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-p27669-cd

Intelligent Semantic Matching Model of Production Safety Hazard and Regulations in the Oil and Gas Industry

JiYu Zhai

Research Institute of Safety and Environment Technology, CNPC, China. E-mail: zhaijiyu@cnpc.com.cn

Yiyue Chen

Research Institute of Safety and Environment Technology, CNPC, China. E-mail: chenyiyue@ cnpc.com.cn

Shunyi Wang

Research Institute of Safety and Environment Technology, CNPC, China. E-mail: wangshunyi@ cnpc.com.cn

Junting Liu

Research Institute of Safety and Environment Technology, CNPC, China. E-mail: liujunting@ cnpc.com.cn

Jiangyang Han

Research Institute of Safety and Environment Technology, CNPC, China. E-mail: cyphernetics@126.com

Addressing the significant semantic gap between the colloquial expressions of on-site hazard descriptions and the professional terminology of safety regulations in the oil and gas industry—which hinders efficient matching—this paper proposes an intelligent semantic matching model based on enhanced contrastive learning. First, a specialized dataset of hazard-regulation pairs was constructed using real-world manual inspection data from oilfields. To tackle the challenges of insufficient negative sample diversity and slow convergence in the late stages of traditional contrastive learning, an initial negative sample selection method based on vector retrieval was designed. Furthermore, a Dynamic Hard Negative Mining (DHNM) mechanism is proposed to optimize the negative sample pool in real-time by monitoring the model's score fluctuations during iterative training. Architecturally, a deep encoder based on a bidirectional attention mechanism is employed to extract robust text embedding features, combined with the InfoNCE loss function to optimize the vector space distribution for precise mapping between informal hazard descriptions and professional regulatory clauses. Case study results demonstrate that the proposed model performs excellently across key metrics such as Recall@K, providing critical technical support for automated compliance auditing and intelligent safety decision-making in oil and gas enterprises.

Keywords: Oil and gas production safety, Hidden danger identification, Safety regulations, Semantic matching, Contrastive learning, Sentence embedding, Bidirectional attention, Compliance management.

1. Introduction

The oil and gas industry is characterized by high temperatures, high pressures, and explosive hazards, marking it as a typically high-risk industrial sector. Against this backdrop, safety compliance management is critical to ensuring the continuous and stable operation of production systems. In practical production workflows, site managers identify and record issues through routine inspections and mandate rectifications to prevent hazards from escalating into accidents. However, the efficacy of hazard identification and governance is heavily contingent upon the inspectors' proficiency in laws, regulations, industry standards, and internal protocols. Due to the complexity and frequent updates of the regulatory framework in the oil and gas domain, traditional methods relying on human memory and manual retrieval are not only inefficient but also prone to misjudgments or omissions when handling massive hazard records. Consequently, exploring

automated semantic matching technology to achieve precise alignment between hazards and regulations has become an urgent requirement for the digital transformation and intelligent safety management of petroleum enterprises.

With the advancement of Industry 4.0, Natural Language Processing (NLP) has been increasingly applied to safety management. Early research primarily focused on classifying accident reports using rule-based matching or shallow machine learning algorithms, such as Support Vector Machines (SVM) and Naive Bayes. In recent years, scholars have begun exploring deep learning techniques to extract risk entities from unstructured text. Nevertheless, existing studies are predominantly concentrated on "entity extraction" or "sentiment analysis," while research specifically targeting the cross-domain retrieval task of "hazard-regulation" alignment remains relatively scarce.

The alignment of hazards and regulations is essentially a semantic matching task that requires retrieving relevant clauses from a vast regulatory corpus. Traditional BM25 algorithms rely on lexical overlap and fail to capture deep semantic nuances. While Transformer-based pre-trained language models have achieved excellent performance on open-domain tasks, they often suffer from performance degradation when applied to specific vertical domains due to insufficient domain-specific knowledge.

To bridge this gap, the academic community has primarily adopted two paths:

(i) Domain-Adaptive Pre-training (DAPT), which involves performing Masked Language Model (MLM) tasks on massive unlabeled domain-specific corpora to adapt the model to professional terminology.

(ii) Supervised Fine-tuning (SFT), which utilizes high-quality labeled pairs to fine-tune model parameters through specific training objectives, enabling the model to learn the mapping from a general semantic space to a vertical-domain semantic space.

