Exploring Data-Centric Strategies for French Patent Classification: A Baseline and Comparisons

You Zuo, Kim Gerdes, Houda Mouzoun, Samir Ghamri Doudane, Benoît Sagot

▶ **To cite this version:**

You Zuo, Kim Gerdes, Houda Mouzoun, Samir Ghamri Doudane, Benoît Sagot. Exploring Data-Centric Strategies for French Patent Classification: A Baseline and Comparisons. CORIA-TALN 2023 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Jun 2023, Paris, France. pp.349-365. ⟨hal-04130188⟩

HAL Id: hal-04130188

https://hal.science/hal-04130188v1

Submitted on 20 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Exploring Data-Centric Strategies for French Patent Classification: A Baseline and Comparisons

You Zuo[1,2] Houda Mouzoun[3] Samir Ghamri Doudane[3]
Kim Gerdes[1,4] Benoît Sagot[2]
(1) Qatent, Paris, France
(2) Inria, Paris, France
(3) Institut national de la propriété industrielle, Courbevoie, France
(4) Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay, Orsay, France
firstname.lastname@inria.fr, hmouzoun@inpi.fr, sghamridoudane@inpi.fr, gerdes@lisn.fr

ABSTRACT

This paper proposes a novel approach to French patent classification leveraging data-centric strategies. We compare different approaches for the two deepest levels of the IPC hierarchy : the IPC group and subgroups. Our experiments show that while simple ensemble strategies work for shallower levels, deeper levels require more sophisticated techniques such as data augmentation, clustering, and negative sampling. Our research highlights the importance of language-specific features and data-centric strategies for accurate and reliable French patent classification. It provides valuable insights and solutions for researchers and practitioners in the field of patent classification, advancing research in French patent classification.

RÉSUMÉ

Exploration des stratégies centrées sur les données pour la classification des brevets français : Une base de référence et des comparaisons

Cet article propose une nouvelle approche de classification des brevets français qui s'appuie sur des stratégies centrées sur les données. Nous comparons différentes approches pour les deux niveaux les plus profonds de la hiérarchie IPC : le groupe IPC et les sous-groupes. Nos expériences montrent que les stratégies d'ensemble simples fonctionnent pour les niveaux peu profonds, mais que les niveaux profonds nécessitent des techniques plus sophistiquées telles que l'augmentation de données, le regroupement et l'échantillonnage négatif. Notre recherche met en évidence l'importance des caractéristiques spécifiques à la langue et des stratégies centrées sur les données pour une classification précise et fiable des brevets français. Elle fournit des informations et des solutions précieuses pour les chercheurs et les praticiens dans le domaine de la classification des brevets, en faisant progresser la recherche en classification des brevets en français.

KEYWORDS : Patent Classification, Extreme Multi-label Text Classification, Deep Learning.

MOTS-CLÉS: Classification de Brevets d'Invention, Classification de Textes Multi-labels Extrême, Apprentissage Profond.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 349–365.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

A patent is a legal document that grants its holder the exclusive right to prevent others from making, using, selling, offering for sale, or importing the patented invention within a certain jurisdiction for a specific period of time. Patent databases are valuable sources of information that reflect global innovations and technological developments. The number of patent applications has increased significantly over the past two decades, which can be attributed to various factors such as the increasing importance of technology in society, the role of patents in business valuation, and the globalization of commerce. Due to the large volume of patents and patent documents, patent analysis and management have become complex and time-consuming. Additionally, patents are granted within a specific domain, and patent classification is critical for patent scope and patent law. For the reasons mentioned above, automated patent classification systems have become essential for patent professionals to manage large collections of patents.

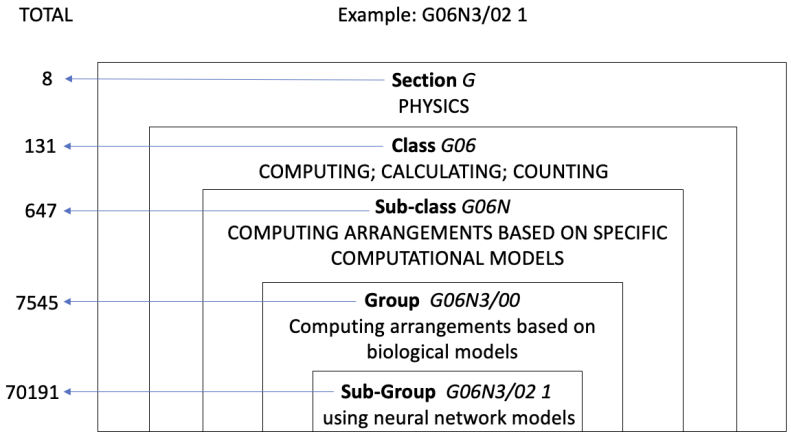

FIGURE 1 – An example of International Patent Classification (IPC) code G06N3/02 1. The IPC scheme is a hierarchical taxonomy with five levels.

The International Patent Classification (IPC) [1] is a widely used standard taxonomy for patent classification. It comprises a hierarchy of five levels, including section, class, subclass, group, and subgroup. The finest level, the subgroup, consists of over 70,000 leaf nodes. An example of an IPC code is shown in Figure 1. When submitting a patent application, the relevant technical fields are indicated by assigning one or multiple IPC codes at the subgroup level. This helps determine the workload of examiners responsible for assessing the application.

