets. Fine-grained data can be obtained through multi-dimensional design. Figure 3(a) shows the accuracy curves of the five models tested on all three corpora: LSTM, SVM, standard BERT, domain-enhanced BERT, and RoBERTa variants. To reduce partition variability, results are obtained from five or more random seeds at a single data point. The performance of the RoBERTa model on the very specialized "maintenance" dataset can be improved by optimizing the embedding layer and input layer. Compared to other models, this is possible. In this dataset, RoBERTa's accuracy reached approximately 93%, which is about 5% higher than that of neural networks and older models. In the compliance and certification corpus, they also have a similar but relatively smaller lead.

Figure 3(b) shows the overall F1 score and displays the neutral, negative, and positive sentiment indicators. The model is more likely to detect low-frequency negative emotions in engineering documents. RoBERTa's F1 score in the negative category exceeded 0.81, surpassing the 0.60 baseline of all non-transformer baselines and the 0.70 threshold set by other transformer configurations. Neutral and positive categories have more clustering, but domain-optimized tokenization and attention mechanisms generally perform better. Each sub-bar contains the average of five cross-validation folds, with a pair showing model-specific variance.

Furthermore, as shown in Figure 3(c), the accuracy is reported separately by dataset and sentiment category, generating a complete subplot with 15 data columns. At this stage, the RoBERTa model demonstrates good stability, with the accuracy of the adverse categories in all datasets exceeding 0.87. Considering the high industrial risk posed by false negatives, this is crucial for operations. BERT and LSTM are domain-specific, but they perform poorly in few-shot and high-term scenarios, with the negative class accuracy of the validation data being below 0.75. Models with domain-specific embeddings are more effective than most RoBERTa settings, but the gain is lower than that of the entire RoBERTa setting.

There is a consistent trend of high accuracy, consistency in F1 scores for each category, and relatively high precision for important categories. In technical sentiment analysis, RoBERTa-based methods outperform other methods in engineering operation safety and quality management.

### Ablation Results and Component Analysis

Figure 4 shows the results of the ablation experiment. In order to evaluate the functionality of components within a specific area, five basic ablation and diagnostic methods were used. Figure 4(a) shows the total F1 scores of the complete model and four selective ablation variants (removing vocabulary expansion, engineering-specific embeddings, hierarchical encoding, refined attention mechanisms, and segment position enhancement) for all three datasets. Removing domain vocabulary and hierarchy, the average F1 score performance on the maintenance and certification sets decreased by more than 10% compared to the original architecture. The

https://doi.org/10.64972/jaat.2026v4.115

impact of removing paragraph and positional features is relatively small. There are still some differences in the compliance reports, indicating that distinguishing similar engineering expressions is helpful.