Although Domain-Adaptive Pre-training (DAPT) can significantly enhance a model's representational capacity, its substantial requirements for computational resources and training data make Supervised Fine-tuning (SFT) the primary vehicle for transferring general models to specific domains. Furthermore, although Transformer-based pre-trained models have greatly enriched word-level representations, directly using their output vectors often results in semantically different sentences showing extremely high cosine similarity in the vector space. To overcome this deficiency, contrastive learning has been introduced into the field of semantic matching. The core logic of contrastive learning lies in explicitly optimizing the distribution structure by pulling positive samples closer and pushing negative samples further apart in the vector space.

However, standard contrastive learning frameworks exhibit limitations when processing fine-grained classification tasks within specific domains. Traditional random negative sampling strategies, while computationally efficient, allow the model to easily distinguish between positive and negative samples in the later stages of training. This leads to gradient vanishing problems, making it difficult to further optimize model parameters. Although In-batch Negatives utilize other samples within the same batch as negative samples, the diversity of these negatives is constrained by the batch size. To address these issues, researchers have recently proposed "Hard Negative Mining" strategies, which involve actively identifying samples that are close to the query in the semantic space but are actually mismatched.

Consequently, achieving precise semantic matching between hazard descriptions and regulatory clauses in the specialized vertical field of oil and gas production safety still faces the following critical challenges:

(i) Significant Semantic Style Disparity: On-site inspection records are typically entered manually by operators, characterized by obvious colloquialism and industry-specific abbreviations (e.g., "well control," "confined space"). In contrast, regulatory clauses are rigorously expressed and highly professional and standardized. Such an asymmetric linguistic style increases the difficulty of alignment.

(ii) Dense Distribution of Technical Terminology: The oil and gas domain contains a vast number of specialized terms. General-purpose semantic models, lacking industry background knowledge, struggle to accurately

capture the deep meanings of these terms within a safety context, leading to a sharp decline in matching precision in professional scenarios.

(iii) Scarcity of High-Quality Labeled Data: The efficacy of algorithms like contrastive learning depends on high-quality positive and negative sample pairs. However, in the field of safety management, although hazard records exist, it is difficult to obtain discriminative "hard negatives," which prevents the model from learning fine-grained compliance boundaries..

This study utilizes a real-world production scenario for illustration. An on-site inspection record states: "The local pressure gauge on the ground medium-pressure steam line is malfunctioning, with the pointer exceeding the full scale." This description contains multi-dimensional information, including the location (steam line), equipment (local pressure gauge), and status (exceeding scale). In contrast, the corresponding compliance requirement in a corporate Static Equipment Management Protocol is phrased as: "Discharge pipelines for high-pressure gas or liquid shall be equipped with stable supporting structures... and undergo regular anti-corrosion inspections." Such a matching process requires the model not only to understand synonymous relationships—such as between "fixed support" and "stable supporting structure"—but also to possess the capability for implicit reasoning (e.g., inferring the necessity of "anti-corrosion inspection" from descriptions of "severe corrosion"). Traditional keyword-based matching methods are only effective in scenarios with high literal overlap and fail to address such deep semantic mapping problems where meanings are similar but terminology differs.

To address the aforementioned challenges, this paper proposes a semantic matching model tailored for the oil and gas safety domain. The model aims to achieve deep semantic alignment between hazards and regulations through a contrastive learning framework and introduces a Dynamic Hard Negative Mining (DHNM) mechanism to enhance performance on domain-specific tasks. The overall technical roadmap is illustrated in Fig. 1. Initially, raw field records from oil and gas production are anonymized and standardized to construct a positive sample set of hazard-regulation pairs, while an initial negative sample set is filtered based on semantic vector similarity. During the model training phase, a deep encoder based on a bidirectional attention mechanism is employed to extract text embedding features. A dynamic negative mining strategy is then introduced to update the negative sample pool in real-time according to the model's training state. The InfoNCE loss function is utilized to drive the model toward precise clustering within the semantic space. Finally, the alignment performance between colloquial hazard descriptions and professional regulations is evaluated using the Recall@K metric.