However, the IPC undergoes regular minor updates and occasionally major restructuring. Developing an automatic patent classification system is thus challenging due to the constantly evolving technological language and patent syntax, requiring ongoing efforts to address this moving target. In addition, the classification of patent text is a widely studied problem, particularly in languages with large patent markets, such as English, Chinese, and Japanese. However, due to variations in technical fields across different countries and markets, it can be challenging to directly transfer models or data trained on English patents to other languages. This problem is particularly challenging in the case of French patents, in particular presents unique challenges that require a sizable annotated corpus and experimentation with various classification models. Despite the critical importance of French patents,

1. https://www.wipo.int/classifications/ipc/

few studies, such as (Zuo *et al.*, 2022), has focused on patent classification in French, and even those have only reported results on IPC subclass or group levels.

To address this gap, this paper proposes a novel approach to French patent classification that builds upon different strategies aimed at addressing several unique challenges, including input length limitation, label imbalance, and data sparsity. Our approach compares and evaluates different strategies for French patent classification at the two deepest levels of the IPC hierarchy : the IPC group and subgroups. To the best of our knowledge, this is the first comprehensive work specifically focused on patent classification at such deep levels for the French language.

2 Background

A patent is a well-structured document that typically includes several sections, such as a title, abstract, background, brief summary of the invention, detailed description, one or more claims, drawings, and classification information. The International Patent Classification (IPC) is a widely used system for uniformly classifying the content of patents, with over 100 countries currently employing it. The IPC scheme, which was established by the World Intellectual Property Organization (WIPO) [2] in 1971, is hierarchical and serves as the preferred classification system for French patent classification. The IPC system consists of eight general categories of patents, represented by the section level of IPC. These categories are hierarchical and represent the broadest possible areas of technology, providing a starting point for patent classification. Table 1 illustrates the eight most general categories of patents.

Section	Title
A	HUMAN NECESSITIES
B	PERFORMING OPERATIONS ; TRANSPORTING
C	CHEMISTRY ; METALLURGY
D	TEXTILES ; PAPER
E	FIXED CONSTRUCTIONS
F	MECHANICAL ENGINEERING ; LIGHTING ; HEATING ; WEAPONS ; BLASTING
G	PHYSICS
H	ELECTRICITY

TABLE 1 – The eight IPC section categories.

In practice, patent offices assign classification codes to patent applications to accurately describe the subject matter of the invention. They typically assign the most specific level of the International Patent Classification (IPC) to classify the documents in their database. By assigning a specific subgroup level of IPC, broader groups and higher levels of the IPC hierarchy that encompass it can be determined, as the IPC hierarchy is organized in a tree-like structure. This avoids any double classification of "parent" and "child" classes in the IPC hierarchy.

In this work, we focus on exploring the effectiveness of classification methods on groups and subgroups separately.

2. https://www.wipo.int

3 Related Work

Prior research on automated patent classification systems has typically relied on traditional algorithms and feature extraction methods (Verberne & D'hondt, 2011; Yun & Geum, 2020; Wu *et al.*, 2010; Cai & Hofmann, 2007; Qiu *et al.*, 2011). However, designing hand-crafted features can be time-consuming and lead to efficiency problems, making them difficult to apply to large patent collections.

More recently, researchers have turned to deep learning techniques to leverage large-scale training data and generalize well to unseen data. For example, (Grawe *et al.*, 2017) used LSTM on 50 IPC subgroups, while DeepPatent (Li *et al.*, 2018) employed a CNN with skip-gram word embeddings. (Risch & Krestel, 2019) trained fastText word embeddings on full-text patents and then used GRU models on top of these embeddings. Pre-trained language models, such as ULMFiT (Hepburn, 2018) and BERT (Lee & Hsiang, 2019), have also been fine-tuned for this specific task. Comparing the performance of different pre-trained models for multi-label patent classification, XLNet (Yang *et al.*, 2020) was found to outperform other models (Roudsari *et al.*, 2021). Moreover, (Zhang *et al.*, 2022) reduced the uncertainty of classification results through the fusion of multiple patent views. Ensemble techniques have also been explored in patent classification. For instance, (Kamateri *et al.*, 2022) found that an ensemble of classifiers with separate inputs for title-abstract, claims, and description achieved the best performance.

Some researchers have explored mapping patent texts to International Patent Classification (IPC) codes using KNN classification and hybrid methods that combine neural feature encoders with KNN (Cai *et al.*, 2010; Bekamiri *et al.*, 2021, 2022). Another approach is to use a Wide and Deep (WnD) network (Cheng *et al.*, 2016) to combine string-level similarity and semantic embeddings of patent text (Niu & Cai, 2019).

The previous research on patent classification has primarily focused on flat classification, where the classification problem is considered at one specific shallow level of the hierarchy, such as the class or subclass level of the IPC. Only a few studies have addressed the more detailed classification of patent data at the group and subgroup levels. For instance, (Chen & Chang, 2012) proposes a three-phase categorization algorithm using SVM classifiers, (Risch *et al.*, 2020) formulates the hierarchical classification problem as a sequence generation task, and (Zuo *et al.*, 2022) formulates the patent classification problem at deeper levels as an extreme multi-label text classification (XMTC) task.

