MDIR-BERT: A Multi-Dimensional Retrieval-Enhanced Language Model for Power Audit Text Understanding

Jia Xiao-liang[1], Li Sen[1], Cui Xia[1], Li Jing[2], Sun Chang-peng[1], Liu Dong-hua[2], Chen Zheng-long[2]
[1]State Grid Tianjin Electric Power Company, Tianjin, 300010, China
[2]State Grid Tianjin Electric Power Company Chengdong power supply branch, Tianjin, 300250, China
E-mail: jxlseuee@163.com
*Corresponding author

Keywords: Power audit text, multi-dimensional information retrieval, large language model (LLM), audit category classification, multi-dimensional information retrieval-based bidirectional encoder representations from transformers (MDIR-BERT)

Received: May 1, 2025

In the rapidly evolving energy sector, efficient access to relevant information from power audit reports is crucial for informed decision-making, regulatory compliance, and operational improvements. However, the intricate language, complex vocabulary, and unstructured format of power audit texts present significant challenges for conventional information retrieval techniques. To address these issues, the research proposes a novel power audit text understanding technology that combines multi-dimensional information retrieval enhancement with a domain-adapted Large Language Model (LLM) to enhance the performance of power audit text processing. The Multi-Dimensional Information Retrieval-based Bidirectional Encoder Representations from Transformers (MDIR-BERT) method captures electric-power-specific morphology, domain-specific vocabulary, and intricate entity relationships more effectively. MDIR-BERT is pre-trained on a huge quantity of electric power audit transcripts utilizing both word-level and entity-level covered language modeling tasks. The model is trained on a curated dataset of annotated electric power audit documents sourced from regulatory and industrial environments. MDIR-BERT integrates domain-specific pre-training with both word-level and entity-level masked language modeling, capturing electric power-specific morphology, terminology, and complex entity relationships. The data preprocessing steps include comprehensive text cleaning, normalization, and tokenization to ensure high-quality input for method training. Experimental results show that MDIR-BERT achieves a classification accuracy of 98.82%, representing a +16.86% improvement over the baseline EPAT-BERT model (81.96%), along with notable gains in precision, recall, and F1-score. These findings highlight the effectiveness of integrating enhanced information retrieval techniques with specialized language modeling for the intelligent understanding of power audit documentation, paving the way for more accurate, scalable, and interpretable audit methods.

Povzetek: MDIR-BERT, izboljšan jezikovni model s večdimenzionalnim iskanjem informacij (MDIR), je razvit za razumevanje revizijskega besedila elektroenergetike. S predhodnim usposabljanjem na besedni in entitetni ravni dosega kvalitetno klasifikacijo revizijskih kategorij.

1 Introduction

The development of Information Retrieval (IR) technology has been intimately linked to the human need for information access. In recent years, IR and associated product systems have expanded significantly as a critical constituent of smart data dispensation tools. The basis of IR technology is the identification of documents related to the customer's search from a big and unorganized collection, which usually leads to a graded catalog of the documents by significance and user requirements [1]. IR plays an essential role in numerous real-world functions, like expert finding, digital libraries, and Web search. IR essentially refers to the task of retrieving information resources related to information required from a large collection of resources [2]. However, user intentions were more complex than simply retrieving information based on similarity [3]. This audit is conducted by a qualified firm with the necessary competencies in line with the requirements established by the Ministry of Energy and Mineral Resources. These criteria apply to businesses or industries that utilize a significant amount of energy. A complete audit evaluates all areas of energy usage, from fuel consumption to the use of generated electrical energy [4]. Lowering electricity costs and cutting down energy waste requires an energy audit. Efforts have to be initiated by governments to require periodic energy audits for industrial buildings. An energy audit is a great way to find the best solution and

assess how much energy a building uses [5]. The generative probability of word sequences, or more generally, the ability to predict forthcoming words conditional on prior words, is a crucial function of language models (LM). LMs were first created for text creation, but they are also being studied for reformulating
a range of NLP issues into different text-to-text challenges in the text of electric power audits [6].
The implementation of Large Language Models (LLMs) marks the most important change in the technical development of electric power audit text [7]. LLMs mark a substantial advancement in Artificial Intelligence (AI) as it makes breakthroughs in generalization and adaptability across tasks, but LLMs generate inaccurate information, misalign with temporal information, struggle to keep context, and struggle to fine-tune each response, leading to serious issues regarding reliability when applied to electric power audit text [8]. In the continually changing energy industry, timely access to essential information from power audit reports is critical for making informed decisions, conforming to regulations, and improving operations. Conventional BERT-based models are not effective in encoding the sophisticated, domain-specific semantics in electric power audit reports. There is a demand for models incorporating domain knowledge and sophisticated retrieval methods to enhance classification and information extraction accuracy. This research explores a new technology for understanding power audit reports that improves multi-dimensional IR and domain-adapted LLM performance by extracting morphology specific to electric power, domain-specific language, and complexities of entities to use the Multi-Dimensional Information Retrieval-based Bidirectional Encoder Representations from Transformers (MDIR-BERT) method.

1.1 Key contributions

- This research aims to develop a multi-dimensional information retrieval for improved classification and understanding of electric power audit texts.
- Initially, Electric power audit reports from energy-intensive sectors, which are obtained from publicly accessible databases from Kaggle, represent various regulatory and operational contexts.
- Utilized preprocessing steps such as stop word elimination, lemmatization, and tokenization to preprocess and normalize intricate technical jargon for optimal model input.
- MDIR-BERT by pre-training on the electric power audit dataset with word-level as well as entity-level masked language modeling to encode domain-specific terminologies and intricate entity relationships.
- Obtained a classification accuracy of 98.82%, representing a +16.86% relative improvement compared to the baseline EPAT-BERT model, in addition to significant boosts in precision, recall, and F1-score.

1.2 Research questions

RQ1: Can a domain-adapted BERT model (MDIR-BERT) enhanced with multi-dimensional information retrieval outperform general-purpose BERT (EPAT-BERT) in power audit text classification?
RQ2: How does multi-dimensional information retrieval improve entity recognition and contextual understanding in regulatory audit texts?
RQ3: What impact does domain-specific pretraining have on the performance of language models in complex, unstructured audit document processing?
The research outline is organized as follows: Section 2 reviews related research, while Section 3 outlines the research methodology. Section 4 presents the results and discussion, and Section 5 concludes the research.

