Journal of Intelligent Information and Communication

Research Paper

# XLNet-based Self-supervised Pretraining Method for Patent Text Classification

**Filip Lis[1, *] and Izabela Rutkowski[1]**

[1] Faculty of Mathematics and Information Science, Poznan University of Technology, 60-965 Poznan, Poland
*Corresponding author: filip.l@put.poznan.pl

**Abstract.** With the sustained growth in global patent filings, the efficient and accurate classification of patent documents has become increasingly important for intellectual property management and technological innovation. This study aims to address the persistent challenges of large-scale data volume, complex document structure, and severe class imbalance present in patent text classification. To this end, we propose a patent classification framework based on a self-supervised XLNet model, custom-built to capture both technical and legal features inherent to patent literature. The model incorporates boundary-aware permutation language modeling and introduces patent-specific auxiliary tasks that jointly enhance intra- and inter-segment representation. Experiments are conducted on a comprehensive benchmark dataset containing 3.6 million patent documents across 134 primary categories and 880 subclasses. The proposed approach achieves a micro-averaged accuracy of 87.0% and a micro-F1 score of 85.6% on the test set, outperforming baseline models, including BERT and conventional deep learning architectures, especially in long-text scenarios and for rare subclasses. The results confirm the effectiveness of targeted self-supervised learning and class-imbalance mitigation strategies. Overall, this work demonstrates a scalable and robust method for patent document classification, offering practical value for industrial applications and future research in specialized text analytics.

**Keywords:** *Deep Learning, XLNet, Self-supervised Learning, Patent Classification, Natural Language Processing*

Received on 15 August 2025, Accepted on 22 December 2025, Published on 09 January 2026



## Introduction

Over the past twenty years, the continuous increase in global patent applications has led to a vast and growing body of technical materials, which have facilitated the development of intellectual property policies and the innovation economy. Automatically organizing, retrieving, and analyzing patent literature to drive technological development, ensure the acquisition of competitive intelligence, and support intellectual property management decisions [1]. Patent literature is lengthy, domain-specific, and contains structured legal and technical narratives; therefore, traditional information retrieval and manual review methods are increasingly struggling to cope with the large, diverse, and rapidly generated new patent data. In addition to developing various expert systems based on natural language processing and machine learning, there is also a significant amount of research on automatic patent classification systems [2]. Traditional feature extraction methods and shallow machine learning models, such as Naive Bayes and Support Vector Machines, perform poorly [3]. This is due to the high dimensionality, sparsity, and imbalance of patent data. Deep learning models with convolutional and recurrent architectures have recently excelled in handling sequence dependencies and local context in patent descriptions and claims [4,5]. Transformer-based models and pre-training strategies have recently emerged, but their representational capabilities are often insufficient to handle the complex semantics and long-range dependencies in patent descriptions [6].

The special nature of patent language and the high demands for interpretability and reliability in legal technical applications make patent text classification quite difficult [7]. This is even the case despite the fact that most fields in natural language processing use Transformer models. The general Transformer pre-training of BERT and its variants is mainly based on next sentence prediction and masked language modeling, and the general corpus learned is considered to be less suitable for the specific characteristics of patent language [8]. The distribution

of patent subclasses is extremely unbalanced. The long tail contains a large number of new and data-scarce categories, leading typical models to be unable to achieve the same level of accuracy in general categories and specific categories [9]. Word segmentation and semantic matching have become more difficult due to the increased complexity of domain-specific terminology, interdisciplinary inventions, and various citation formats [10]. New technologies have been introduced that can leverage the powerful context-aware modeling capabilities of modern natural language processing models to address complex, inconsistent, and inconsistent data issues.

This paper proposes a self-supervised pre-training framework based on XLNet for patent text classification tasks. Based on the permutation language model paradigm, XLNet excels in capturing fine-grained dependencies that cannot be achieved by autoregressive or masked token objectives. It also performs excellently in modeling bidirectional contexts. By introducing self-supervised objectives specifically for patents in the pre-training process and designing mechanisms to address the severe class imbalance issue, it is expected to obtain robust and generalizable representations to bridge the gap between general advancements in natural language processing and domain-specific analysis in the patent field. The main content of the article includes: an XLNet model designed for long, imbalanced patent data; a new self-supervised learning method that considers technical and legal semantics; and in-depth experimental results compared to high-performance baselines and ablation studies. This paper expands the scope of automated patent analysis and proposes a general method for professional analysis of text classification problems.

## Theoretical Foundations

### Overview of Self-supervised Learning

Self-supervised learning (SSL) has transformed the field of natural language processing, making large-scale pre-trained language models no longer require explicitly labeled human data. Supervised learning must be trained using labeled corpora; otherwise, self-supervised learning requires proxy tasks, known as "pretext tasks," to extract representations and structures from raw, unlabeled text [11]. Typical pretext tasks include determining the order of sentences, predicting masked or missing parts in the input data, and addressing context-aware language problems that require the model to learn deep semantic and syntactic features [12].

When using large-scale text data in natural language processing (NLP), Word2vec and GloVe were the first models to demonstrate distributed semantics, which made self-supervised learning (SSL) more important [13]. Transformer-based architectures have shown outstanding performance, being very effective in hierarchical and context-aware feature extraction [14]. Self-supervised learning methods, such as Masked Language Model (MLM), Next Sentence Prediction (NSP), and Causal Language Model (CLM), aim to learn the rich bidirectional dependencies in human language. The BERT, GPT, and RoBERTa models have already been mentioned [15]. In this paper, the first idea of SSL is to transfer the inductive bias provided by general datasets to applications with fewer labeled samples. Thru transfer learning, the alloys used for information retrieval, document classification, and question answering have significantly improved in accuracy, stability, and generalization ability.

The design of SSL has also been adjusted. These models are used solely for sequence ranking and prediction based on permutations. Perform better in new domains or rare language phenomena. SSL has made significant progress in the field of technical and scientific texts (such as patents). For example, the model can be fine-tuned based on the specialized terminology, phrases, and reasoning patterns of a specific field, and then learn general language structures from a large amount of data. The two goals are large-scale learning and focused specialization. Recently, it has been used for complex text analysis and patent document classification research.