Despite these efforts, many of these works still rely on feature extraction and model architecture improvements, which only partially address the challenges of accurate patent classification. In contrast, our work focuses on enhancing the quality of the training set and its input format for automatic patent classification, leveraging the same dataset proposed in (Zuo *et al.*, 2022), to improve model performance specifically at the most detailed levels of the IPC hierarchy.

4 Methodology

The task can be framed as follows : given a patent document x, it is assigned with one or more IPC codes $l \in \mathcal{L} = \{l_1, l_2, \ldots, l_L\}$, where L represents the total number of predefined IPC codes at a specific level. Our training set $\mathcal{D} = \{(x_i, y_i) | x_i \in \mathcal{X}, y_i \in \{0,1\}^L, i = 1, 2, \ldots, N\}$ comprises instances x_i and their corresponding labels y_i, with the objective of learning a scoring function f. This function maps an input x_i and a label l to a score $f(x_i, l) \in \mathbb{R}$ and is optimized to maximize

the score when $y_{i,l} = 1$ (i.e., the label l is relevant to the instance x_i) and minimize the score when $y_{i,l} = 0$ (i.e., the label l is irrelevant to the instance x_i).

4.1 Baseline Model

We selected LightXML (Jiang *et al.*, 2021) as our baseline model because it has shown the best performance on INPI-CLS according to (Zuo *et al.*, 2022). LightXML is a transformer-based model specifically designed for multi-label text classification in English. It employs multiple pre-trained language models, such as BERT (Devlin *et al.*, 2018), Roberta (Liu *et al.*, 2019), and XLNet (Yang *et al.*, 2020), and concatenates the representations of the special token [CLS] in the last five hidden states to create a text representation. To handle negative sampling during training, LightXML utilizes a label-recalling network to dynamically sample negative samples during training, followed by a label-ranking network to separate positive from negative labels. We obtained the codes we used from the online extreme classification repository (Bhatia *et al.*, 2016) [3].

Since our corpus is in French, we also used three BERT-like language models specifically designed for French (Martin *et al.*, 2020) or multilingual contexts (Pires *et al.*, 2019; Conneau *et al.*, 2019). The final result of LightXML is the ensemble of the three classifiers. In Table 2, we provide a preliminary comparison of the selected pre-trained models.

Checkpoint	camembert-base (Martin *et al.*, 2020)	bert-base-multilingual-cased (Pires *et al.*, 2019)	xlm-roberta-base (Conneau *et al.*, 2019)
Pre-training Dataset	OSCAR(Suárez *et al.*, 2019)	Wikipedia	cleaned CommonCrawl
Pre-training Tasks	MLM	MLM + NSP	MLM
# Languages	1	104	100
# Parameters	110M	110M	125M
Tokenization	SentencePiece	WordPiece	SentencePiece

TABLE 2 – Overview of pre-trained models we use for LightXML for the classification of French patents.

The methods presented in the remainder of this paper aim to enhance the performance of the baseline approach. We conduct a comprehensive ablation study to offer additional insights for future research on the patent classification problem.

4.2 Weighted Sum Ensemble

An ensemble is a collection of models designed to improve the performance of individual base models by combining their predictions. In this study, we employed the weighted sum ensemble method, which assigns a weight to each base model based on its performance on a validation set. We evaluated the performance of each base model using the $precision@k(k = 1)$ score, which measures the proportion of correct predictions among the top-ranked k categories suggested by the model. This method assumes that some models in the ensemble are more effective than others.

To build our ensemble, we first trained multiple base models using different architectures or training strategies. Next, we evaluated the performance of each model using a separate validation set and

3. http://manikvarma.org/downloads/XC/XMLRepository.html

selected the top-performing ones. Finally, we combined the predictions of the selected models by taking a weighted sum of their output probabilities, where the weight of each model is proportional to its precision@1 score.

4.3 Information Extraction

Due to the maximum input length restrictions of BERT-like models (512 tokens), we explored different approaches to extract the most critical information from patent content.

TextRank (Mihalcea & Tarau, 2004). TextRank is a graph-based ranking model that identifies the most relevant sentences and keywords in text. We considered using TextRank because we observed that using a patent description as input yielded good results. However, the lengthy description text posed a challenge since only the first few hundred words could be entered. To address this issue, we used TextRank [4] to extract the most critical sentences from the patent description as input.

SAO (Subject-Action-Object) Extraction. The SAO structure is a widely used approach in patent analysis for representing technology concepts in a subject-action-object format (Choi *et al.*, 2010; Radauer & Walter, 2010). This structure can be extracted through grammatical processing of patent text and enables a more systematic understanding of the central functional properties of the patent application. For example, consider the sentence, "The super-capacitor electrode further comprising a silane coupling agent." In this sentence, the subject is "super-capacitor electrode," the action is "comprise," and the object is "silane coupling agent." By analyzing many SAO structures in a patent's text, we can extract the underlying functional relationships, identify key features of the invention, and reformulate it as input for classifier models.

To make use of the extracted SAO structures, we typically reformulate them as simple sentences for easier interpretation. For example, the SAO triplet extracted from the above sentence can be reformulated as "The super-capacitors comprise a silane coupling agent." This method [5] has been applied to the claims section of a patent, which is in a strict syntactic format and hence more amenable to rule-based extraction. The extracted SAO structures can then be used as input for our classifier models to improve their performance.