2 Related work

The transformational effects of LLMs on IR research were investigated in the research [9]. The method comprised synthesizing findings from a strategy workshop organized by the Chinese IR community. It suggested a new IR technological paradigm involving IR models, LLMs, and humans, but faces computational trade-offs, trustworthiness concerns, domain boundaries, and implications. An analysis of e-commerce customer reviews on drum washing machines using Robotic Process Automation (RPA) was demonstrated [10]. It combined ROST Content Mining System 6 (ROSTCM6) and LOGCONTROL-BLOCK systems to extract sentiment and correct audit robot paths. While effective in revealing customer sentiments and guiding e-commerce strategies, limitations include reliance on predefined keywords and the need for improved automated sentiment analysis accuracy.
The Mistral 8x7B LLM's current Mixture of Experts (MoE) architecture was combined with Retrieval Augmented Generation (RAG) to improve on challenging IR and reasoning tasks, which were investigated in [11]. In the quantitative and qualitative evaluation of the model using the Google BIG-Bench dataset, notable gains were observed in F1 score, accuracy, precision, and recall. Limitations include computing needs and dataset breadth. Integrating LLMs with Knowledge Graphs (KGs) enhanced intelligent fault detection and IR for new energy vehicles (NEVs) [12]. It developed an intelligent fault retrieval system, a structured knowledge graph, and an optimized BERT model for fault classification,

demonstrating exceptional performance in Q&A situations for NEVs, but facing scalability issues.

To evaluate the Word to Vector (Word2Vec) model for document compliance detection by comparing it with Term Frequency–Inverse Document Frequency (TFIDF), Latent Dirichlet Allocation (LDA), and Bidirectional Encoder Representations from Transformers (BERT) as described [13]. Results showed that Word2Vec effectively captures semantic similarity with higher efficiency and simplicity. However, it performs slightly lower than BERT in handling complex semantics and domain-specific terminology. A self-retrieval framework [14] that leverages self-supervised learning was developed to improve retrieval efficiency and model simplicity. It internalized a retrieval corpus, improved downstream LLM applications, and outperformed conventional IR systems. However, it faced high computing costs and scaling challenges, despite maintaining real-time efficiency and cross-domain generalizability. Predictive Analytics (PA) in Current Research Information Systems (CRIS) to predict research trends through machine learning is used [15]. In this research, k-Nearest Neighbor had the best performance. Limitations include moderate AUC scores and dependence on historical metadata to generate predictions. The Financial BERT (FinBERT) model, specialized in the finance industry, was developed to enhance sentiment analysis in financial writings [16]. FinBERT model outperformed traditional dictionaries in classifying context-dependent sentiment and Environmental, Social, and Governance (ESG)-related talks with minimal training data, but faced limitations in domain-specificity and potential decreased generalizability. The use of LLMs in auditing was investigated in [17], with an emphasis on compliance checks and report production. LLMs effectively handle unstructured data, address compliance concerns, and provide excellent audit reports, despite challenges like data security and model interpretability. The research [18] enhanced LLM privacy audits by creating more robust sequences that allow for more successful membership inference assaults under realistic threat models. It demonstrated a significant improvement in detection and True Positive Rate (TPR) with optimal sequences, achieving a 49.6% TPR on Qwen2.5-0.5B, compared to 4.2% earlier, but has drawbacks due to reliance on model access without shadow models or gradient insertion. To compare forecasting MASI trends with ARDL with trend and seasonality (Long short-term memory (LSTM), and extreme gradient boosting (XGBOOST) was determined [19]. ARDL, with trend and seasonality, returns the lowest MAPE, at 26.7%. Limitations include LSTM and XGBOOST executing higher error rates and taking longer to process. The Two Sliding Windows Graph Neural Network (TSW-GNN) architecture for text classification was introduced, which works around limitations of corpus-level graph approaches that suffer from continuous memory usage and are completely contextually agnostic, was introduced [20]. The TSW-GNN model addresses this issue by introducing TSW into the GNN architecture with a new dynamic global sliding window and a new dynamic local sliding window, increasing contextual memory and representation of semantics. Tests from the seven datasets reveal that the classification accuracies were improved, though at increased complexity of the two sliding windows and their associated GNN parameters. To explore the independent role of internal auditors at the Swedish Police Authority and to illustrate their relational struggles within the organization was described [21]. The research adopts a narrative framework in the study of auditor independence and introduces stories of auditors highlighting psychological distress, ambiguity in legitimacy, and attempts to negotiate competing demands. The picture painted by these narratives can be viewed as a tragedy where auditors were unable to resolve tensions that manifested themselves as professional dilemmas. Results showed LLMs perform well in noise handling but struggle with falsehood management. An overview of the related work is given in Table 1.

Table 1: Overview of the related works

Ref. No.	Objective	Task Type	Domain	Model Used	Method	Limitations
Ai et al., [9]	Investigate the role of LLMs in IR research	Information Retrieval	General	Not specified	Strategic workshop proposing IR-LLM-human paradigm	Computational trade-offs, trust concerns, and ethical issues
Sun and Huo, [10]	Analyze e-commerce reviews using automation	Sentiment Analysis	E-commerce	RPA, ROSTCM6, LOGCONTROL-BLOCK	Keyword extraction, path correction, sentiment classification	Relies on predefined keywords, limited sentiment accuracy