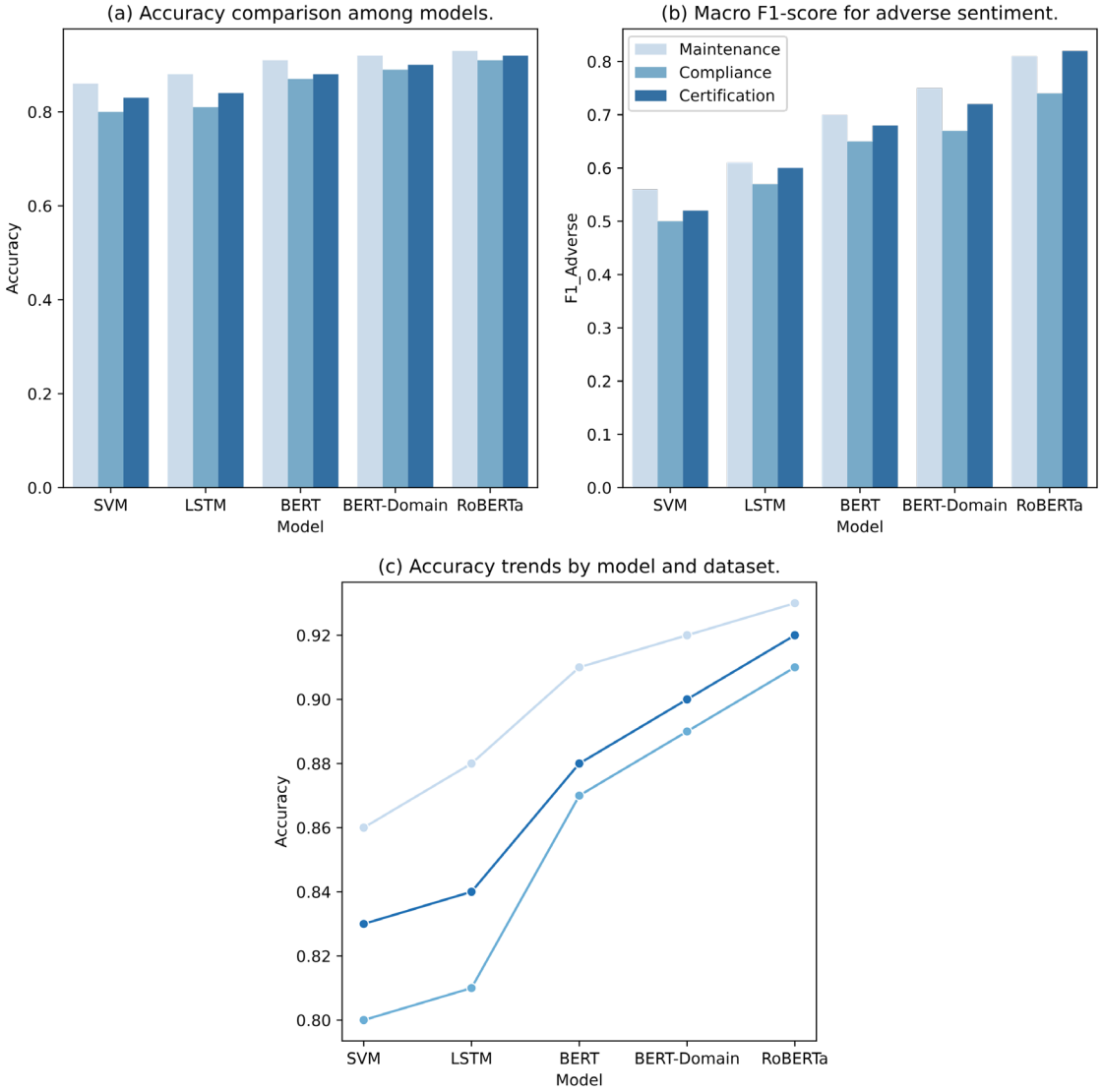


**Figure 3.** Overall Model Performance: (a) Accuracy; (b) Macro F1-score; (c) Precision for each model and dataset

Figure 4(b) shows the model's sensitivity to the reduction in technical vocabulary richness. Only when retaining vocabulary and embeddings does the high-term report show a certain level of robustness; when the input term density is artificially reduced, the F1-score significantly decreases, especially in negative sentiment detection. Figure 4(c) shows that, in the absence of the target module, the precision of negative emotions remains relatively high, and the increase in false positives is significantly reduced. Figure 4(d) shows the recall rate of the negative category after ablation. By using special embeddings and hierarchical context, subtle alert signals can be found in complex texts.

Figure 5 shows the term connection diagram and the sentiment recognition structure. Figure 5(a) shows the accuracy of different reports under varying term densities. In high-density terms, models with domain-specific components can still maintain an accuracy of over 90%. Figure 5(b) depicts the relationship between chapter structure complexity and F1, such as the presence of nested bullet points or appendices. The F1 score of the RoBERTa variant is higher than that of the general model. The error heatmap overlaid on the document layout is Figure 5(c). The error hotspots in the section divisions and conclusion summaries are due to the omission of structural coding, and these areas contain important evaluation metrics.