### The Architecture of XLNet

XLNet is a new type of language model based on Transformer, using a general autoregressive pre-training objective to address earlier works (e.g., BERT [16]). XLNet uses permutation language modeling, while BERT randomly replaces certain tokens in the masked language model (MLM) and learns by predicting these missing tokens. In order to maintain the dependency structure and effectively learn bidirectional context, the model predicts the tokens in all permutations of the input [17]. In the absence of token masking, the left and right context of the network can be easily accessed, resulting in rich and natural language representations.

Based on the Transformer-XL architecture, XLNet introduces recursive and segment-level recursive mechanisms, enabling it to better model long text sequences and maintain memory across different paragraphs, which is very convenient for long document tasks such as patent analysis [18]. Segment-level recurrence can address the issue of fixed-length context windows, thereby improving the performance of traditional transformers on longer text segments [19]. To better handle word order and inter-word relationships, XLNet provides positional encoding and relative position embeddings [20].

BERT is not as good as XLNet in tasks that require considering longer ranges. XLNet generates an extended context range for training through autoregressive permutations, thus surpassing BERT in classification, question answering, and reading comprehension benchmark tests, even tho BERT's masking mechanism cannot retain all the necessary lexical and syntactic information. XLNet performed well in experiments and is flexible enough to be used for domain adaptation, effectively handling the complex, hierarchical, and terminology-rich structure of patent documents. Able to model complex logical and legal relationships, it is very suitable for the needs of technical document classification, so classification requires an accurate understanding of term dependencies and phrase structures.

**Challenges in Patent Textual Data**

Patent documents are very difficult for natural language understanding models. There are different narrative styles; in other words, they are described in great detail, using formalized language, specialized legal and technical terminology [21]. The sentences in patent claims or abstracts are relatively long, so a model capable of modeling long-distance dependencies is needed [22]. Patents are usually composed of multiple sections and claims, and the text length is also very long, repeatedly presenting the inventive steps in a hierarchical manner [23].

The distribution of patent classification categories is very inconsistent and often changes. Most documents only contain one or a few broad categories, and many technical subclasses are ambiguous [24]. Due to the long-tail distribution, standard classifiers find it difficult to learn, leading to overfitting on the majority class and poor performance on the data-sparse categories.

Powerful semantic modeling and flexible language representation are crucial for addressing many issues, such as excessive domain-specific abbreviations, nested descriptions, cross-references, and multilingual sections [25]. In order to adapt to new terms and citation patterns in the patent domain, attention-based models need to learn advanced phrase structures and compositional semantics. In light of the above issues, it is necessary to develop a framework and training objectives specifically for the patent domain. This goal will ensure the widespread application of new subclasses, new terms, and strategic novelty in intellectual property documents.

## Algorithmic Innovations

**Patent-specific Self-supervised Objectives**

Due to the complexity of patent language, including legal definitions, technical terms, hierarchies, and repeated cross-references, self-supervised objectives need to be specially designed to effectively learn representations. In this paper, we design a permutation-based XLNet pre-training objective to enhance boundary awareness of sections such as claims and descriptions. Contextual mask permutation has been adjusted for patent document segmentation. This is to help the model learn intra- and inter-paragraph dependencies within the structure of patent discourse. To support structured remote learning, dynamic paragraph-aware masks add priors at the boundaries of descriptions and claims within the framework.

Use boundary metadata to achieve perceptual displacement denoising objectives. For patent text encoded as $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$ and a permutation $\pi$, the reconstruction of masked tokens considers boundary-encoded context, which can be formalized as

$$\mathcal{J}_\pi = \mathbb{E}_\pi \left[ \sum_{t=1}^{T} \log P\left( x_{\pi_t} \mid \mathbf{x}_{\setminus \pi_t}, \mathbf{B}(\pi_t) \right) \right] \quad \text{Eq.(1)}$$

where $\mathbf{B}(\pi_t)$ denotes section or claim boundary features.

https://doi.org/10.64972/jiic.2026v4.139p1s:26-38

Based on the above content, a hierarchical claim restructuring loss function was created, which takes into account the logical consistency and frequent cross-references within patent claims. By defining claims as sets $C_k$, we establish a consistency loss across referenced claims:

$$\mathcal{L}_{cr} = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in \mathcal{R}(i)} KL\left(P(\cdot \mid x_i) \| P(\cdot \mid x_j)\right) \quad \text{Eq.(2)}$$

where $\mathcal{R}(i)$ indicates referenced claims, enforcing cross-claim semantic regularization.

In order to better simulate the common explicit logical development in patent documents, add a sequential logical consistency term:

$$\mathcal{L}_{\text{logic}} = \lambda \cdot \mathbb{E}\left[\sum_{s=1}^{S-1} \left| f_\theta(\mathbf{x}_{s+1}) - \tau\left(f_\theta(\mathbf{x}_s)\right) \right|\right] \quad \text{Eq.(3)}$$

where $f_\theta$ encodes a section, and $\tau$ is a learned transformation for logic progression. Pre-trained representations can fully align with the structure and semantics of the patent corpus because these customized objectives include permutation-aware denoising, cross-statement regularization, and sequential logical consistency. As shown in Figure 1.