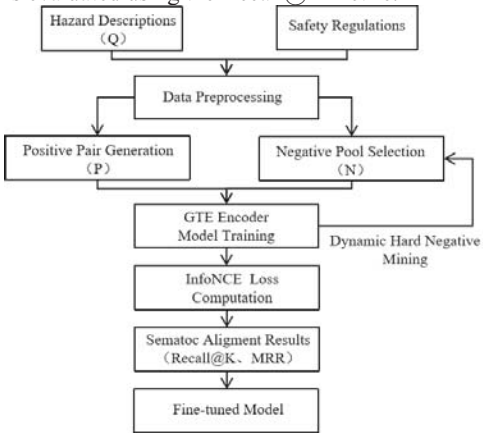

Fig. 1. The overall technical roadmap.

2. Preliminaries

2.1.*Text Embedding Model*

Pre-trained language models, having been trained on large-scale Chinese corpora, possess robust semantic representation capabilities. Through targeted fine-tuning and contrastive learning strategies, these models can effectively extract key feature vectors from hazard descriptions and regulatory clauses within the oil and gas safety domain. In this study, a pre-trained model based on an Encoder-only architecture with a bidirectional attention mechanism is employed as the core for semantic extraction.

For an input hazard or regulation sentence X, the model first generates tokenized representations. To capture long-distance semantic dependencies, a transformer-based multi-head self-attention mechanism is utilized. For the *l*-th layer's hidden state $H^{(l)}$, the

computation process is defined as shown in Eq.(1):

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where Q, K, and V represent the Query, Key, and Value matrices, respectively, obtained through linear transformations of the input vectors, and d_k is the scaling factor. This mechanism allows the model to automatically focus on relevant entities such as "pipeline" or "wall thickness" at the distal end of a sentence when encoding a feature like "corrosion," thereby generating word vectors with global contextual information. Finally, to obtain a fixed-dimensional vector representation \$h\$ for the sentence, a lightweight mean pooling technique is applied after the encoder's output layer, as expressed in Eq.(2):

$$h = \frac{1}{N}\sum_{i=1}^{N} v_i \quad (2)$$

where N denotes the token length of the sentence, and v_i represents the output vector of the i-th token.

2.2.*Contrastive Learning Technology*

In the context of oil and gas regulation matching, a vast majority of irrelevant clauses can be easily distinguished by the model. Consequently, the model struggles to learn complex compliance boundaries from these easy negatives.

To address this, this study incorporates contrastive learning into the training process. The primary objective of fine-tuning is to maximize the cosine similarity between the hazard description vector and its corresponding positive regulation vector, while simultaneously minimizing its similarity with negative samples. We employ the InfoNCE loss, as defined in Eq.(3):

$$\text{InfoNCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\frac{q_i \cdot pos_i}{\tau})}{\sum_{j=1}^{N}\exp(\frac{q_i \cdot neg_{i,j}}{\tau})}\right) \quad (3)$$

where N denotes the number of training samples, q_i is the encoding vector of the i-th query, pos_i represents the positive sample vector for the i-th query, and $neg_{i,j}$ represents the j-th negative sample vector associated with the i-th query. The parameter τ is the temperature coefficient. By maximizing the log-likelihood, the model pulls correlated hazards and regulatory clauses closer in the vector space, thereby achieving semantic alignment between colloquial descriptions and professional standards.

3.Method

3.1.*Data Sources and Cleaning*

The foundational data for this research is derived from authentic safety supervision and inspection records of an onshore oil and gas production site. Since the raw inspection data were primarily entered by frontline operators via mobile terminals, they exhibit prevalent issues such as colloquial expressions, non-standard abbreviations. Consequently, a systematic data cleaning process was conducted:

(i) Data Anonymization: To protect commercial confidentiality and privacy, all sensitive information within the inspection records underwent de-identification. This included replacing specific personnel names, unique equipment IDs, and detailed geographic coordinates with generic placeholders. Specific facility names were substituted with category identifiers, such as "Station A" or "Drilling Platform B."

(ii) Deduplication and Filtering: Redundant entries and invalid records with insufficient word counts or lacking substantive safety semantics were removed. This process resulted in a refined dataset containing tens of thousands of high-quality hazard description texts.

(iii) Structuring of the Regulatory Corpus: In parallel, current national and industry safety standards were compiled and organized. Each regulatory clause was assigned a unique index number and formatted into a standardized textual representation, providing a normative baseline for the subsequent construction of matching pairs.