4.4 Vocabulary Enlargement

Another limitation of using pre-trained language models for patent classification is the presence of a large number of scientific and technological terminologies that are seldom encountered in the pre-training corpus, leading to suboptimal model performance. To address this challenge, we propose incorporating external vocabularies from other models. Specifically, we leverage features from sparse classifiers, such as logistic regression with TF-IDF, to extract its 100 most important terms and add them to the vocabulary of the neural encoder. Examples of lemmatized terms can be found in Appendix A.

4. https://summanlp.github.io/textrank/
5. Codes used for SAO extraction : https://github.com/ZoeYou/SAO-extraction

4.5 Sampling Strategies

Dynamic Negative Sampling. LightXML(Jiang *et al.*, 2021) offers a dynamic negative sampling approach that incorporates generative cooperation networks to recall and rank labels from recalled label clusters and dynamically sample negative labels during label ranking. However, for datasets with a small label space, the paper suggests that there is no need to build label clusters and the label recall and re-ranking module degenerates to a linear layer. It can be challenging to determine the appropriate size of the label space to decide whether to set aside the recall and ranking modules. In our study, we compare the classification performance of LightXML with and without the recall and ranking modules at the group and subgroup levels of the IPC.

Oversampling. Class imbalance is a common problem in patent classification datasets that can lead to biased models and poor performance on underrepresented classes. Oversampling is a technique that can address class imbalance by increasing the number of samples in the minority class, thereby providing a better representation of rare concepts. Different oversampling techniques are available for patent classification, such as weighted oversampling and SMOTE (Chawla *et al.*, 2002). In our study, we use the weighted oversampling strategy, where the weight of each label is represented by its inverse frequency in the dataset. We leave the exploration of other oversampling methods for future studies.

4.6 Data Augmentation

The scarcity of training data is a significant challenge for accurate French patent classification. To overcome this obstacle, we explore the possibility of leveraging annotated data from external sources. Although multilingual patent datasets, such as MAREC/IREC (Piroi, 2021) and (Roda *et al.*, 2009; Piroi, 2010; Piroi *et al.*, 2011), have been previously proposed, we opted not to use them because their publication dates significantly differ from our test set. Instead, we obtained a more recent dataset of annotated patent texts published between 2010 and 2019 from the European Patent Office (EPO). This dataset closely resembles our target test set in terms of label distribution and format.

We present a comprehensive analysis of the label distribution across IPC sections in various datasets in Appendix B. Table 3 provides a summary of the statistics of the EPO data [6] that we used.

Language	# Title	# Abstract	# Claims	# Description
French	1,087,313	352,410	507,998	29,539
English	1,082,679	591,045	1,099,062	981,128

TABLE 3 – Statistics of EPO in English and French.

Introduction of EPO French Data. We incorporated French patent data published by the EPO from 2010-2019 into our training set to address the issue of limited training data for French patent classification. Our test set remained the same throughout our work.

Translation of EPO English Data. Introducing French data from the EPO alone did not yield satisfactory performance in deeper levels of IPC classification. To overcome this challenge, we fine-tuned a T5 model (Raffel *et al.*, 2020) using various strategies to translate EPO English patents into French, which augmented the French training data. We utilized the parallel European patent dataset

6. EPO data extracted from EP full-text data for text analytics

EuroPat (Heafield *et al.*, 2022) to train our models. Further details of the fine-tuning experiments can be found in Appendix C.

5 Experiments and Results

5.1 Dataset Description

In this study, we employ the French Patent corpus INPI-CLS (Zuo *et al.*, 2022) as the dataset for our patent classification task. This corpus consists of patents extracted from the internal database of the French National Institute of Industrial Property (INPI) [7] and covers patent texts from 2002 to 2021, including the title, abstract, claims, and description. Each patent in the corpus has been labeled with IPC codes at all levels, from the most general sections to the more specific subgroup labels. In this study, we specifically focus on the group and subgroup levels of IPC, which have been identified as more challenging in previous research. We conducted a time-based split of the corpus to create separate training and test sets. The training set includes patents published between 2002 and 2019, while the test set is composed of patents published between 2020 and 2021.

Dataset	**N**	$\mathbf{L_6}$	$\mathbf{\bar{L}_6}$	$\mathbf{\hat{L}_6}$	$\mathbf{L_8}$	$\mathbf{\bar{L}_8}$	$\mathbf{\hat{L}_8}$
Train	268,254	6,788	2.21	39.52	48,932	2.73	5.48
Test	28,017	4,351	2.20	6.44	19,593	2.64	1.43

TABLE 4 – Basic Statistics of INPI-CLS dataset used for our experiments. L indicates the label count, $\bar{L}$ stands for the average number of IPC labels per patent document, and $\hat{L}$ represents the average number of patent documents per label. The subscripts 6 and 8 indicate the IPC code's length in characters for the IPC's group and subgroup levels, respectively.

5.2 Experimental Setup

We evaluate the performances of models with the rank-based metrics $Precision@K$ and $Recall@K$ $(k = 1, 3, 5)$. $Precision@K$ and $Recall@K$ are calculated for each test patent and then averaged over all the patent documents. Theoretically, each patent document is assigned firstly a primary IPC code, followed by an unlimited number of secondary IPC codes. However, during our evaluation, we did not take into account the order in which the predicted IPC codes are assigned. This aspect is left for future work.