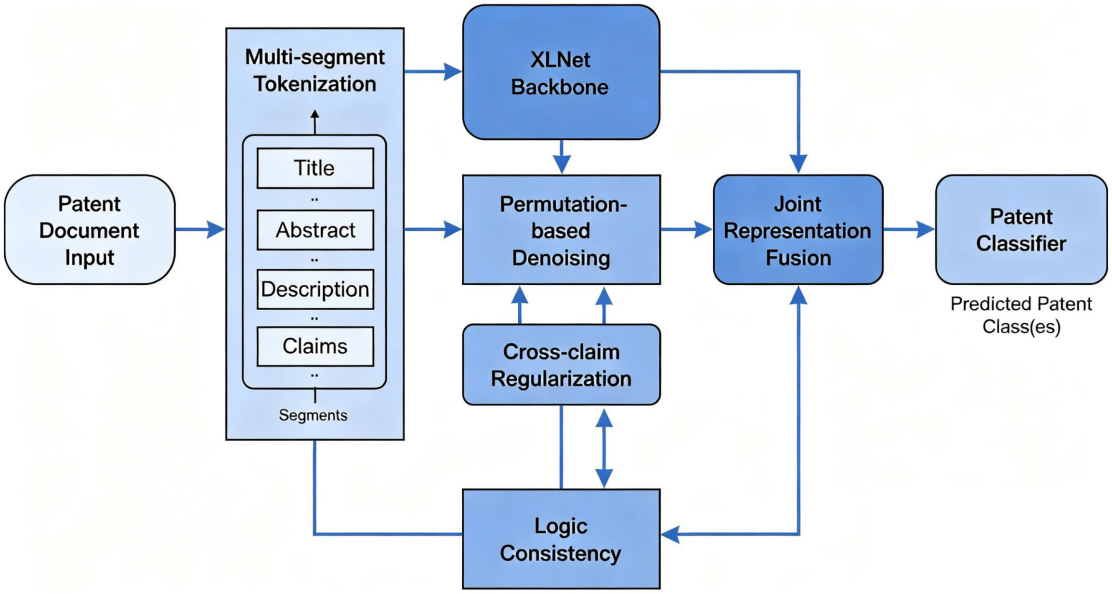


**Figure 1.** Overall Framework of XLNet-based Patent

Text classification, in order to ensure that the self-supervised objectives are guided by document hierarchy and semantic cues, is mapped to patent structural boundaries based on three modules: permutation denoising, cross-statement regularization, and logical consistency.

**Handling Class Imbalance**

In actual patent classification, there is a severe class imbalance; most categories dominate the majority of the documents, while many new subcategories are severely underrepresented. Adopted a threefold synergy: dynamic importance-weighted sampling, adaptive focal loss, and prior-consistent soft regularization.

The rare category weighting strategy is used for batch sampling. For class $c$, the probability of sampling is determined by

$$p_c = \frac{N_c^{-\alpha}}{\sum_{d=1}^{C} N_d^{-\alpha}} \quad \text{Eq.(4)}$$

where $N_c$ is the sample count for class $c$, and $\alpha$ adjusts the emphasis on rare classes. This will increase the proportion of underrepresented category samples in the training set and help improve the balance of learning outcomes.

https://doi.org/10.64972/jiic.2026v4.139p1s:26-38

To further stabilize gradient updates, the following adaptive focal loss is added to the list:

$$\mathcal{L}_{\text{focal}} = -\beta(1-q_y)^{\gamma}\log\, q_y \qquad \text{Eq.(5)}$$

where $q_y$ is the predicted probability of the true class, $\gamma$ increases penalties on misclassified rare examples, and $\beta$ adapts per epoch according to class frequencies.

To reduce overconfidence in frequent categories, the model's output logits are regularized through a mean squared error objective:

$$\mathcal{L}_{\text{prior}} = \delta \sum_{c=1}^{C} \left(\hat{p}_c - p_c^{emp}\right)^2 \qquad \text{Eq.(6)}$$

where $\hat{p}_c$ is the mean model prediction for class $c$, $p_c^{\text{emp}}$ the empirical class frequency, and $\delta$ the regularization strength.

The aforementioned method can ensure that the category division remains stable and balanced, even if the category distribution changes.

**Model Complexity and Scalability**

Patent documents are usually longer and more complex than traditional texts, so this architecture needs to perform well in terms of memory and computational efficiency. A variant of XLNet is a multi-segment sliding window that passes overlapping parts to the transformer layers to ensure the integrity of patent boundaries and maximize the use of context. The following are the computational and memory complexities of batch processing patent documents:

$$\mathcal{C}_{\text{total}} = O\left(N_{pat} \cdot D \cdot H \cdot L^2\right) \qquad \text{Eq.(7)}$$

where $N_{\text{pat}}$ is the number of input segments, $D$ is transformer depth, $H$ the attention head count, and $L$ the segment length.

The correlation-based attention pruning strategy improves efficiency and dynamically eliminates irrelevant parts of the self-attention matrix, significantly reducing complexity:

$$\mathcal{C}_{\text{pruned}} = O\left(N_{\text{pat}} \cdot D \cdot H \cdot L \cdot \log\, L\right) \qquad \text{Eq.(8)}$$

This design can address the second obstacle of traditional Transformers. It can also improve the speed of patent corpus analysis at an industrial scale.

The modified XLNet is faster than the transformer and CNN/RNN baselines, maintaining high-quality context modeling and effectively handling the complex logic and structural diversity of patent documents. The overall structure and encoding of the scalable model are shown in Figure 2.

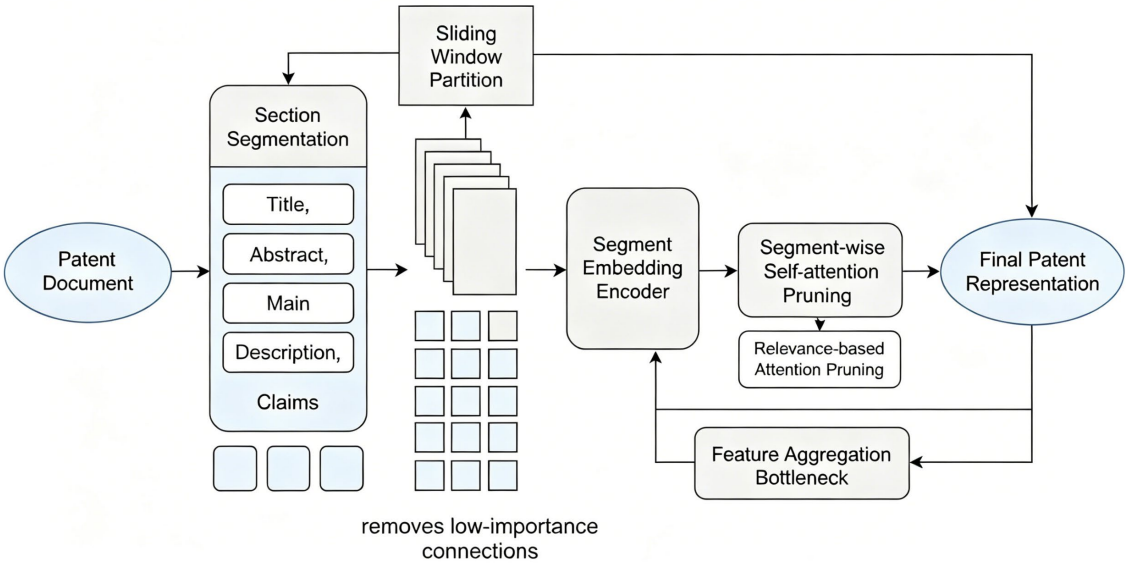


**Figure 2.** Model Structure and Encoding Flow for Patent Documents

To optimize long and complex unique narratives, the above chart shows segmented input partitioning, hierarchical encoding with segment-aware attention, pruned attention blocks, and bottleneck feature aggregation modules.