Xiong and Zheng, [11]	Improve IR and reasoning	IR + Reasoning	General	Mistral 8x7B with RAG	RAG + Mixture of Experts evaluated on BIG-Bench	High computing needs, limited dataset
Zhang et al., [12]	Enable intelligent IR for NEVs	Classification + Retrieval	New Energy Vehicles	Optimized BERT + KG	Fault classification using KG-enhanced BERT	Scalability issues
Wen et al., [13]	Evaluate Word2Vec for document compliance detection	Document Similarity	Legal, Audit	Word2Vec, TFIDF, LDA, BERT	Semantic similarity via vector models	Slightly lower performance than BERT in complex semantics
Tang et al., [14]	Merge IR functionality within a single LLM	IR	General	Self-Retrieval LLM	Self-supervised corpus-internal IR	High computational cost, scaling complexity
Azeroual et al., [15]	Predict research trends in CRIS	Trend Forecasting	Research Management	kNN, SVM, Random Forest	Predictive analytics with machine learning	Moderate AUC, dependent on historical metadata
Huang et al., [16]	Domain-specific sentiment analysis	Sentiment Classification	Finance	FinBERT	Domain-adapted BERT for financial sentiment	Limited generalizability
Gan, [17]	Automate audit compliance	Report Generation + Classification	Auditing	LLM-based	Process unstructured audit data for reporting	Data security, interpretability
Panda et al., [18]	Enhance privacy auditing	Membership Inference	General	Qwen2.5-0.5B	Robust canaries for audit testing	Requires model access, no shadow models
Oukhouya et al., [19]	Compare forecasting models for MASI trends	Financial Forecasting	Stock Market	ARDL, LSTM, XGBOOST	Time series modeling with trend and seasonality	Higher error and processing time in LSTM and XGBOOST
Li et al., [20]	Improve text classification with sliding windows	Text Classification	NLP	TSW-GNN	Local and global sliding window graph construction	Increased model complexity and parameter tuning
Nordin et al., [21]	Explore internal auditors' independence challenges	Organizational Behavior Analysis	Public Sector / Audit	Narrative Framework	Story-based analysis of auditor roles	Unresolved tensions, emotional strain, and ambiguous legitimacy

Existing approaches have various difficulties, including computational complexity, ethical problems, scaling challenges, decreased accuracy, a limited dataset scope, low generalizability, data security threats, and reliance on embedding quality. These constraints impede real-time, domain-specific, and reliable information retrieval in specialist sectors, such as electric power auditing. To address these issues, the research explores a new technology for understanding power audit reports that improves multi-dimensional IR and domain-adapted LLM performance by extracting morphology specific to electric power, domain-specific language, and complexities of entities to use the MDIR-BERT Model.

3 Multi-dimensional information retrieval (MDIR)

MDIR is an advanced retrieval technique that expands the traditional keyword-based search into a multidimensional framework, incorporating semantic meaning, contextual relevance, domain-specific lexicon, relationships between entities, and user intent, thereby enabling retrieval that is precise and comprehensive. MDIR increases the power of a text understanding process for audit text, allowing for the MDIR-BERT model to better capture the complex, technical, unstructured nature of power audit documents and their underpinning. This section gathers the electric power audit text data and preprocesses the data using techniques, such as data cleaning using Stop Words Removal and data normalization using lemmatization and tokenization. Finally, classification and information retrieval were performed using BERT. Figure 1 depicts the System Design of the MDIR-BERT Model.

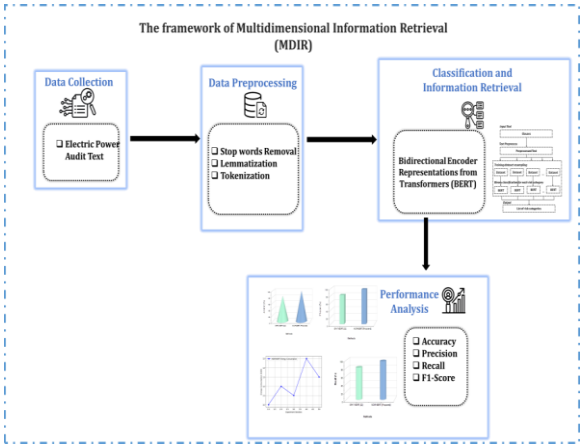

Figure 1: System design of the multi-dimensional information retrieval-enhanced BERT model

3.1 Data collection

The data is obtained from the Kaggle link: https://www.kaggle.com/datasets/zoya77/power-audit-report-and-entities-dataset. The dataset comprises 1,001 audit report entries collected from Kaggle. Each entry includes an audit report ID, audit text, a list of extracted named entities, and a category label. The technical power audit reports cover equipment, energy systems, and compliance, supporting tasks like entity recognition and classification across categories such as safety, efficiency, and regulation. Audit texts range from 15–40 tokens, averaging around 25 tokens. Entities cover standard equipment (e.g., Load Balancer) and locations (e.g., Control Room), enabling comprehensive analysis of energy systems. The obtained dataset is split in 80:20 ratios for training and testing performance.

3.2 Data preprocessing

Data preprocessing is the procedure of converting fresh data into an organized and cleansed form to improve model performance. It cleans, normalizes, and tokenizes electric power audit texts to provide high-quality input for model training while also improving classification and information retrieval accuracy. It includes approaches, such as stop word removal, lemmatization, and tokenization, to arrange power audit texts in a structure that reduces noise while providing high-quality information. It allows for efficient and accurate IR, entity recognition, and classification operations in the system.

3.2.1 Data cleaning using stop words removal

Data cleaning is the process of removing unnecessary or noisy elements from raw data to make it more accurate. It is utilized to eliminate extraneous or noisy information, resulting in high-quality input for training and improved overall performance of electric power audit classification. In the research, stop word removal reduces frequent, unnecessary words from audit text so that the model focuses on important content for better IR and classification. This is produced by establishing a frequency threshold. This threshold was simply set as the average frequency of all terms gathered for the language in Equation (1).

$$\sigma = \frac{\alpha}{n}\sum_{j=1}^{n} t_j \qquad (1)$$

Where t_j is the frequency of the j^{th} term, Equation (1), α is defined as a smoothing adjustment factor to 1.25, empirically validated in validation experiments to moderately increase the average threshold and dampen noise from low-frequency terms. This value $\frac{\alpha}{n}\sum_{j=1}^{n} t_j$selected to optimize in entity recognition by preventing inclusion of excessively rare or excessively common terms.

3.2.2 Data normalization using lemmatization

Normalization refers to the process of converting text to a uniform state, often by reducing words to their standard forms or original structures. The normalization process allows the different variations of words to be standardized,

which permits the method to better process and comprehend province-precise language in power audit texts. This process of normalization helps to standardize variations of words to allow for treating different versions of a word as equivalent terms. Lemmatization helps to determine the organizational meanings of words, which assists in the analysis of text, and naturally, the processing of this text. It is valuable in many text analysis projects, especially those focusing on IR, sentiment, and text classification.

3.2.3 Tokenization

It is the procedure in which input text is divided into minor units of meaningful units (tokens), which can be meaningful individual words, phrases, or sentences. Tokenization is a key step towards breaking down the raw power of the audit text into portions that will ultimately be meaningfully analyzed by the model. By tokenizing text, the model can better interpret the relationships, structure, and contextualization of words. Tokenization will be used to confirm the conducting of tasks, like IR and entity recognition, where tokens are identified and labeled. It enhances the ability of the model to yield valuable and informative information from sophisticated and complex unstructured audit documents.