To improve the training efficiency, we fix the maximum input length of each encoder to 128. For other configurations of LightXML, we fix $learning\_rate = 1e-4$ with warm-up, $batch\_size = 16$, and $number\_epoch = 3$.

5.3 Results and Discussion

We evaluated the performance of the LightXML model with baseline configurations at the IPC group and subgroup level, and present the results in Table 5 and Table 6, respectively. Our experiments

7. https://data.inpi.fr/recherche_avancee/brevets

show that the choice of dataset has a significant impact on the performance of the LightXML model for patent classification. In particular, we found that the title+abstract dataset achieved the best performance in terms of $Precision@k$ and $Recall@k$ values for IPC group classification, while the description dataset yielded the highest performance for IPC subgroup classification.

Moreover, our results indicate that the LightXML algorithm performs better at the IPC group level than at the IPC subgroup level. This is because the latter level requires more detailed information about the invention, and the subgroup classes are often imbalanced, making it more challenging to achieve high performance. However, we believe that our approach can be further improved by addressing these issues, such as by incorporating more relevant information or using more advanced modeling techniques.

Dataset	P@1	P@3	P@5	R@1	R@3	R@5
title+abstract	61.08	37.24	26.84	27.67	50.61	60.80
claims	58.60	35.61	25.83	26.54	48.39	58.50
description	58.99	36.30	26.26	26.72	49.33	59.48

TABLE 5 – Baseline performance at IPC group level.

Dataset	P@1	P@3	P@5	R@1	R@3	R@5
title+abstract	12.59	8.80	7.12	4.76	9.99	13.47
claims	15.34	10.65	8.50	5.80	12.09	16.07
description	18.52	12.54	9.87	7.01	14.23	18.67

TABLE 6 – Baseline performance at IPC subgroup level.

Furthermore, in our experiments comparing the performance of classifiers based on different encoders (as shown in Figure 2), we consistently found that `mbert` outperformed other encoders. These results suggest that `mbert` is highly effective in capturing the nuances of the French language in patent documents. One possible reason for its superior performance is that it is trained on a large and diverse corpus, which includes the Wikipedia corpus that contains a vast range of technical terms and topics relevant to patent texts.

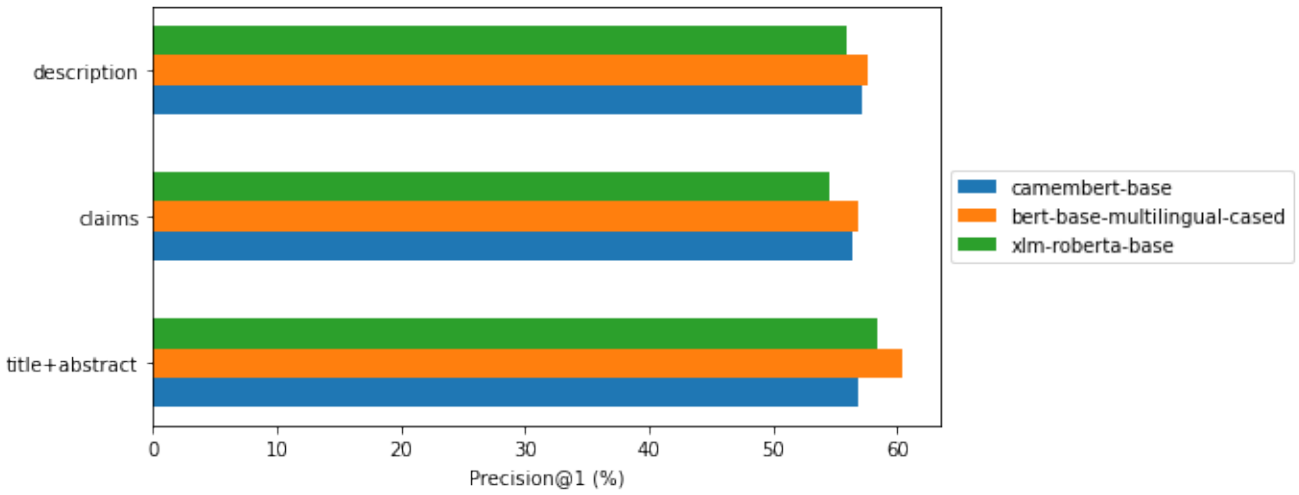

FIGURE 2 – P@1 of classifiers based on different encoders (models trained on IPC group level).

Our experiments comparing different strategies for French patent classification are summarized in Tables 7 and 8. Notably, our results indicate that the weighted sum ensemble approach was not always effective in improving classifier performance, particularly for the subgroup level when the ensemble includes a basic classifier with poor performance. We also observed that methods for extracting the most essential information from patent texts, such as TextRank and SAO extraction, led to a significant decrease in classification performance, particularly for the claims dataset. One possible reason for this is that these methods may have removed important details and nuances from the patent texts that are necessary for accurate classification. In addition, TextRank may have selected sentences from different subsections of the patent description that serve different functions, resulting in a loss of coherence among sentences.

Furthermore, we discovered that enlarging the vocabulary did not enhance the baseline performance. One possible reason for this could be that the additional vocabulary was not well-tuned during the pre-training process, and fine-tuning only the last five feature layers did not provide accurate semantic meaning for the added vocabulary.