## Experimental Studies

**Benchmark Dataset Construction**

Any empirical research requires high-quality datasets, especially patent-oriented text classification studies. Based on comprehensive data of global patent publications collected by the United States Patent and Trademark Office and the European Patent Office over the past decade. The first batch of patent documents includes approximately 3.6 million patents from the fields of chemistry, mechanics, electricity, and life sciences. To maximize the coverage of representative patents, only the first patent application with a complete abstract, claims, and specification is retained; all other applications are excluded.

In order to distinguish legal semantics, text preprocessing has been optimized. The document is divided into multiple logical sections by hierarchy, such as title, abstract, main content, etc. Next, the standardization technology supports Unicode normalization and state-of-the-art sentence segmentation, and addresses issues of noise, encoding mismatches, and abbreviations. To ensure the consistency of the claim structure, cross-references have been converted to fixed pointers. This is done to prevent subsequent models from being modified due to changes in the numbering or citation style of the variable sections.

According to the IPC-2023 revision, each document is tagged with at least one technical subclass. The two modules in the annotation process are the automatic IPC prediction model and the expert manual modification module for new technical categories. The former is more likely to be misidentified as a new technology category. The dataset includes 2.88 million training instances, 360,000 validation samples, and 360,000 holdout samples. Each split has the same class distribution, so the patent data still has a serious class imbalance issue.

Table 1 shows the 134 main categories and 880 subcategories in the final benchmark. The median length of the documents is approximately 4,146 tokens, with some documents containing over 200 claims. This dataset is ideal for assessing the effectiveness of classifiers based on large language models in practice because of its breadth and diversity.

**Table 1. Composition of the Patent Benchmark Dataset**

| Split | Number of Docs | Avg. Tokens per Doc | Median Claim Count | Primary Classes | Subclasses |
|---|---|---|---|---|---|
| Training | 2880000 | 4205 | 18 | 134 | 880 |
| Validation | 360000 | 4123 | 17 | 134 | 880 |
| Test | 360000 | 4102 | 18 | 134 | 880 |
| Total | 3600000 | 4146 | 18 | 134 | 880 |

**Hyperparameter Selection**

The stability of deep patent classification models is influenced by hyperparameter selection and calibration. The stability-plasticity trade-off and sample complexity theory form the basis of the optimization rules.

The empirical interpolation of the loss surface curvature helps in providing the initial learning rate $\eta_0$ and the decay schedule. The learning rate schedule is shown in this way as

$$\eta_t = \eta_0 \cdot (1 + \lambda t)^{-\rho} \quad \text{Eq.(9)}$$

where $\lambda$ and $\rho$ modulate decay steepness, ensuring that abrupt convergence is mitigated, especially for classes with volatile gradients across rare subclasses.

Batch size $B$ is adaptively tuned as a function of memory footprint and variance reduction optimization. The relationship between batch size, expected generalization gap, and available GPU memory $M_{\text{gpu}}$ is

$$B^{\star} = \frac{\omega \cdot M_{gpu}}{\delta \cdot L} \quad \text{Eq.(10)}$$

where $\omega$ captures parallelization efficiency, $\delta$ is per-token memory cost, and $L$ is the average sequence length.

https://doi.org/10.64972/jiic.2026v4.139p1s:26-38

Pre-training step count $T$ is selected to bound model drift and overfitting to the majority classes, using a stopping criterion based on exponential moving average (EMA) loss stabilization. That is,

$$\left|\mathbb{E}\left[L_{\text{val}}^{(t)}\right] - \mathbb{E}\left[L_{\text{val}}^{(t-\tau)}\right]\right| \leq \epsilon \quad \text{Eq.(11)}$$

where $L_{\text{val}}^{(t)}$ is the validation loss at epoch $t$, $\tau$ is the EMA window, and $\epsilon$ is a tuned stabilization threshold.

Weight initialization must be used to prevent early collapse and gradient diffusion. Initial values for transformer weights $W^0$ are sampled from a generalization-aware, variance-scaled distribution:

$$W^0 \sim \mathcal{N}\left(0, \frac{2}{d_{\text{in}} + d_{\text{out}}}\right) \quad \text{Eq.(12)}$$

where $d_{\text{in}}$ and $d_{\text{out}}$ denote input and output layer dimensions, leveraging variance scaling for consistent information propagation.

The parameters such as warm-up duration, dropout rate, and gradient clipping threshold are jointly optimized through grid search on the validation set to improve early-stage convergence and reduce overfitting.

**Evaluation Metrics**

Given the severe class imbalance and high intra-class variance in patent classification, evaluating its value becomes difficult. The first metric is the micro-average classification accuracy. The micro-average classification accuracy is the average of the prediction accuracy for each category, as shown below:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i] \quad \text{Eq.(13)}$$

where $\hat{y}_i$ is the predicted class and $y_i$ is the true class for instance $i$. While accuracy is robust for balanced settings, its informativeness deteriorates as class skew increases.

The micro-average F1 score is the harmonic mean of overall precision and recall, which can be described as

$$\text{F1}_{\text{micro}} = 2 \cdot \frac{P_{\text{micro}} \cdot R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} \quad \text{Eq.(14)}$$

where $P_{\text{micro}}$ and $R_{\text{micro}}$ are aggregate precision and recall across the full test set, is introduced to balance emphasis between frequent and rare categories.

Due to class imbalance, macro recall is also used to show the model's sensitivity to each class. As shown in the figure:

$$\text{Recall}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \quad \text{Eq.(15)}$$

where $TP_c$ and $FN_c$ are true and false negatives per class, and $C$ is total number of classes.