3.3 Classification and information retrieval using bidirectional encoder representations from transformers (BERT)

Classification is the process of assigning text data into predefined categories based on its content, and IR is the task of finding and extracting significant data from a huge collection of structured or unstructured information. In the research, classification helps to organize and label power audit texts into specific, meaningful categories for easier analysis, while IR enables quick and accurate extraction of relevant insights from large volumes of audit documents to support informed decision-making. These methods are boosted by BERT, which captures deep contextual power and meaning of the text to enhance classification accuracy as well as retrieval precision. After tokenization, BERT uses the electric power audit text to gather contextual relations for accurate classification. It further enhances IR through accurate detection and retrieval of audit-specific features and patterns.

3.3.1 Overview of MDIR-BERT

MDIR-BERT is based on the basic architecture of BERT (Bidirectional Encoder Representations from Transformers), which encodes the bidirectional context of words in a sentence through self-attention mechanisms. This helps the BERT model comprehend word semantics concerning the previous and next words, and hence, BERT is very effective for tasks like classification and information retrieval. While general BERT is pre-trained with the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks on general corpora, MDIR-BERT takes this further by adding domain adaptation for the electric power audit domain. In particular, MDIR-BERT is additionally pre-trained on a massive dataset of electric power audit transcripts to capture more domain-specific vocabulary, morphological forms, and intricate named entity relations. To facilitate this domain-specific adaptation, two domain-specific pre-training tasks are utilized: Word-Level Masked Language Modeling (W-MLM), likewise the standard MLM, but with modifications to focus on domain tokens that usually appear within audit texts, including audit procedures, voltage types, compliance, and equipment-related terms. Entity-Level Masked Language Modeling (E-MLM): This task entails masking named entities determined through a domain-tuned NER system and having the model predict them in their respective contextual environments. This assists MDIR-BERT in capturing hierarchical and relational dependencies between domain-specific entities more effectively. With these enrichments, MDIR-BERT gains a better grasp of electric-power-specific semantics and structure for more accurate and context-sensitive classification and retrieval.

3.3.2 BERT for classification

BERT processes input text using its transformer layers while performing categorization jobs. After the text has been analyzed, the output representation is sent through a classification head to forecast the text's proper category, as shown by Equation (2).

$$Output_{class} = Softmax(Dense(BERT(Input))) \quad (2)$$

Where $BERT(Input)$ represents the BERT model processing the input text, $Dense$ is the classification layer, and $Softmax$ is used to transform logits into probabilities for classification.

3.3.3 BERT for information retrieval

BERT is used in IR to discover the documents that are relevant to a given query. BERT recognizes the context of a query and a group of documents, which improves retrieval accuracy. BERT's bidirectional nature assists in identifying more semantically relevant documents even when keywords fail to match perfectly, as shown by Equation (3).

$$Relevance\ Score = Similarity(BERT(Query), BERT(Document)) \quad (3)$$

Where $BERT(QUERY)$ and $BERT(Document)$ are the query and document's context-aware embeddings, respectively.

3.3.4 Fine-Tuning BERT

Fine-tuning is the method of modifying the pre-trained BERT model to a precise goal, such as classification or IR, by training it on a labeled dataset. This involves adapting BERT's weights by the task requirements, enabling it to learn domain-specific jargon and nuances represented by Equation (4).

$$Loss = \sum_{j=1}^{N} Cross - Entropy(True_j, Predicted_j) \quad (4)$$

Where $True_j$is the definite tag for the j^{th} model, $Predicted_j$ is the forecasted tag for the j^{th}sample, and $Cross - Entropy$ is the loss function used during training. BERT has the advantage of considering the complete background of words in a phrase, which significantly enhances classification and IR efficiency. This two-way context enables BERT to recognize subtle semantic links, making it particularly useful for processing complex and specialized language in power audit reports. Furthermore, due to its pre-training on great corpora and fine-tuning capabilities, BERT could be trained to perform specific tasks with less data.

4 Results and discussion

The research objective is to improve electric power audit text categorization and IR performance by introducing a new MDIR-BERT model. The experimental setting and performance assessment measures used in the research improve the electric power audit text categorization and IR performance.

The experiments were run on a machine with an Intel Core i7 processor, 32GB RAM, and an NVIDIA RTX 3080 graphics card. The models were run in Python 3.9 with PyTorch as the base, with the BERT model developed atop this framework.
The model proposed took around 4.2 hours for training, using 4.2 GPU hours. It comprises about 110 million parameters and occupies a storage size of 420 MB. All the models were trained under the same settings to ensure fairness when comparing the process.

4.2 Hyper-parameters

Table 2 represents the hyperparameters utilized in the power audit text understanding research.

Table 2: Hyperparameters

Hyperparameter	**Value**
Learning Rate	$2e-5$
Batch Size	32
Number of Epochs	30
Optimizer	AdamW
Warmup Steps	500
Max Sequence Length	128
Gradient Clipping	1.0
Weight Decay	0.01
Random Seed	42
Dropout Rate	0.1

4.3 Performance metrics

The Performance Metrics, including Execution time, Energy consumption, and speed of convergence, are utilized to enhance the performance of electric power audit text classification.

Energy consumption

Energy consumption refers to the energy needed by a model to execute inputs and generate outputs. It is an important metric in energy-limited systems, like internet-enabled edge devices or mobile devices. Lower energy usage renders the system more efficient and sustainable, particularly for large-scale AI deployments. Figure 2 depicts the Energy Usage seen in the MDIR-BERT Model Execution.