Our experimental results for dynamic negative sampling indicate that this technique, proposed in the original LightXML paper, had varying effects on the classifiers' performance at different levels of the IPC hierarchy. Specifically, while this technique led to a decrease in performance for classifiers trained at the IPC group level, it resulted in significant performance improvements for classifiers trained at the subgroup level. These results suggest that in the context of multi-label classification, where long-tail distribution is a significant challenge, clustering techniques and effective negative sampling methods can greatly enhance classification performance.

Moreover, our experiments also showed that the oversampling technique, which involved duplicating minority samples to address class imbalance, did not always improve performance. This suggests that weighted oversampling of long-tail labels may not be an effective approach for addressing label imbalance in French patent classification.

To improve our classification performance, we also investigated the use of supplementary training data, including French data sourced from the EPO, as well as English data translated to French using neural machine translation. We focused on incorporating claims data from the EPO dataset, as it is typically longer and more complete than other sections of patents such as abstracts and descriptions. Our experiments show that the introduction of additional French data from the EPO led to a significant improvement in performance, as demonstrated in Table 8. Furthermore, our models were able to benefit from the increased volume and variety of training examples when augmenting the EPO French data with EPO English data, even when part of the latter dataset was translated from English.

Methods	**title+abstract**	**claims**	**description**
Baseline	61.08	58.60	58.99
Ensemble		63.97	
TextRank		47.11	56.35
SAO Extraction		47.57	
Vocabulary Enlargement	60.98	58.52	58.08
Dynamic Negative Sampling	54.16	48.90	52.38
Oversampling	56.86	53.62	55.47

TABLE 7 – Overall Precision@1 of proposed methods on IPC group level.

Methods	title+abstract	claims	description
Baseline	12.59	15.34	18.52
Ensemble		17.40	
Dynamic Negative Sampling	28.93	26.91	28.18
++data EPO_fr		32.43	
++data EPO_fr & EPO_en by NMT		34.06	

TABLE 8 – Overall Precision@1 of proposed methods on IPC subgroup level.

6 Conclusion

In this paper, we proposed and compared various data-centric approaches for French patent classification. Through extensive experiments, we demonstrated that an ensemble strategy can significantly improve patent classification at shallower levels, such as the IPC group level. However, for deeper levels of classification, such as the IPC subgroup level, where data scarcity and long-tail label distribution are common problems, we recommend using data augmentation techniques, clustering, and negative sampling during the training process to improve model performance.

Our research makes a valuable contribution to the development of automated patent classification systems in the French language, and we hope that our findings will inspire further research in this area. In summary, our work highlights the potential of data-centric strategies to overcome the challenges associated with patent classification and lays the groundwork for future studies in this field.

Acknowledgements

This work was funded and conducted within the framework of a cooperation agreement between the French National Industrial Property Office (INPI) and the French National Institute for Research in Computer Science and Automation (INRIA) for the development of a patent classification model. We would like to express our gratitude to the CLEPS infrastructure at Inria Paris for providing resources and support. The last author's contribution was also supported by his chair in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

BEKAMIRI H., HAIN D. S. & JUROWETZKI R. (2021). Hybrid model for patent classification using augmented SBERT and KNN. *CoRR*, **abs/2103.11933**.

BEKAMIRI H., HAIN D. S. & JUROWETZKI R. (2022). A survey on sentence embedding models performance for patent analysis. *arXiv preprint arXiv :2206.02690*.

BHATIA K., DAHIYA K., JAIN H., KAR P., MITTAL A., PRABHU Y. & VARMA M. (2016). The extreme classification repository : Multi-label datasets and code.

CAI L. & HOFMANN T. (2007). Exploiting known taxonomies in learning overlapping concepts. In *IJCAI*, volume 7, p. 708–713.

CAI Y. L., JI D. & CAI D. (2010). A knn research paper classification method based on shared nearest neighbor. In *NTCIR*.

CHAWLA N. V., BOWYER K. W., HALL L. O. & KEGELMEYER W. P. (2002). Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, **16**, 321–357.

CHEN Y.-L. & CHANG Y.-C. (2012). A three-phase method for patent classification. *Information Processing & Management*, **48**(6), 1017–1030.

CHENG H.-T., KOC L., HARMSEN J., SHAKED T., CHANDRA T., ARADHYE H., ANDERSON G., CORRADO G., CHAI W., ISPIR M. *et al.* (2016). Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, p. 7–10.

CHOI S., LIM J., YOON J. & KIM K. (2010). Patent function network analysis : A function based approach for analyzing patent information. In *19th International conference for the international association of management of technology, Cairo, Egypt, March*, p. 8–11.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv :1911.02116*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

GRAWE M. F., MARTINS C. A. & BONFANTE A. G. (2017). Automated patent classification using word embedding. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, p. 408–411 : IEEE.

HEAFIELD K., FARROW E., VAN DER LINDE J., RAMÍREZ-SÁNCHEZ G. & WIGGINS D. (2022). The europat corpus : A parallel corpus of european patent data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 732–740.

HEPBURN J. (2018). Universal language model fine-tuning for patent classification. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, p. 93–96, Dunedin, New Zealand.

JEHL L. & RIEZLER S. (2018). Document-level information as side constraints for improved neural patent translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1 : Research Track)*, p. 1–12.