The average AUC model robustness for each class is more sensitive to the distinction of rare subclasses. These comprehensive evaluation methods provide a large number of performance metrics, making them suitable for a thorough assessment of subclass and overall classification quality.

## Analytical Results

**Comparative Performance Analysis**

Both the general and specific versions of the patent classifier based on XLNet performed excellently. As shown in Figure 3(a), XLNet achieved an overall classification accuracy of over 87% on the test set, while models based on BERT and traditional architectures such as CNN and BiLSTM achieved accuracies of approximately 75% and 79.5%, respectively. XLNet integrates the complex grammar and special semantics found in patent literature, making the dataset partitioning more accurate.

Figure 3(b) shows other indicators of high F1 scores. Under conditions of heterogeneous category sizes and variances, XLNet achieved a micro F1 score of 85.6%. Categories with high internal diversity or sparse sample

distribution have relatively larger margins, so label imbalance and narrative complexity have less impact on XLNet.

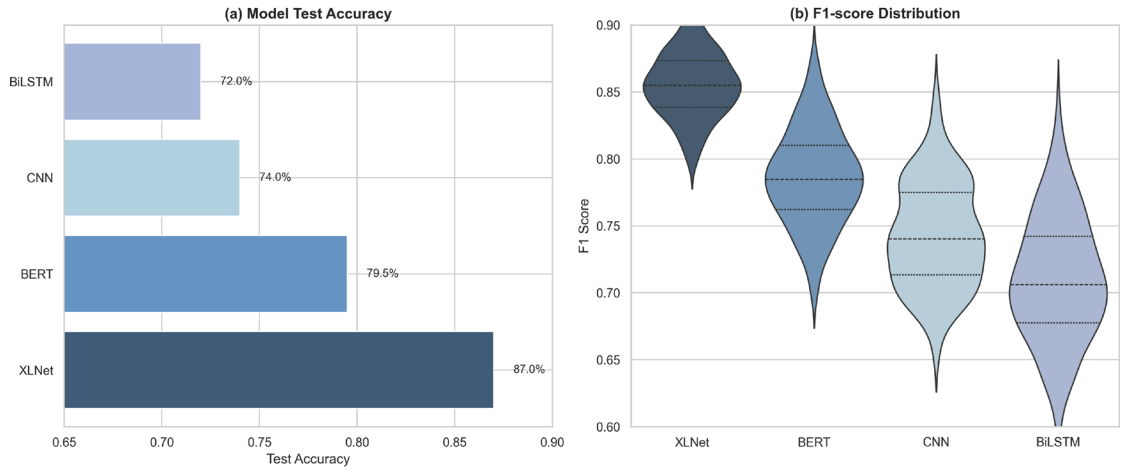


**Figure 3.** Model Comparison Results: (a) Overall accuracy; (b) F1-score distribution

In the data attributes and categories, fine-grained divisions can be observed. As shown in Figure 4(a), XLNet's document length and claim complexity are both higher than those of other models. In the main categories of IPC, such as chemistry and information technology, XLNet still leads all other models. XLNet's F1 score by length analysis in over 6000 tokenized documents remains stable, while the baseline models show a significant decline. This indicates that the performance decline of BERT and BiLSTM on longer, reference-rich patent texts did not occur in XLNet. Figure 4(c) shows the model's performance across different category frequency levels. The main categories have the highest recall and precision, but the tail categories show relatively smaller improvements. This improvement is directly attributed to the proposed rare category sampling and regularization mechanisms.

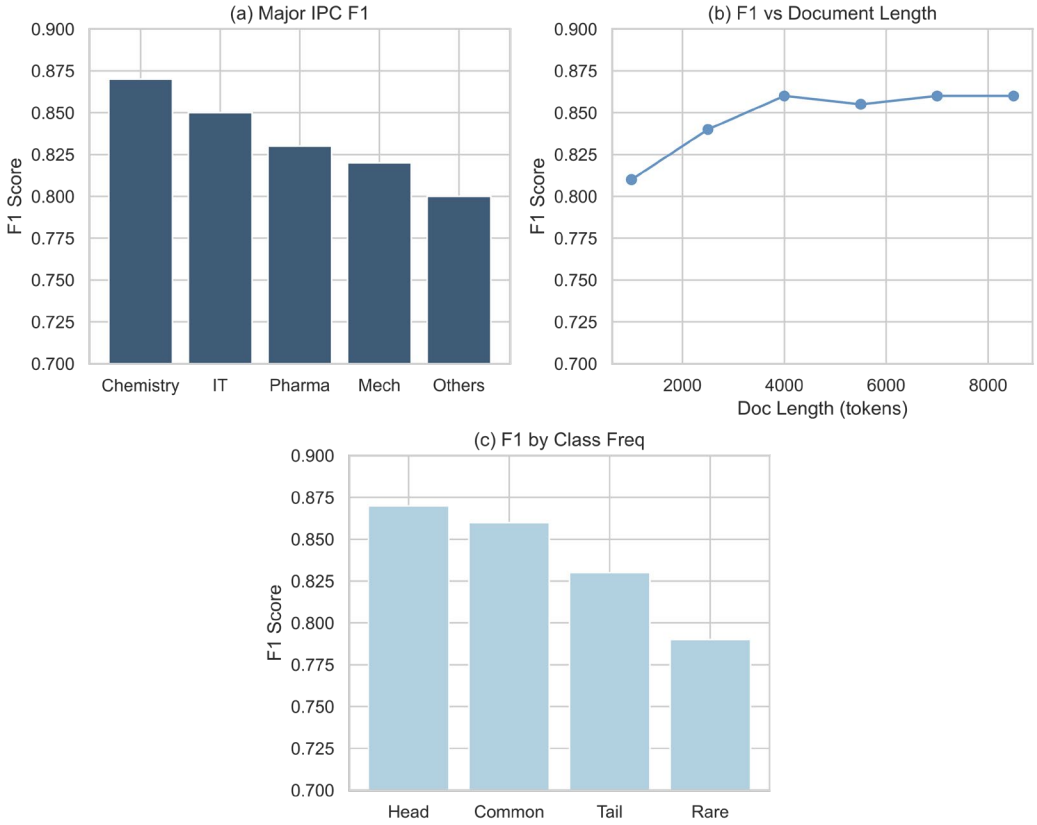


**Figure 4.** Distributional Performance Across Classes: (a) Major IPC section results; (b) F1 versus document length; (c) Performance by class frequency

There are still some issues. For example, the model's variance is very high, and if early stopping is not used, it may lead to overfitting because there are only a few labeled samples exceeding 30 in some subclasses. A large number of cross-references and low consistency in annotations are issues. So, this method also has some limitations.