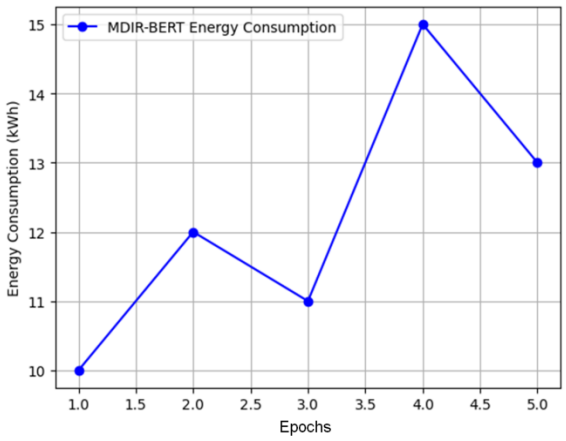

Figure 2: Energy Consumption Observed in the MDIR-BERT Model Execution

MDIR-BERT model energy consumption or moderate consumption rates varied between 10 and 15kWh over five test repetitions, which converts to moderate consumption of resources. In the central sets of repetition, there was an increase in utilization, attributed to the complexity in processing or the size of the data. The model's average utilization was more uniform and efficient, demonstrating its potential for real-world applications. For epoch 1, the model reaches 10kWh, 12kWh in epoch 2, 11kWh in epoch 3, 15kWh in epoch 4, and 13kWh in epoch 5. The proposed MDIR-BERT method shows extreme performance in epoch 4 with 15kWh.

Execution time

Execution time refers to the number of times it takes a model to consume an input, process it, and produce an output. It is a significant metric for real-time or time-sensitive applications, like autonomous systems or internet applications. Lower execution times are preferable so that users can experience the best, and the system's efficiency is enhanced overall. Figure 3 illustrates the visualization of the MDIR-BERT's execution time.

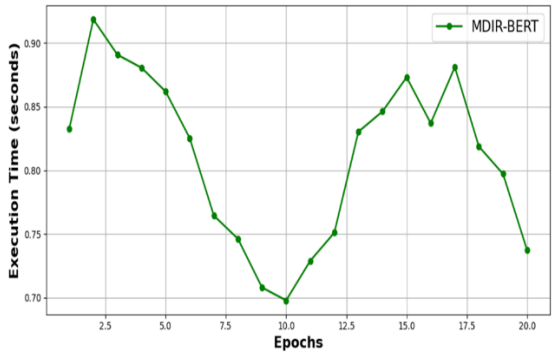

Figure 3: Visualization of the execution time of the MDIR-BERT

The execution time of the MDIR-BERT model replicates its performance over 20 epochs with moderate variances due to computation and environmental conditions. The execution duration varies around an average of 0.8 seconds, with peaks reaching roughly 0.89 seconds and troughs around 0.72 seconds, driven by a sinusoidal pattern and small random noise.

Accuracy and loss

Accuracy is the number of correct predictions made by a classical model to the total number of predictions, whereas loss is the difference between expected and actual values, which measures how well the model performs throughout training. The loss curve shows how the model converged during training, with lower values representing better performance, while the accuracy curve shows how well it captures electric-power-specific morphology, domain power, and intricate entity relationships more effectively. The accuracy and loss characteristics of the training for the MDIR-BERT technique are shown in Figure 4.

The resulting MDIR-BERT model demonstrates good performance: training loss goes down from 0.95 to almost 0.01 after 30 epochs, and training accuracy increases steeply from 0.1 to about 0.97, which indicates good convergence and high learning efficiency.

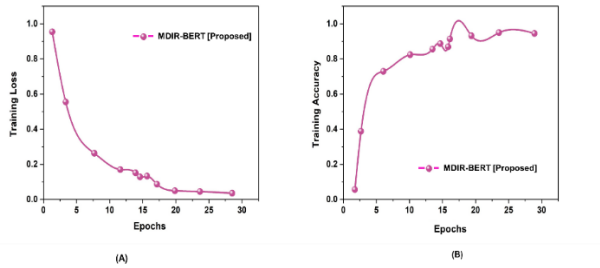

Figure 4: Graphical outcome of (a) loss and (b) accuracy

Statistical Significance

The confidence interval for model accuracy is the normal distribution curve, where the shaded region highlights the most likely accuracy range. It visually represents the reliability and accuracy of the model's performance estimate. Figure 5 shows the Graphical outcome of a 95% confidence interval for accuracy (MDIR-BERT).

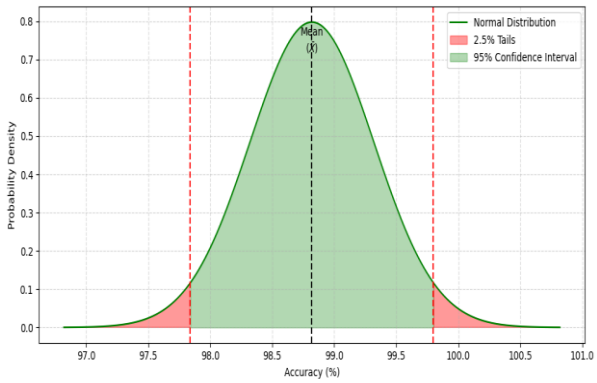

Figure 5: Graphical outcome of 95% confidence interval for accuracy (MDIR-BERT)

The image shows a 95% confidence interval for the accuracy of MDIR-BERT, represented as a normal distribution curve. The x-axis indicates accuracy percentages ranging from 97.0% to 101.0%. The shaded area under the curve represents the 95% confidence interval, meaning there is a 95% probability that the true accuracy of the model lies within this range. The 2.5% tails on either side of the distribution are excluded, highlighting the central 95% region. The peak of the curve corresponds to the most probable accuracy value, with the density decreasing as values move away from the center.

Confusion Matrix

A confusion matrix compares the expected and actual values for a dataset to show the effectiveness of a classification model (Figure 6). The confusion matrix shows how well MDIR-BERT classified data in five different power audit categories. Considering its high overall prediction accuracy, the model occasionally

misclassifies objects, especially between closely similar classes like "Energy Efficiency" and "System Upgrade." Through multi-dimensional information retrieval and domain-adapted language modeling, this demonstrates the domain complexity and the model's efficacy in comprehending electric power audit material.