https://doi.org/10.64972/jaat.2026v4.115

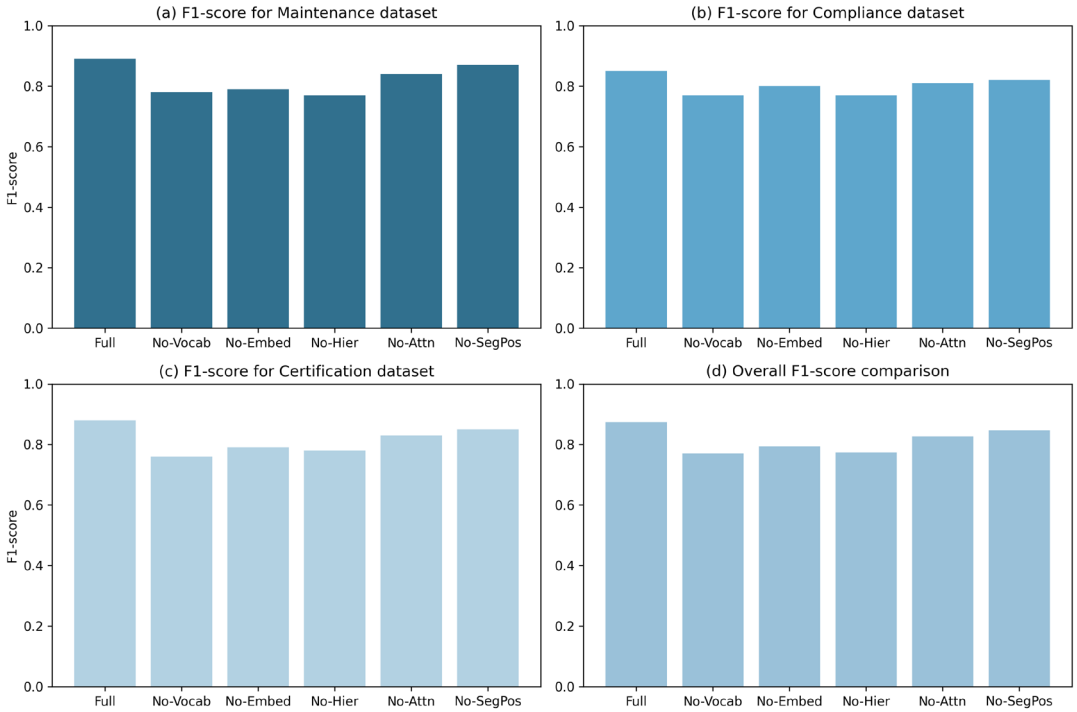


**Figure 4.** Ablation and Component Effects: (a) F1-score for full model and five ablation settings across datasets; (b) Sensitivity to jargon density reduction; (c) Precision drops under component removal; (d) Recall for adverse class by ablation