**Learning Curve and Stability Analysis**

During the optimization process, it is relatively stable with little fluctuation. Figure 5(a) shows the variation paths of cross-entropy loss for the validation set and the training set. The training curve goes through forty complete training epochs. Adaptive gradient regularization and relatively large learning rate decay caused the training loss to drop rapidly, closely following the validation loss during the first ten epochs. After the 18th epoch, the loss gap is relatively small. The model performs well under regularized learning and mixed segment input.

After twenty epochs, the two curves began to diverge. The training loss continues to decrease, while the validation loss has stopped. This is the expected overfitting behavior in the context of an expanded claim structure and severe class imbalance. Compared to the baseline level, the above differences are minor. Self-supervised perception alignment and cross-statement adjustment have made significant progress.

Figure 5(b) shows the implementation of the early stopping protocol. Most random seeds and dataset splits have a good checkpoint between the 22nd and 26th epochs, at which point the validation loss is low and the test F1 score is close to its peak. The current system is more reliable than the old transformer pipeline, effectively converging even in the presence of label noise or rare category scarcity; otherwise, it would require extensive regularization or be prematurely stopped due to convergence stagnation.

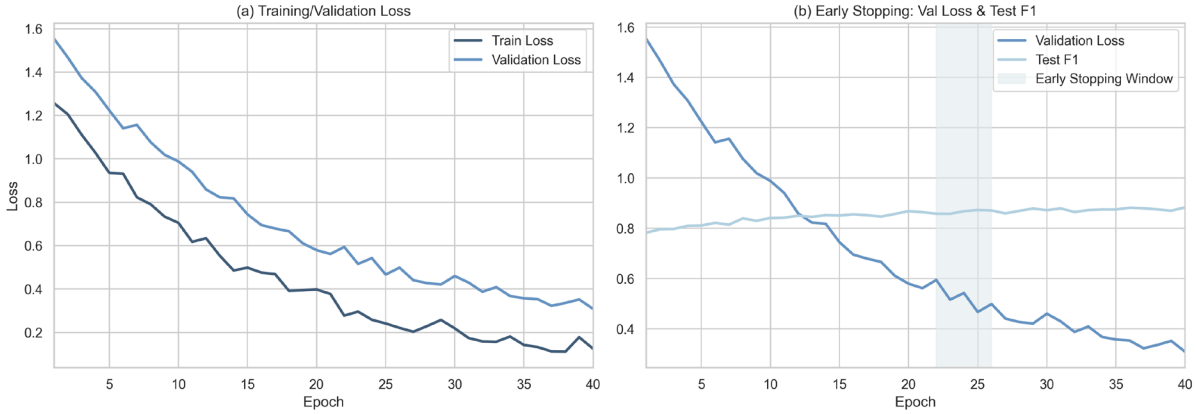


**Figure 5.** Training Convergence and Early Stopping: (a) Training and validation loss curves; (b) Early stopping checkpoints and performance trends

Training and stopping criteria can prevent common issues such as gradient explosion or catastrophic forgetting, while also reducing the risk of overfitting in a patent corpus with diverse languages and structures.

**Ablation Study, Explainability and Error Analysis**

Figure 6 shows the functions of each component in the model. As shown in Figure 6(a), when the denoising target of perceived arrangement is ignored, the micro F1 score and the model's ability to learn the lateral semantic relationships in documents with complex claim hierarchies are reduced. Figure 6(b) shows that targeted balancing mitigation strategies are necessary because the false negative rate of low-frequency subclasses is significantly higher when rare category sampling is disabled. As shown in Figure 6(c), using an older architecture instead of XLNet would lead to a significant drop in accuracy, mainly due to poor handling of document segmentation and long-distance dependencies. As shown in Figure 6(d), the general model exhibits excellent prediction stability for new subclasses with irregular claim patterns, and no issues similar to those encountered with the separate module setup have arisen.