3.2.*Construction of Hazard-Regulation Positive Samples*

To meet the training requirements of the contrastive learning model, the raw data were transformed into a triplet structure consisting of a query, a positive sample, and negative samples. A single training instance, denoted as $data_i$, is formulated as shown in Eq.(4):

$$data_i = \{query, pos, neg\} \quad (4)$$

where *query* represents the textual description of the on-site hazard, *pos* denotes the corresponding positive regulatory clause, and *neg* signifies the non-correlated negative samples. Positive samples are derived from archived records in the safety supervision system, where the "hazard-regulation" correspondences, verified by safety experts, serve as ground-truth pairs. Negative samples are generated through the selection mechanism proposed in this study.

3.3.*Similarity Interval-based Negative Sample Selection*

Given the specificity of the oil and gas safety domain, random negative sampling often results in "easy negatives," which prevents the model from distinguishing between semantically similar but fundamentally different violations. Therefore, this paper designs a negative sample selection strategy based on semantic similarity intervals:

(i) Vector Representation: A pre-trained text embedding model is utilized to map all clauses in the regulatory corpus into a high-dimensional semantic vector space.

(ii) Interval Filtering Mechanism: The cosine similarity between the *query* and non-correlated regulatory clauses is calculated and ranked by score.

(iii) Hard Negative Initialization: This study specifically selects 20 regulatory clauses with similarities within the [0.4, 0.7] interval as the initial negative samples for a given hazard. Compared to easy negatives with extremely low similarity, samples in this interval exhibit partial overlap with the *query* in terms of vocabulary or scenario description.

This design aims to force the model to learn subtle semantic distinctions in the oil and gas domain from the early stages of training, thereby enhancing its capability for compliance boundary delimitation.

3.4.*Dynamic Hard Negative Mining*

Traditional contrastive learning typically employs static negative samples. This leads to a scenario in the later stages of training where the model can easily distinguish between positive and negative samples, causing the gradient contribution to drop sharply and resulting in low training efficiency. To address this, we propose a state-aware dynamic negative sample update strategy.

(i) Initial State Setting: For each *query*, 20 clauses within the [0.4, 0.7] similarity interval are retrieved from the full regulatory corpus to form the initial negative set N_{init}.

(ii) State Monitoring and Updating: A sliding window of size W=100 steps is defined. Every 100 steps, the system calculates the average similarity score S_{avg} of the current negative set N_{cur} relative to the current model state. The update threshold is set based on the initial average score S_{origin}. The update logic is defined as shown in Eq.(5):

$$Update\_Flag = \begin{cases} True, & \text{if } S_{avg} < S_{origin} \times \alpha \\ Flalse, & \text{otherwise} \end{cases} \quad (5)$$

where α is the decay coefficient (set to 0.9 in this study). When $Update_{Flag}$ is true, it indicates that the current negative samples are no longer sufficiently challenging for the model.

(iii) Hard Negative Resampling: When an update is triggered, signifying that the model has fully learned and identified the features of the current negatives, the system slides the selection window down the initial similarity ranking to select 20 new negative samples to replace the old set. This mechanism ensures a consistent presence of hard negatives throughout the training process, compelling the model to learn increasingly refined semantic boundaries.

4. Experiments

4.1.*Experimental Setup and Evaluation Metrics*

To evaluate the practical effectiveness of the proposed semantic matching model in oil and gas safety compliance management, this study established a standardized experimental benchmark. The experiments were implemented using the PyTorch 2.1 framework on an Ubuntu 22.04 operating system.

(i) Environmental Configuration: The model utilized GTE-zh-small as the foundational encoder. Hardware acceleration was provided by four NVIDIA A100 GPUs (80GB VRAM each). To ensure reproducibility, the random seed was fixed to 42 for all initialization and data shuffling processes.

(ii) Training Hyperparameters: The model was fine-tuned for 10 epochs using the AdamW

optimizer. The learning rate was set to 2e-5 with a weight decay of 0.01. The batch size was set to 64 and the maximum sequence length was truncated to 512 tokens. For the contrastive loss function (InfoNCE), the temperature coefficient τ was set to 0.05. During the dynamic negative mining phase, the decay coefficient α in Equation (5) was maintained at 0.9.