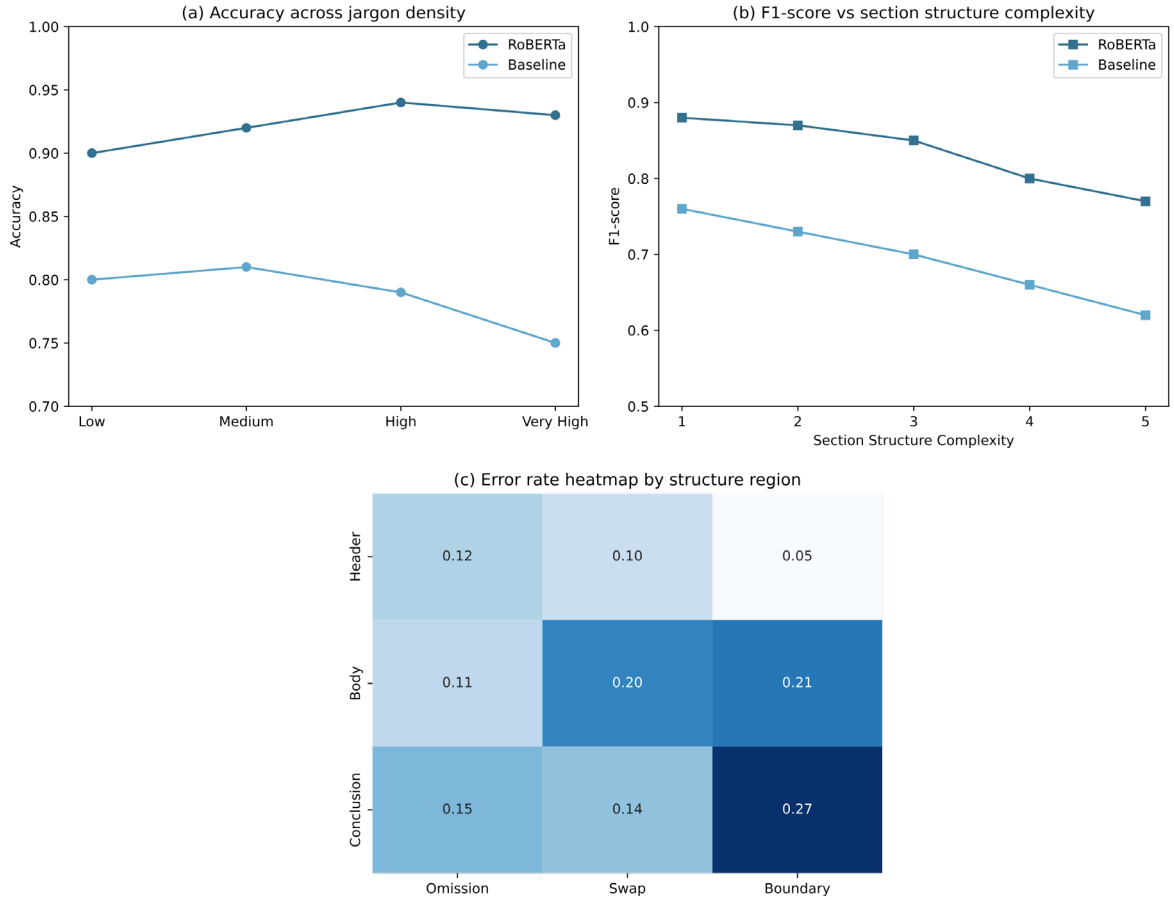


**Figure 5.** Jargon and Structure Analysis: (a) Accuracy across varying jargon densities; (b) F1-score versus report section complexity; (c) Error heatmap overlays by structural region

https://doi.org/10.64972/jaat.2026v4.115

No single feature is universally superior to others. In the context of low signal-to-noise ratio complexity, combining domain-specific knowledge, multi-level input decomposition, and attention mechanisms is very useful. In order to improve the comprehensiveness and fine-grained accuracy of sentiment detection, it is necessary to make multi-level adjustments and integrations of the features of engineering reports.

**Error, Attention, and Interpretability**

As shown in Figures 6 and 7, the detailed information on the defects of the proposed domain-adaptive RoBERTa model includes interpretability mechanisms and error cases, showing the attention positions of the model during the inference process and various types of failures. Relatively transparent, used for safety or quality-critical engineering.

Figure 6(a) shows the attention heatmap of samples with strong negative emotional characteristics, such as process warnings or explicit alerts. RoBERTa is a large language model that gives more weight to structured technical terms in risk, cause, countermeasure, and dense evaluation summary statements. For example, phrases like "critical deviation detected" or "electrical fault unresolved" are often used in maintenance records. The model's interpretive heuristic methods in this field are only applicable to this specific application. As shown in Figure 6(b), the focus of the neutral report is on the program description and measurement readings, with these weights evenly distributed.The neutral report contains less evaluative information. All these different types of attention patterns can provide a linguistic foundation for model decisions and offer practical methods for post-audit, especially in regulatory or engineering supervision cases.

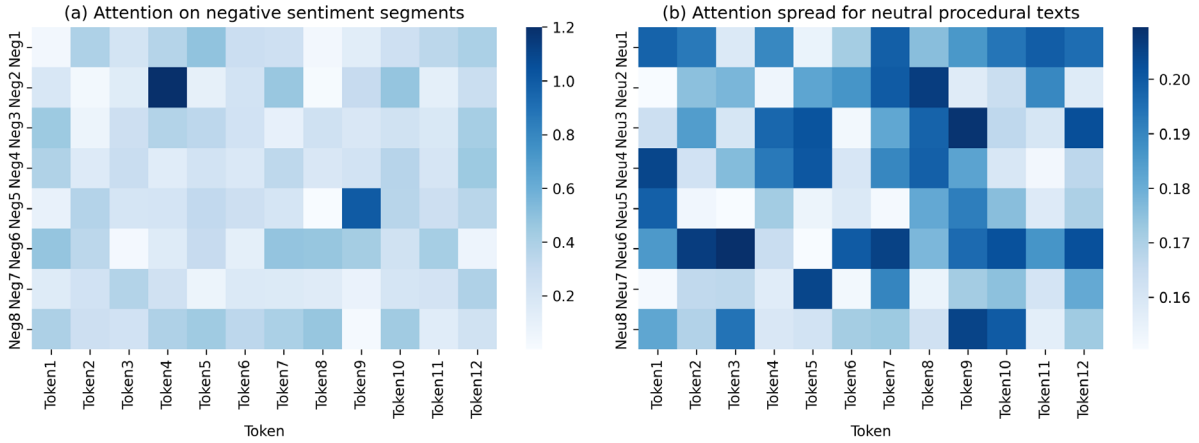


**Figure 6.** Attention Heatmap Visualization: (a) Model attention on key negative sentiment segments; (b) Attention spread for neutral procedural texts