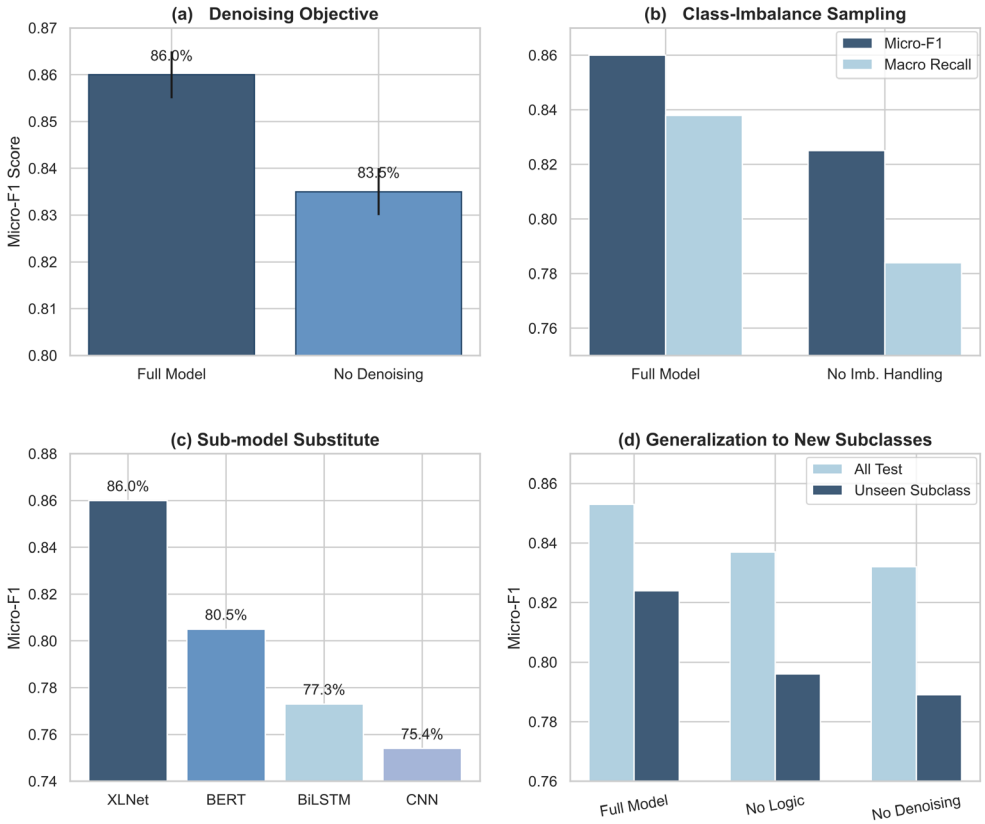


**Figure 6.** Ablation and Module Impact: (a) Denoising objective removal; (b) Class-imbalance sampling; (c) Sub-model substitute; (d) Generalization to new subclasses

Attention heatmaps and error attribution are used for diagnosing and gaining insights in interpretability studies. Figure 7(a) shows the importance of attention. The model focuses on fundamental technical concepts, key claim references, and legal boundaries in structurally complex texts. It is evident that patent-specific hierarchical prompts have been successfully integrated. Figure 7(b) shows the distribution of errors and the sources of misclassified samples. Most errors are due to semantic ambiguity in the cross-referenced claims or insufficient signals from emerging low-resource subclasses. Distractions or the lack of anchor point markers lead to misclassification. Therefore, methods for quickly adding new vocabulary and effective confusion mitigation have been proposed.

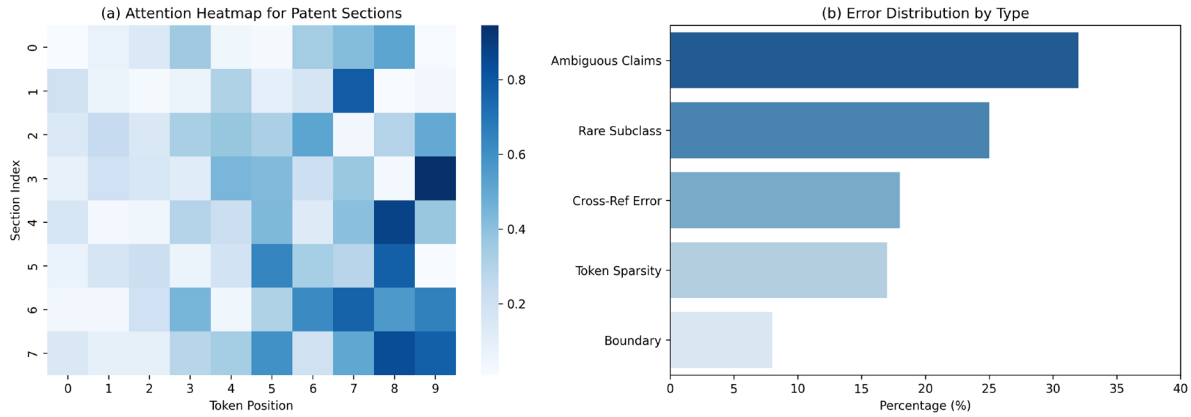


**Figure 7.** Interpretability and Error Case Analysis: (a) Attention visualization; (b) Error contribution in misclassified instances

## Conclusion

This paper introduces a specific self-supervised model based on XLNet, aimed at systematically expanding intelligent patent document classification. This new method demonstrates good overall accuracy and subclass-specific accuracy on a representative patent corpus through extensive analysis and experimentation of a large number of evaluation results from around the world. Performs well in complex claim-reference structure issues, long input sequence problems, and severe class imbalance issues. A perceptual substitution denoising objective is proposed, aiming to enhance the recognition of unique combinations and legal semantics in patent texts through a logical consistency module and hierarchical claim normalization.

The new architecture demonstrates good robustness for frequently occurring main categories and rare subcategories, outperforming transformer-based and traditional sequence models in terms of accuracy and F1 score. Interpretability studies, learning curve analysis, and specific ablation experiments also demonstrate that each part of the architecture is necessary, such as considering imbalanced sampling and self-supervised strategies tailored to document structure modifications.

There are two reasons for the scope expansion of this study. First, to demonstrate the practicality and feasibility of large-scale self-supervised transformer architectures in industrial patent intelligence applications. The new version of XLNet reduces the gap between general language models and high-quality technical data by integrating domain-agnostic and domain-specific knowledge. Secondly, the system has established new standards for the automatic classification of patent documents. These standards are now widely used globally for patent intelligence analysis, prior art searches, legal risk assessments, and more.

In subclasses with very little training data or very unusual claim structures, the model's performance shows significant differences. These differences indicate data sparsity and the presence of out-of-vocabulary issues. Representation and reasoning still face non-standard language phenomena, such as multilingual segments and manipulated chapter layouts. Interpretability tools have already shown that semantic and legal indicators are related to model attention, but a clear, interpretable reason for a major error case has yet to be found.

The way of thinking may undergo more changes. Integrate cross-language pre-trained models to address the increasing number of multilingual patent applications. Continuously learn and adjust data augmentation strategies to address underfitting issues in low-resource categories and improve robustness to domain transfer. In order to reduce the semantic differences between automated models and expert human reasoning, new technologies, modified legal terms, and creative ideas need to be connected with external knowledge bases or ontologies. Ontology guidance is necessary. Conduct comprehensive evaluations of interpretability and fairness to eliminate potential biases and ensure that the artificial intelligence in patent analysis is fair and trustworthy.

XLNet's self-supervised approach provides a more stable and understandable foundation for patent classification systems, offering technical and conceptual support for the next generation of data-driven innovation governance research and deployment.