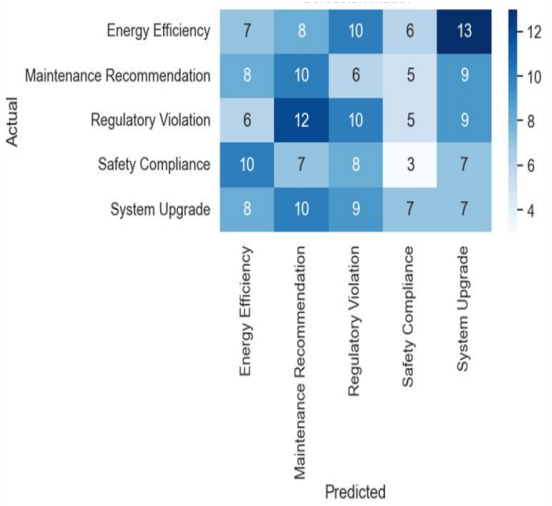

Figure 6: Confusion matrix of MDIR-BERT performance

Precision-recall curves

A binary classification model's effectiveness is represented graphically by a Precision-Recall curve, which is particularly beneficial for unbalanced datasets (Figure 7). The precision-recall curve validates MDIR-BERT's efficacy in domain-dependent audit text understanding by demonstrating its superior classification performance in power audit categories, with an average accuracy. Whereas the energy efficiency is 0.89, the maintenance recommendation is 0.91, the regulatory violation is 0.96, safety compliance is 0.97, and the system upgrade is 0.89.

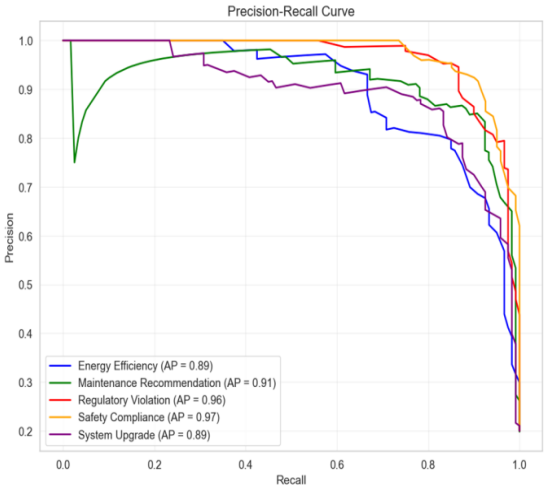

Figure 7: Efficiency of MDIR-BERT with precision-recall curves

The research evaluates the macro/micro F1. Evaluating all classes equally, Macro F1 computes the F1 score for each class separately before the results. To create a single F1 score, Micro F1 provides all the real positives, incorrect positives, and incorrect negatives for each class, which gives each occurrence equal weight. The proposed model demonstrates 0.964 of macro F1 and 0.963 of micro F1.

4.4 Comparison phase

The performance metrics used to compare the performance of electric power audit text classification are accuracy, F1-score, recall, and precision. The MDIR-BERT was compared with the existing methods like Text Convolutional Neural Networks (Text CNN) [22], BERT [22], and Electric Power Audit Text–BERT (EPAT-BERT) [22].

Accuracy

Accuracy: Accuracy indicates how well the model accurately recognizes relevant and irrelevant information in the power audit text. To indicates the ratio of correct to incorrect predictions across all cases, giving a total view of classification performance across different audit documents. Accuracy measures the proportion of all correct power audit text classifications performed by a model. It can be helpful in assessing overall MDIR-BERT's performance. Table 3 depicts the accuracy of the MDIR-BERT.

Table 3: Performance summary of MDIR-BERT

Methods	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**F1-score (%)**
Text CNN [22]	71.65	69.01	74.27	71.56
BERT [22]	77.91	77.94	78.23	78.08
EPAT-BERT [22]	81.96	81.62	80.79	81.20
MDIR-BERT [Proposed]	**98.82**	**97.81**	**96.48**	**97.34**

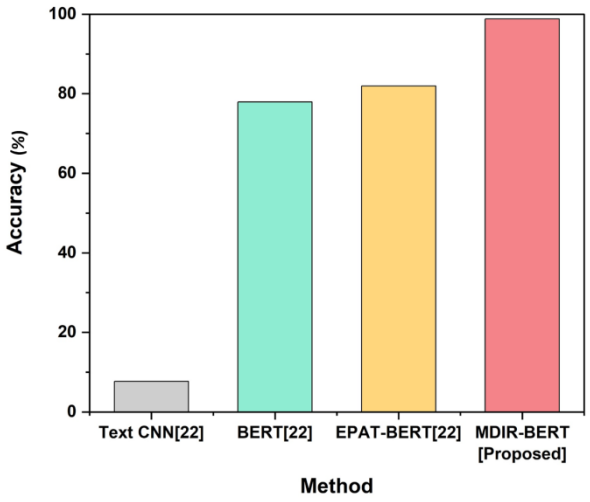

Figure 8: Graphical representation of accuracy for the MDIR-BERT

Figure 8 demonstrates consistent improvement in accuracy, which improves to 71.65% for Text CNN, 77.91% for BERT, and 81.96% with EPAT-BERT. The proposed method receives a significant increase to 98.82%, suggesting that power audit texts are classified very well overall.

Recall

Recall represents how well the model collects all the relevant audit information in the documents. Recall fits with a focus of reducing missed important content, which is key to holistic regulatory compliance and decision support in power audit. Recall is the ratio of True Positives (TP) to TP with the False Negatives (FN). Recall is important if the cost of misclassifying a positive instance is high, as in the case of a diagnostic method.

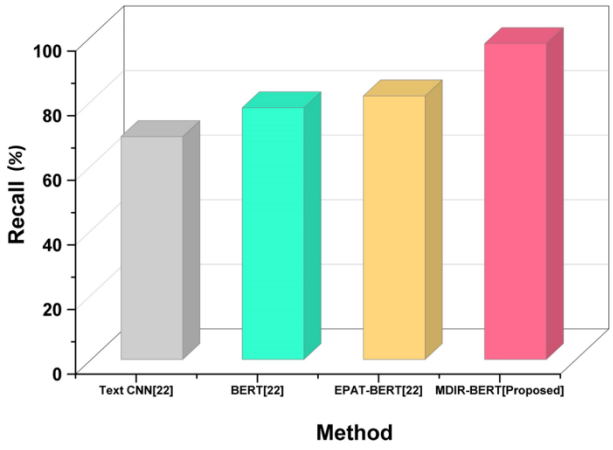

Figure 9: Visual depiction of classification recall achieved by MDIR-BERT

Figure 9 indicates the extent to which each model identifies all relevant content. Text CNN (69.01%) and BERT (77.94%) demonstrate moderate ability to identify relevant content, while EPAT-BERT shows a refined ability (81.62%), and the proposed method achieved 97.81%.