(iii) Dataset Partitioning: From the pool of preprocessed records, 10% were randomly held out as an independent test set, with the remaining 90% used for training and validation (80/10 split). This partitioning ensures that test samples remained entirely unseen during the optimization phase, thereby strictly validating the model's generalization capability.

(iv) Evaluation Metrics: Since the ultimate objective is to recommend the most relevant regulatory clauses to safety managers, Recall@K (K=1, 3, 5) and Mean Reciprocal Rank (MRR) were adopted as the core metrics. These metrics measure the proportion of correct regulatory clauses within the top *K* results and the average rank of the first correct match, respectively, directly reflecting the model's ranking quality in practical business scenarios.

4.2.*Comparative Experiment Design*

Several groups of comparative experiments were conducted to evaluate the performance gains of the deep semantic matching model over traditional methods:

(i) Statistical Matching Model (BM25): Serving as the baseline for traditional text retrieval, this model performs matching based on term frequency statistics. Experimental results indicate that due to the lexical disparity between hazard descriptions and regulatory texts, BM25 exhibits limited effectiveness when processing colloquial expressions.

(ii) Baseline Semantic Model (Vanilla GTE): This involves directly using the general-purpose pre-trained embedding model without any fine-tuning. Although it can capture certain semantic features, it performs poorly in understanding specialized oil and gas safety terminology.

(iii) Proposed Model: This model integrates the Dynamic Hard Negative Mining (DHNM) mechanism with InfoNCE contrastive learning. To further validate the effectiveness of the proposed strategy, we compared the model without dynamic mining (w/o Dynamic Mining) against the full model (Full). The experimental results demonstrate that by continuously introducing challenging negative samples throughout the training process, the model's ability to learn compliance boundaries is significantly enhanced.

4.3.*Experimental Analysis and Discussion*

The results of the comparative experiments are summarized in Table 1.

Table 1. Statistical results of the comparative experiments.

Model	Recall @1	Recall @3	Recall @5	MRR
BM25 (Keyword-based)	0.321	0.458	0.512	0.402
Vanilla GTE-zh-small	0.584	0.723	0.789	0.665
Proposed Model (w/o Dynamic Mining)	0.742	0.856	0.894	0.798
Proposed Model (Full)	**0.805**	**0.902**	**0.924**	**0.852**

By comparing the performance of different models on the test set, several key conclusions can be drawn:

(i) The Prerequisite of Deep Semantic Embedding: Compared to the keyword-based BM25, embedding-based models achieved a performance gain of over 40% across all metrics. This proves that in the oil and gas safety domain, semantic alignment between hazard descriptions and regulatory clauses must transcend literal matching and delve into the deep contextual environment.

(ii) Enhancement in Matching Precision: The proposed model significantly outperformed the baseline models in the Recall@1 metric. The ablation experiments demonstrate that the introduction of the Dynamic Hard Negative Mining (DHNM) mechanism leads to more stable convergence in the later stages of training. This mechanism effectively mitigates the overfitting issues caused by overly simplistic negative samples.

(iii) Practical Application Potential: The proposed model achieved a Recall@5 of 92.4%. In real-world business scenarios, this implies that safety managers have an extremely high probability of finding the accurate compliance

basis by reviewing only the top 5 recommendations provided by the model. This significantly reduces the workload associated with manual retrieval and demonstrates substantial practical implications for industrial safety management.

4.4.*Case Study*

To further illustrate the practical efficacy of the proposed model compared to the traditional BM25 model, we conducted a qualitative analysis of matching results in real-world production scenarios.

Case 1: Misalignment due to keyword bias.

- Hazard Description: "Gas cylinders in the doghouse on the drill floor are missing anti-vibration rings."
- BM25 Result: Erroneously matched a clause regarding "drill floor handrail height," primarily because the keyword "drill floor" was assigned disproportionately high weight.
- Proposed Model Result: Successfully matched the relevant clause in the *Safety Technology Administration Regulation for Gas Cylinders* regarding "regulations for gas cylinder accessories and anti-vibration rings."

Analysis: This demonstrates that through the attention mechanism, the proposed model effectively suppresses the interference of situational locators (e.g., "drill floor") and accurately captures the semantic features of core entities (e.g., "gas cylinders" and "anti-vibration rings").

Case 2: Identification of latent logical associations.