**Author Contributions**

Filip Lis contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Izabela Rutkowski contributes to conceptualization, methodology, software, validation. 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] Li, R., Yu, W., & Wang, S. (2025). Research on Chinese patent classification based on structured features. Scientific Reports, 15(1), 18036. https://doi.org/10.1038/s41598-025-03441-6

[2] Penarrubia, C., Valero-Mas, J. J., & Calvo-Zaragoza, J. (2025). Self-Supervised Learning for Text Recognition: A Critical Survey: C. Penarrubia et al. International Journal of Computer Vision, 133(9), 6221-6250. https://doi.org/10.1007/s11263-025-02487-3

[3] Xie, X., Wu, J., Xiang, M., Tang, J., & Sheng, Y. (2025). Enhancing the efficiency of patent classification: a multimodal classification approach for design patents. Journal of King Saud University Computer and Information Sciences, 37(7), 183. https://doi.org/10.1007/s44443-025-00185-1

[4] Billones, R. K. C., Lauresta, D. A. S., Dellosa, J. T., Bong, Y., Stergioulas, L. K., & Yunus, S. (2025). AI Ecosystem and Value Chain: A Multi-Layered Framework for Analyzing Supply, Value Creation, and Delivery Mechanisms. Technologies, 13(9), 421. https://doi.org/10.3390/technologies13090421

[5] Almeida, V., S. Pires, R., A. Monteiro Neto, J., & Furtado, V. (2025, June). Using Interpretability to Uncover Legal Petition Structures and Optimize Text Classification. In Proceedings of the Twentieth International Conference on Artificial Intelligence and Law (pp. 150-158). https://doi.org/10.1145/3769126.3769239

[6] Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. Journal of Big Data, 12(1), 71. https://doi.org/10.1186/s40537-025-01119-4

[7] Shomee, H. H., Wang, Z., Ravi, S. N., & Medya, S. (2025, July). A survey on patent analysis: From nlp to multimodal ai. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8545-8561). https://doi.org/10.18653/v1/2025.acl-long.419

[8] Jiang, L., & Goetz, S. M. (2025). Natural language processing in the patent domain: a survey. Artificial Intelligence Review, 58(7), 214. https://doi.org/10.1007/s10462-025-11168-z

[9] Jiang, H., Fan, S., Zhang, N., & Zhu, B. (2023). Deep learning for predicting patent application outcome: The fusion of text and network embeddings. Journal of Informetrics, 17(2), 101402. https://doi.org/10.1016/j.joi.2023.101402

[10] Zhao, C., Sun, X., Yang, X., Kang, L., Shen, L., Gao, J., & Wang, Y. (2025). Enhancing multi-source cross-domain sentiment classification with generative adversarial networks and transfer learning. Cluster Computing, 28(14), 904. https://doi.org/10.1007/s10586-025-05653-x

[11] Ji, T., Self, N., Fu, K., Chen, Z., Ramakrishnan, N., & Lu, C. T. (2024). Citation forecasting with multi-context attention-aided dependency modeling. ACM Transactions on Knowledge Discovery from Data, 18(6), 1-23. https://doi.org/10.1145/3649140

[12] Xu, S., Zhang, Y., An, X., & Pi, S. (2024). Performance evaluation of seven multi-label classification methods on real-world patent and publication datasets. Journal of Data and Information Science, 9(2), 81-103. https://doi.org/10.2478/jdis-2024-0014

[13] Liu, X., Sun, F., Wang, X., & Sun, T. (2025, February). Legal Core Element Recognition Based on XLNet with Correlation Matrix. In Proceedings of the 2025 3rd International Conference on Communication Networks and Machine Learning (pp. 24-31). https://doi.org/10.1145/3728199.3728204

[14] Lu, Y., Tong, X., Xiong, X., & Zhu, H. (2024). Knowledge graph enhanced citation recommendation model for patent examiners. Scientometrics, 129(4), 2181-2203. https://doi.org/10.1007/s11192-024-04966-9

[15] Wu, H., Zhang, L., Zhu, H., Liu, Q., Chen, E., & Xiong, H. (2025). Examination process modeling for intelligent patent management: a multi-aspect neural sequential approach. ACM Transactions on Management Information Systems, 16(3), 1-23. https://doi.org/10.1145/3712309

[16] Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. Artificial Intelligence Review, 57(4). https://doi.org/10.1007/s10462-023-10652-8

[17] Haghighian Roudsari, A., Afshar, J., Lee, W., & Lee, S. (2022). PatentNet: multi-label classification of patent documents using deep learning based language understanding. Scientometrics, 127(1), 207-231. https://doi.org/10.1007/s11192-021-04179-4

[18] Chen, N., & Liu, X. (2025). Research on World Models for Connected Automated Driving: Advances, Challenges, and Outlook. Applied Sciences, 15(16), 8986. https://doi.org/10.3390/app15168986

[19] Compare the performance of Legal-BERT and XLNet in classification and feature extraction tasks for large-scale patent text mining. https://doi.org/10.5753/jidm.2022.2547

[20] Kang, D. M., Lee, C. C., Lee, S., & Lee, W. (2020, August). Patent prior art search using deep learning language model. In Proceedings of the 24th Symposium on International Database Engineering & Applications (pp. 1-5). https://doi.org/10.1145/3410566.3410597

[21] Kim, H., & Gim, G. (2025). Enhancing Patent Document Similarity Evaluation and Classification Precision Through a Multimodal AI Approach. Applied Sciences, 15(17), 9254. https://doi.org/10.3390/app15179254

[22] Huang, X., Wu, Z., Wang, G., Li, Z., Luo, Y., & Wu, X. (2024). ResGAT: an improved graph neural network based on multi-head attention mechanism and residual network for paper classification. Scientometrics, 129(2), 1015-1036. https://doi.org/10.1007/s11192-023-04898-w

[23] Upadhya, R., & Santosh, T. Y. S. S. (2025, July). LexCLiPR: Cross-Lingual Paragraph Retrieval from Legal Judgments. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 13971-13993). https://doi.org/10.18653/v1/2025.acl-long.683

[24] Tan, S., Zhang, T., Zhao, S., & Zhang, Y. (2023). Self-supervised scientific document recommendation based on contrastive learning. Scientometrics, 128(9), 5027-5049. https://doi.org/10.1007/s11192-023-04782-7

[25] Geng, B. (2022). Text segmentation for patent claim simplification via bidirectional long-short term memory and conditional random field. Computational Intelligence, 38(1), 205-215. https://doi.org/10.1111/coin.12455