Precision

Precision assesses the extent to which each piece of text identified as relevant contains useful audit content. That is, it signifies the degree to which the model is able to avoid false positives and is a matter of importance for limiting irrelevant or misleading content through audit analysis. Precision measures the number of true positives (TP divided by the total number of TP, with the False Positives (FP). Precision is important if the cost of an FP is high, for example, misclassifying a legitimate user as a spammer or a fraudster.

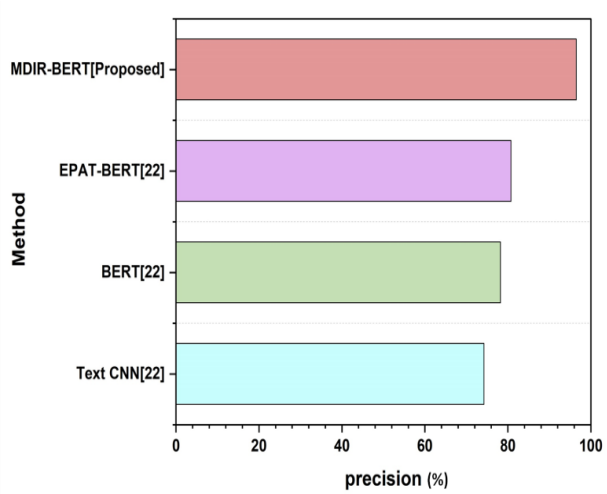

Figure 10: Precision analysis chart of MDIR-BERT model

Figure 10, indicating the correctness of predicted relevant pieces of information, is highest for the proposed method at 96.48%. This demonstrates a low false-positive rate. In the study, a measure of the precision could be performed with Text CNN, showing a decent 74.27%, BERT achieving a better performance of 78.23%, and EPAT-BERT showing 80.79%.

F1-Score

The F1-score balances precision and recall to deliver a single metric of model performance at comprehending an audit text. It can be especially helpful when it is as important to avoid false alarms as it is to capture every detail necessary. The F1-score is the harmonic average of recall and precision, which balances $precision$ and $recall$. It is especially useful in problems involving imbalanced data, especially when FP and FN are equally important.

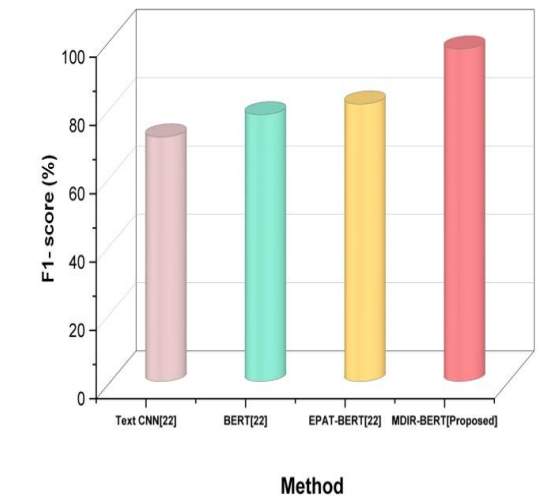

Figure 11: Performance visualization of MDIR-BERT in terms of F1-score

Figure 11 summarizes overall performance, balancing recall and precision. It ranges from 71.56% (Text CNN) to 78.08% (BERT) and 81.20% (EPAT-BERT). In contrast, the proposed method is 97.34%, and affirms our approach's clear superiority in accurately and consistently extracting meaningful audit content.

4.5 Training and testing splits

The training and testing validation of the proposed MDIR-BERT method's performance in 80:20 validations is compared with 70:30 splits to determine the efficiency of the proposed model in the field of power audit text understanding research. Table 4 explores the training and testing validation of the proposed model with 80:20 and 70:30 splits.

Table 4: Performance of proposed MDIR-BERT model with training and testing splits

Metrics	Training and Testing splits	
	80:20	70:30
Accuracy (%)	98.82	97.6
Precision (%)	96.48	95.5
Recall (%)	97.81	96.72
F1-score (%)	97.34	96.13

Based on the performance of various training and testing validations, the proposed MDIR-BERT model shows more significance in 80:20 validations than in 70:30 validation assessments.

The comparative results showed notable weaknesses in Text CNN [22], BERT [22], and EPAT-BERT [22] in their suitability to power audit text classification. Text CNN struggles with long-distance/contextual knowledge and domain-specific vocabulary and language due to its inability to layer information in a multi-dimensional way. Thus, Text CNN [22] will return lower recall and precision for this corpus. BERT [22] improved contextual understanding but did not adapt to the language structures and entities specific to power audits, which ultimately limited its performance on complicated auditor narratives. While EPAT-BERT [22] is adapted for domain use, it does not sufficiently model multi-dimensional relationships and detailed audit semantics. MDIR-BERT is superior to EPAT-BERT by the pretraining within a domain and multi-dimensional information retrieval (IR) boost, allowing for more in-depth electric power audit language comprehension and enhanced entity/context identification. It's a +16.86% accuracy improvement, indicating enhanced classification and retrieval. Whereas the success of the model is domain-specific and will not generalize to other audit types unless the model is retrained. In contrast to domain-specific transformers (such as FinBERT) and RAG-based models, MDIR-BERT has better structured text understanding but without the generative ability. Future research will focus on RAG integration for summarization and improving cross-domain adaptability by transfer learning or model compression. The proposed MDIR-BERT method in the research makes it possible for researchers, utilities, and regulatory organizations to modify and evaluate models for certain auditing conditions while providing innovation. High-stakes audit environments are secured by compliance with energy regulations and data management systems.

The proposed MDIR-BERT method in the research makes it possible for researchers, utilities, and regulatory organizations to modify and evaluate models for certain auditing conditions while providing innovation. High-stakes audit environments are secured by compliance with energy regulations and data management systems.

5 Conclusions

The aim was to build a multi-dimensional information retrieval for enhanced classification and comprehension of electric power audit texts. MDIR power audit text comprehension technology using the integration of multi-dimensional enhancement and a domain-adapted LLM. An end-to-end data preprocessing method was utilized, which involved data cleaning to eliminate unwanted symbols and noise, normalization via lemmatization to normalize word forms, and tokenization to split text into useful units appropriate for model input. MDIR-BERT model, being pre-trained on electric power audit texts, efficiently learned domain-specific terms, morphological phenomena, and entity relationships. These preprocessing operations considerably enhanced the textual data quality and uniformity utilized for training and fine-tuning. The model achieved significant accuracy (98.82%), recall, precision, and F1-score improvements, signifying robust performance. It also exhibited very high efficiency through lower energy expenditure, a quicker execution time, and improved convergence rate.