JIANG T., WANG D., SUN L., YANG H., ZHAO Z. & ZHUANG F. (2021). Lightxml : Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, p. 7987–7994.

KAMATERI E., STAMATIS V., DIAMANTARAS K. & SALAMPASIS M. (2022). Automated single-label patent classification using ensemble classifiers. In *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, p. 324–330.

LEE J.-S. & HSIANG J. (2019). Patentbert : Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv :1906.02124*.

LI S., HU J., CUI Y. & HU J. (2018). Deeppatent : Patent classification with convolutional neural networks and word embedding. *Scientometrics*, **117**(2), 721–744. DOI : 10.1007/s11192-018-2905-5.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MIHALCEA R. & TARAU P. (2004). Textrank : Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, p. 404–411.

NIU M. & CAI J. (2019). A label informative wide & deep classifier for patents and papers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3438–3443.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How multilingual is multilingual bert ? *arXiv preprint arXiv :1906.01502*.

PIROI F. (2010). Clef-ip 2010 : Retrieval experiments in the intellectual property domain.

PIROI F. (2021). The marec/irec data set. DOI : 10.48436/2zx6e-5pr64.

PIROI F., LUPU M., HANBURY A. & ZENZ V. (2011). Clef-ip 2011 : Retrieval in the intellectual property domain.

QIU X., HUANG X.-J., LIU Z. & ZHOU J. (2011). Hierarchical text classification with latent concepts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 598–602.

RADAUER A. & WALTER L. (2010). Elements of good practice for providers of publicly funded patent information services for smes–selected and amended results of a benchmarking exercise. *World Patent Information*, **32**(3), 237–245.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **21**(1), 5485–5551.

RISCH J., GARDA S. & KRESTEL R. (2020). Hierarchical document classification as a sequence generation task. In *Proceedings of the Joint Conference on Digital Libraries (JCDL)*, p. 147–155.

RISCH J. & KRESTEL R. (2019). Domain-specific word embeddings for patent classification. *Data Technologies and Applications*.

RODA G., TAIT J., PIROI F. & ZENZ V. (2009). Clef-ip 2009 : Retrieval experiments in the intellectual property domain. volume 1175, p. 385–409. DOI : 10.1007/978-3-642-15754-7_47.

ROUDSARI A. H., AFSHAR J., LEE W. & LEE S. (2021). Patentnet : multi-label classification of patent documents using deep learning based language understanding. *Scientometrics*.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)* : Leibniz-Institut für Deutsche Sprache.

VERBERNE S. & D'HONDT E. (2011). Patent classification experiments with the linguistic classification system lcs in clef-ip 2011. volume 1177.

WU C.-H., KEN Y. & HUANG T. (2010). Patent classification system using a new hybrid genetic algorithm support vector machine. *Applied Soft Computing*, **10**(4), 1164–1177. Optimisation Methods Applications in Decision-Making Processes, DOI : https ://doi.org/10.1016/j.asoc.2009.11.033.

YANG Z., DAI Z., YANG Y., CARBONELL J., SALAKHUTDINOV R. & LE Q. V. (2020). Xlnet : Generalized autoregressive pretraining for language understanding.

YUN J. & GEUM Y. (2020). Automated classification of patents : A topic modeling approach. *Computers & Industrial Engineering*, **147**, 106636.

ZHANG L., LIU W., CHEN Y. & YUE X. (2022). Reliable multi-view deep patent classification. *Mathematics*, **10**(23), 4545.

ZUO Y., MOUZOUN H., DOUDANE S. G., GERDES K. & SAGOT B. (2022). Patent classification using extreme multi-label learning : A case study of french patents. In *SIGIR 2022-PatentSemTech workshop-3rd Workshop on Patent Text Mining and Semantic Technologies*.

A Important Features from Sparse Classifiers

We utilized a Logistic Regression model with TF-IDF sparse features, in which each category corresponds to a classifier. In our data preprocessing step, we first removed stop words and then lemmatized each remaining word. We set the dimensionality of the features to 10,000 and excluded words that occurred in more than 90% of the documents. The most important features are represented by the coefficients in the z-equation of the logistic regression, denoted by w_1 to w_n in the following equation :

$$y = \frac{1}{1 + e^{-z}}$$
$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n$$

To select the most important features for each classifier, we took the top five features with the highest coefficient values. We then identified the 100 words that appeared most frequently across all classifiers and added them to our final encoder. These words will be used to improve the model's accuracy in classification tasks.

title+abstract	**claims**	**description**
composition	outil	moteur
moteur	pourcent	véhicule
outil	composition	outil
composé	moteur	eau
machine	véhicule	composé
roue	formule	polymère
eau	machine	roue
signal	atome	composition
formule	dispositif	fibre
polymère	signal	signal
véhicule	roue	combustion
combustion	groupe	machine
fabrication	couche	gaz
produit	acide	acide
gaz	polymère	air
air	gaz	électrique
commande	air	atome
fibre	eau	mesure
fluide	combustion	fluide
mesure	fibre	piston

TABLE 9 – Examples of important features of logistic regression models trained on different patent parts

B Visualization of IPC distributions

Given that different countries have varying priorities in protecting different technical fields, it is important to consider the distance between training and testing data when performing patent classification. Simply adding more data for training without considering the distribution of the data across countries may lead to suboptimal classification results. To demonstrate this phenomenon, we visualize the distribution of labels at the IPC subclass level in Figure B.