Figure 7 shows the complex structure of model errors. Figure 7(a) categorizes the errors into the following types: neutral/unfavorable, neutral/favorable, neutral/favorable, and confusion between ambiguous annotation cases. Most errors are not due to large-scale misclassification, but rather due to granularity loss at the category boundaries. This is especially true for nested conditional statements or complex technical fuzzy expressions. Figure 7(b) shows that there are relatively more errors in areas such as tables, appendices, and other irregular chapter titles. Due to frequently being out of sequential context, they are not suitable for special embeddings and hierarchical models. As shown in Figure 7(c), the distribution of errors is also affected by different model variant architectures. The baseline transformer model lacks explicit structural encoding and exhibits a wide range of unpredictability in error localization, whereas the errors in RoBERTa-based variants are limited to a small number of structurally irregular samples.

https://doi.org/10.64972/jaat.2026v4.115

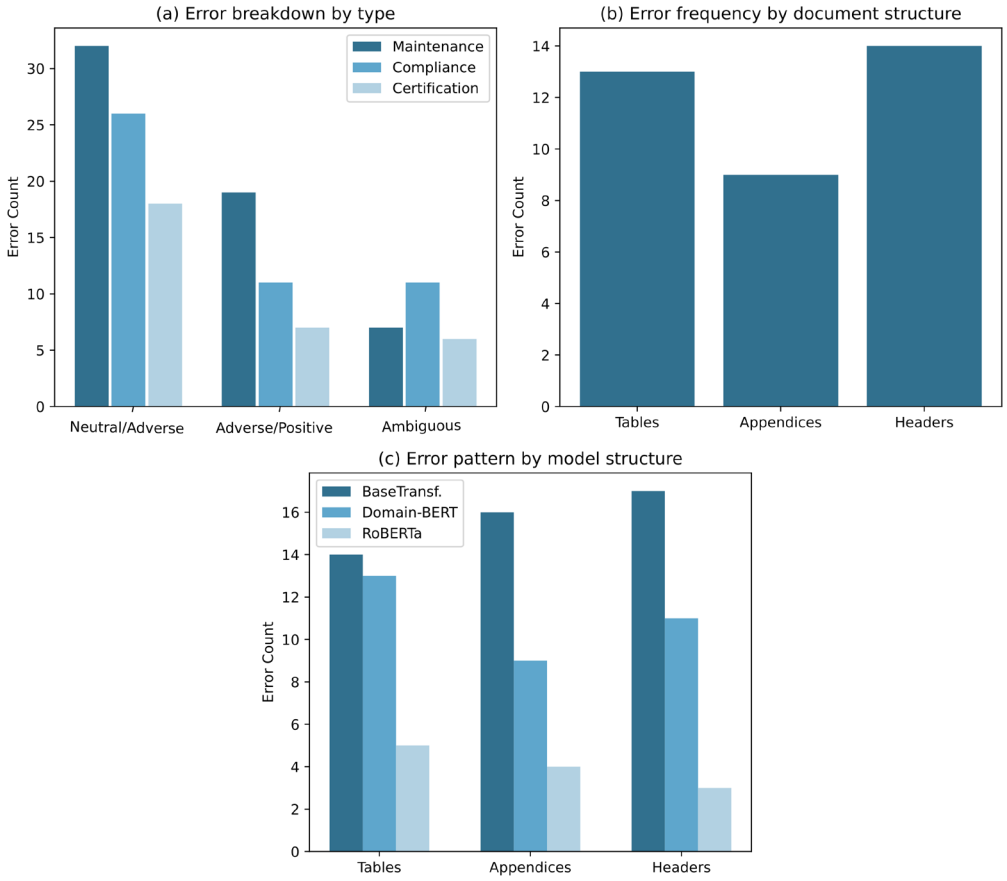


**Figure 7.** Failure Type Analysis: (a) Error breakdown; (b) Error localization by document structure; (c) Comparison of error patterns between models

Based on the multifaceted failure analysis of Figures 6 and 7, it can be determined that the primary cause of the framework's failure is not generalization collapse, but rather the inherent complexity of engineering language, including interdisciplinary terminology, fine-grained conditional statements, and noise in document layout. This concept can provide a method for iterative optimization and improve the reliability of model output under high-performance conditions. The attention heatmap provides a detailed description of the distribution of evaluation cues in the actual data, and the errors found are usually interpretable and can be managed through improved structure-aware encoding or targeted corpus annotation.