5.1 Limitations and future scopes: In uncertain circumstances, MDIR-BERT gets biased or hallucinatory findings with its performance, which leads to incorrect regulatory decisions. The integrity of an audit could be affected by misuse or a misconception that lacks domain expertise. These limitations show the significance of management and the consistency of power sector control regulations. The quality of domain-specific information is a potential factor that could be further investigated. Additional research intends to develop reasoning capabilities and extend the model's ability to process a broader range of document types. Future research should focus on the generalization of the model to various sectors.

Funding:
This work was supported by Technology Project of State Grid Tianjin Electric Power Company (Grant no. Chengdong Yanfa 2024-05).

References

[1] Pan M, Liu Y, Chen J, Huang EA, & Huang JX (2024). A multi-dimensional semantic pseudo-relevance feedback framework for information retrieval. *Scientific Reports*, *14*(1), 31806. https://doi.org/10.1038/s41598-024-82871-0

[2] Guo J, Fan Y, Pang L, Yang L, Ai Q, Zamani H & Cheng X (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management*, *57*(6), 102067. https://doi.org/10.1016/j.ipm.2019.102067

[3] Wang X, Wang J, Cao W, Wang K, Paturi R, & Bergen L (2024). Birco: A benchmark of information retrieval tasks with complex objectives. *arXiv preprint arXiv:2402.14151*. https://doi.org/10.48550/arXiv.2402.14151

[4] Hambarde KA, & Proenca H (2023). Information retrieval: recent advances and beyond. *IEEE Access*, *11*, 76581-76604.https://doi.org/10.3390/1010000

[5] Taherzadeh-Shalmaei N, Rafiee M, Kaab A, Khanali M, Rad MAV and Kasaeian A (2023). Energy audit and management of environmental GHG emissions based on multi-objective genetic algorithm and data envelopment analysis: An agriculture case. *Energy Reports,* 10, pp.1507-1520.

[6] Quispe EC, Viveros Mira M, Chamorro Díaz M, Castrillón Mendoza R & Vidal Medina JR (2025). Energy Management Systems in Higher Education Institutions' Buildings. *Energies*, 18(7), 1810. https://doi.org/10.3390/en18071810

[7] Rios FC, Al Sultan S, Chong O & Parrish K (2023). Empowering Owner-Operators of Small and Medium Commercial Buildings to Identify Energy Retrofit Opportunities. *Energies*, 16(17), 6191. https://doi.org/10.3390/en16176191

[8] Gunasegaran MK, Hasanuzzaman M, Tan C, Bakar AHA & Ponniah V (2023). Energy Consumption, Energy Analysis, and Solar Energy Integration for Commercial Building Restaurants. *Energies*, 16(20), 7145. https://doi.org/10.3390/en16207145

[9] Ai Q, Bai T, Cao Z, Chang Y, Chen J, Chen Z ... & Zhu X (2023). Information retrieval meets large language models: a strategic report from chinese ir community. *AI Open*, *4*, 80-90. https://doi.org/10.1016/j.aiopen.2023.08.001

[10] Sun B & Huo F (2025). Analysis of Customer Comment Data on E-commerce Platforms Based on RPA Robots. *Informatica*, *49*(10). https://doi.org/10.31449/inf.v49i10.5908

[11] Xiong X & Zheng, M. (2024). Merging mixture of experts and retrieval augmented generation for enhanced information retrieval and reasoning. https://doi.org/10.21203/rs.3.rs-3978298/v1

[12] Zhang H, Zhao Y, Sun B, Wu Y, Fu Z & Xiao X (2025). Large Language Model Based Intelligent Fault Information Retrieval System for New Energy Vehicles. *Applied Sciences*, *15*(7), 4034. https://doi.org/10.3390/app15074034

[13] Wen B, Wang T, Xu J, Liu Y, Li J & Lin S (2025). File Compliance Detection Using a Word2Vec-Based Semantic Similarity Framework. *Informatica*, 49(18). https://doi.org/10.31449/inf.v49i18.7421

[14] Tang Q, Chen J, Yu B, Lu Y, Fu C, Yu H ... & Li Y (2024). Self-retrieval: Building an information retrieval system with one large language model. *arXiv e-prints*, arXiv-2403. https://doi:10.48550/arXiv.2403.00801

[15] Azeroual O, Nacheva R, Nikiforova A & Störl U (2025). A CRISP-DM and Predictive Analytics Framework for Enhanced Decision-Making in Research Information Management Systems. *Informatica*, 49(18). https://doi.org/10.31449/inf.v49i18.5613

[16] Huang AH, Wang H & Yang Y (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806-841. 10.1111/1911-3846.12832 DOI:10.1111/1911-3846.12832

[17] Gan Z (2024). LARGE LANGUAGE MODELS EMPOWERING COMPLIANCE CHECKS AND REPORT GENERATION IN AUDITING. *World Journal of Information Technology*, 35.10.61784/wjit3003

[18] Panda A, Tang X, Nasr M, Choquette-Choo CA & Mittal P (2025). Privacy auditing of large language models. *arXiv preprint arXiv:2503.06808*. https://doi.org/10.48550/arXiv.2503.06808

[19] Oukhouya MH, Angour N, Aboutabit N & Hafidi I (2025). Comparative Analysis of ARDL, LSTM, and XGBoost Models For Forecasting The Moroccan Stock Market During The COVID-19 Pandemic. *Informatica*, 49(14). https://doi.org/10.31449/inf.v49i14.5751

[20] Li X, Wu X, Luo Z, Du Z, Wang Z & Gao C, (2023). Integration of global and local information for text classification. *Neural Computing and Applications*, 35(3), pp.2471-2486. https://doi.org/10.1007/s00521-022-07727-y

[21] Nordin IG (2023). Narratives of internal audit: The Sisyphean work of becoming "independent". *Critical Perspectives on Accounting*, 94, p.102448. DOI:10.1108/MEDAR-01-2022-1584

[22] Meng Q, Song Y, Mu J, Lv Y, Yang J, Xu L ... & Meng Q (2023). Electric power audit text classification with multi-grained pre-trained language model. *IEEE Access*, *11*, 13510-13518. https://doi.org/10.1109/ACCESS.2023.3240162