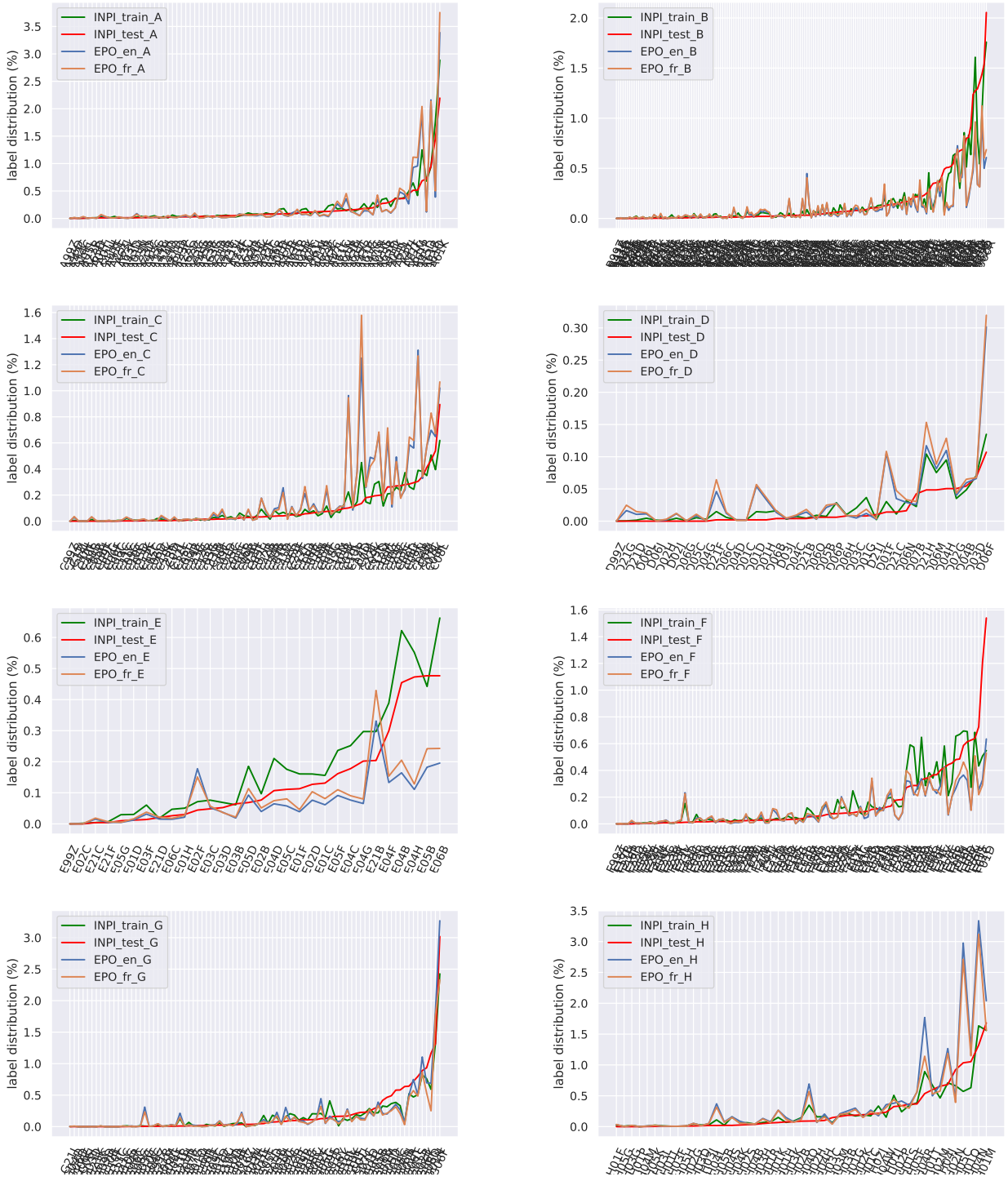

FIGURE 3 – Distribution of labels per IPC section.

C Machine Translation en-fr

In order to improve the classification of French patents, we used a T5 (Raffel *et al.*, 2020) model that was trained on aligned English-French data. Specifically, we leveraged the first release of the EuroPat dataset (Heafield *et al.*, 2022), restricting our analysis to patents that were published after 2010. This allowed us to compile a corpus of 2 million sentence pairs for training our model.

We drew inspiration from (Jehl & Riezler, 2018) to investigate the effectiveness of incorporating special tokens to introducing information as patent sections (<A>, <B>,. . .,<H>) or text types (<title>, <abstract>, <claims>, <description>). Our study compares various methods to determine the optimal approach.

For each approach, we fine-tuned the `t5-base` with 220 million parameters for a single epoch, using a batch size of 16 and a learning rate of 1e-4. The maximum input and output lengths were set to 256. This configuration was selected to optimize the balance between training time and model performance. To evaluate the performance of our model, we relied on the widely used machine translation metric, BLEU score.

	BLEU
Original Text	79.15
IPC1	79.10
Text Type	78.85
IPC1, Text Type	**79.19**

TABLE 10 – Performances of translation models.

We can see that the translation model achieves the best performance when it differentiates between text types and text domains. Therefore, in our main experimental results, we demonstrate the use of the model with special tokens of IPC1s and text types during training for translation and data augmentation.