- Hazard Description: "The work permit for confined space operations has not been approved."
- Proposed Model Result: In addition to matching the primary regulations for work permit management, the model also recommended clauses related to "gas testing in confined spaces" (ranked 3rd).

Analysis: This proves that the model has learned the logical interconnectivity inherent in safety management workflows. It can provide safety managers with extended compliance recommendations, moving beyond literal matching to identify functionally related safety requirements.

5. Conclusion

This paper proposes and implements a contrastive learning-based semantic matching model to address the challenge of intelligently aligning colloquial hazard descriptions with professional regulatory clauses in the oil and gas safety domain. The primary conclusions are as follows:

(i) Efficacy of Semantic Alignment: Fine-tuning a deep embedding model (e.g., GTE-zh-small) with a bidirectional attention mechanism and InfoNCE loss effectively captures semantic correlations across disparate linguistic styles, significantly outperforming traditional statistical matching methods.

(ii) Innovative Value of DHNM: The proposed dynamic negative update mechanism selects hard negatives via real-time score monitoring and a sliding window. This successfully overcomes performance bottlenecks in late-stage training and substantially improves key metrics such as Recall@1.

(iii) Industrial Application Prospects: Case studies demonstrate that the model exhibits significant potential in enhancing compliance auditing efficiency and provides essential technical support for intelligent HSE management systems.

Despite the promising results, this study has certain limitations. First, the model's performance is highly dependent on the quality of the initial regulatory corpus; highly ambiguous or overlapping clauses may still pose challenges for precise disambiguation. Second, while the DHNM strategy enhances accuracy, it introduces additional computational overhead during the training phase. Furthermore, as the dataset is specifically tailored to the oil and gas industry, its generalizability to other industrial sectors requires further validation.

Future research will focus on exploring multimodal matching technologies that integrate on-site imagery with textual descriptions to achieve a more comprehensive and objective assessment of hazard compliance. To ensure the reproducibility of this work, the core algorithm architecture and hyperparameter configurations

have been standardized and detailed in the experimental section of this paper.

Acknowledgments

This work was supported by the following research projects:

- China National Petroleum Corporation (CNPC) Scientific Research Projects (Nos. 2025ZG82, 2023DJ6508);
- Scientific Research Project of the CNPC Research Institute of Safety and Environmental Technology (No. 25AQHBSC006);
- PetroChina Company Limited Scientific Research Project (No. 2024ZZ57).

References

Aven, T. (2016). Risk Assessment and Risk management: Review of Recent Advances on Their Foundation. European Journal of Operational Research, [online] 253(1), pp.1–13.

Chen, Y., Zhai, J., Wu, S., Tian, K. and Song, X. (2025). Knowledge Graph Construction of Large Language Model Retrieval Augmented Generation for Oil and Gas HSE Professional Q&A System. 35th European Safety and Reliability Conference (ESREL 2025) and the 33rd Society for Risk Analysis Europe Conference (SRA-E 2025), pp.1200–1207.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. and Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. Findings of the Association for Computational Linguistics: EMNLP 2020.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North, [online] 1, pp.4171–4186.

Gao, T., Yao, X. and Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. [online] ACLWeb.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. and Smith, N.A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. [online] ACLWeb.

He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. Thecvf.com, [online] pp.9729–9738.

Lison, P., Pilán, I., Sanchez, D., Batet, M. and Øvrelid, L. (2021). Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Reimers, N. and Iryna Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.

Tixier, A.J.-P. ., Hallowell, M.R., Rajagopalan, B. and Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. Automation in Construction, 62, pp.45–56.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, [online] 30, pp.5998–6008.

V.V. Kuleshov, T.G. Aygumov, A.L. Zolkin and O.V. Saradzheva (2024). Assessing and managing risks in the oil and gas industry with machine learning and digital technologies. E3S web of conferences, 474, pp.02011–02011.

Wang, J., Qu, Z., Wu, S., Skitmore, M. and Ma, L. (2025). Automatic classification of construction accident reports using BERTopic-GLDA approach. Engineering Construction & Architectural Management, pp.1–30.

Wu, M., Mosse, M., Zhuang, C., Yamins, D. and Goodman, N. (2020). Conditional Negative Sampling for Contrastive Learning of Visual Representations. ArXiv (Cornell University).

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J. and Overwijk, A. (2020). Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. arXiv:2007.00808 [cs]. [online]