## Conclusion

This paper conducts a comprehensive and technically demanding study on sentiment mining in engineering documents. A RoBERTa model for deployment and optimization in the field has been proposed. The new model can address some issues in technical papers, such as dense jargon and unexplained evaluations, by adding custom vocabulary construction, hierarchical paragraph encoding, and centralized attention mechanisms. It can also handle the different structures within these documents. The three emotions (negative, neutral, and positive) have been enhanced in terms of accuracy and detection capability because some new technological innovations have been introduced into the actual engineering corpus.

This is not the result of simple adjustments or area expansions. This is the result of careful consideration at each stage (including tokenization and embedding, paragraph context modeling, attention distribution optimization, etc.). The RoBERTa framework outperforms traditional machine learning benchmarks and many non-specialized BERT variants, even when the training data is complex, context-dependent, or has rare sentiments. By using advanced attention visualization and error map diagnostics, the model's interpretability has been shown to be directly related to the training methods of domain language and structural specifications. Errors often occur in very irregular areas or at boundaries. Fine-tuning and adaptive learning will be guided in the future.

The modular design of the new RoBERTa pipeline improves work efficiency and provides important support for enterprise operations management, compliance, and security in an open and reliable manner. Scientific publishing databases, process monitoring records, and adaptive regulatory filing systems are the application areas for future research outcomes. Expanding these technologies to integrate structured data sources and multimodal signals, or improving the semantic tracking of relationships between documents, are all directions for future work. Establishing a new high standard for sentiment analysis in complex engineering lays the foundation for future research and applications.

**Author Contributions**

Piotr Aleksander Kamiński, Natalia Joanna Dąbrowska contribute to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Anna Maria Nowicka and Natalia Woźniak contribute to conceptualization, methodology, software and project administration. All authors have read and agreed with the manuscript before its submission and publication.

**Funding**

This research received no specific financial support from any funding agency.

**Institutional Review Board Statement**

Not applicable.

## References

[1] Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022). A fine-tuned BERT-based transfer learning approach for text classification. Journal of healthcare engineering, 2022(1), 3498123. https://doi.org/10.1155/2022/3498123

[2] Abdullah, T., & Ahmet, A. (2022). Deep learning in sentiment analysis: Recent architectures. ACM Computing Surveys, 55(8), 1-37. https://doi.org/10.1145/3548772

[3] Huang, Y., Chen, J., Zheng, S., Xue, Y., & Hu, X. (2021). Hierarchical multi-attention networks for document classification. International Journal of Machine Learning and Cybernetics, 12(6), 1639-1647. https://doi.org/10.1007/s13042-020-01260-x

[4] Karim, M., Missen, M. M. S., Umer, M., Sadiq, S., Mohamed, A., & Ashraf, I. (2022). Citation context analysis using combined feature embedding and deep convolutional neural network model. Applied sciences, 12(6), 3203. https://doi.org/10.3390/app12063203

[5] Guo, Y., Ai, X., & Luo, W. (2024). A multi-task learning risk assessment method for the chemical process industry. Process Safety and Environmental Protection, 186, 980-994. https://doi.org/10.1016/j.psep.2024.04.030

[6] Sugu, N., & Babu, N. V. (2024, March). Text-Based Emotion Recognition with Hybrid Feature Selection and Ensemble Classification: A BERT and RoBERTa Approach. In International Conference On Health Informatics, Intelligent Systems And Networking Technologies (pp. 611-623). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-96-4008-9_46

[7] Cui, L. (2025). Construction and implementation of knowledge enhancement pre-trained language model for text sentiment analysis. Systems and Soft Computing, 7, 200293. https://doi.org/10.1016/j.sasc.2025.200293

[8] Zou, H., & Wang, Y. (2024). A novel automated framework for fine-grained sentiment analysis of application reviews using deep neural networks. Automated Software Engineering, 31(2), 43. https://doi.org/10.1007/s10515-024-00444-x

[9] Maitama, J. Z., Idris, N., Abdi, A., Shuib, L., & Fauzi, R. (2020). A systematic review on implicit and explicit aspect extraction in sentiment analysis. IEEE Access, 8, 194166-194191. https://doi.org/10.1109/ACCESS.2020.3031217

[10] Li, Z., Weng, S., Xia, Y., Yu, H., Yan, Y., & Yin, P. (2024). Cross-domain damage identification based on conditional adversarial domain adaptation. Engineering Structures, 321, 118928. https://doi.org/10.1016/j.engstruct.2024.118928

[11] Vithanage, D., Yu, P., Wang, L., & Deng, C. (2024). Contextual word embedding for biomedical knowledge extraction: a rapid review and case study. Journal of healthcare informatics research, 8(1), 158-179. https://doi.org/10.1007/s41666-023-00157-y

[12] Azari, M. S., Flammini, F., Santini, S., & Caporuscio, M. (2023). A systematic literature review on transfer learning for predictive maintenance in industry 4.0. IEEE access, 11, 12887-12910. https://doi.org/10.1109/ACCESS.2023.3239784

[13] Xie, Q., Tiwari, P., & Ananiadou, S. (2023). Knowledge-enhanced graph topic transformer for explainable biomedical text summarization. IEEE journal of biomedical and health informatics, 28(4), 1836-1847. https://doi.org/10.1109/JBHI.2023.3308064

[14] Shi, X., Zhang, Y., Yu, M., & Zhang, L. (2025). Deep learning for enhanced risk management: a novel approach to analyzing financial reports. PeerJ Computer Science, 11, e2661. https://doi.org/10.7717/peerj-cs.2661

[15] Li, Z., Dai, Y., & Li, X. (2022). Construction of sentimental knowledge graph of Chinese government policy comments. Knowledge management research & practice, 20(1), 73-90. https://doi.org/10.1080/14778238.2021.1971056

[16] Lei, Y., Qu, K., Zhao, Y., Han, Q., & Wang, X. (2024). Multimodal sentiment analysis based on composite hierarchical fusion. The Computer Journal, 67(6), 2230-2245. https://doi.org/10.1093/comjnl/bxae002

[17] Wang, S., Zhao, D., Zhang, C., Guo, Y., Zang, Q., Gu, Y., ... & Jiao, L. (2022). Cluster alignment with target knowledge mining for unsupervised domain adaptation semantic segmentation. IEEE Transactions on Image Processing, 31, 7403-7418. https://doi.org/10.1109/TIP.2022.3222634

[18] Demir, S., & Topcu, B. (2022). Graph-based Turkish text normalization and its impact on noisy text processing. Engineering Science and Technology, an International Journal, 35, 101192. https://doi.org/10.1016/j.jestch.2022.101192

[19] Wang, H., Liu, M., & Shen, W. (2023). Industrial-generative pre-trained transformer for intelligent manufacturing systems. IET Collaborative Intelligent Manufacturing, 5(2), e12078. https://doi.org/10.1049/cim2.12078

[20] Zhang, F. (2022). A hybrid structured deep neural network with Word2Vec for construction accident causes classification. International Journal of Construction Management, 22(6), 1120-1140. https://doi.org/10.18653/v1/P18-1032

[21] Wang, Q., Su, T., Lau, R. Y. K., & Xie, H. (2023). DeepEmotionNet: Emotion mining for corporate performance analysis and prediction. Information Processing & Management, 60(3), 103151. https://doi.org/10.1016/j.ipm.2022.103151

[22] Sun, T., Wang, S., & Zhong, S. (2022, September). Multi-granularity feature attention fusion network for image-text sentiment analysis. In Computer Graphics International Conference (pp. 3-14). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-23473-6_1

[23] Du, H., Jia, Q., Gehringer, E., & Wang, X. (2024). Harnessing large language models to auto-evaluate the student project reports. Computers and Education: Artificial Intelligence, 7, 100268. https://doi.org/10.1016/j.caeai.2024.100268

[24] Krishnamoorthy, P., Sathiyanarayanan, M., & Proença, H. P. (2024). A novel and secured email classification and emotion detection using hybrid deep neural network. International Journal of Cognitive Computing in Engineering, 5, 44-57. https://doi.org/10.1016/j.ijcce.2024.01.002

[25] Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D Management, 50(3), 329-351. https://doi.org/10.1111/radm.12408