Patent Classification using Extreme Multi-label Learning: A Case Study of French Patents

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

▶ **To cite this version:**

You Zuo, Houda Mouzoun, Samir Ghamri Doudane, Kim Gerdes, Benoît Sagot. Patent Classification using Extreme Multi-label Learning: A Case Study of French Patents. SIGIR 2022 - PatentSemTech workshop - 3rd Workshop on Patent Text Mining and Semantic Technologies, Jul 2022, Madrid, Spain. ⟨hal-03850405⟩

HAL Id: hal-03850405

https://hal.science/hal-03850405v1

Submitted on 14 Nov 2022

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

Patent Classification using Extreme Multi-label Learning: A Case Study of French Patents

You Zuo
Inria Paris
Paris, France
you.zuo@inria.fr

Houda Mouzoun
Institut national de la propriété industrielle
Paris, France
hmouzoun@inpi.fr

Samir Ghamri Doudane
Institut national de la propriété industrielle
Paris, France
sghamridoudane@inpi.fr

Kim Gerdes
LISN, CNRS and University Paris-Saclay
Orsay, France
gerdes@lisn.fr

Benoît Sagot
Inria Paris
Paris, France
benoit.sagot@inria.fr

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Information systems** → *Document representation*; • **Social and professional topics** → *Patents.*

KEYWORDS

IPC prediction, Clustering and classification, Extreme Multi-label Learning, French

1 INTRODUCTION

The number of patent applications has risen sharply over the past 20 years. As a result, automatic patent classification systems have become essential for patent specialists to analyze and manage large collections of patents. There are several standard classification structures, the most commonly used being the IPC (International Patent Classification) and the CPC (Cooperative Patent Classification), which have hierarchical structures with five different levels: sections, classes, subclasses, groups, and subgroups.

Most previous approaches [1, 6, 11, 12, 20–22] treat the patent classification task as a general text classification task and apply commonly used text classification methods. Some have attempted to implement XML (Extreme Multi-label Learning) methods to handle large numbers of classes [5, 24], but they focus only on the IPC subclasse level, which is far from "extreme" with less than 700 labels.

In this paper, we present a French Patents corpus, named **INPI-CLS**, with IPC labels at all levels, and we test different models at the subclass and group levels on it. Our published French patents are extracted from the INPI[1] internal database, and contain all parts of patent texts (title, abstract, claims, description) published from 2002 to 2021, each patent being annotated with all levels from sections to the IPC subgroup labels. A statistical overview of the data is given in Tables 1 and 2. The training set is constructed from patent documents published before 2020, while the test set includes patents published in 2020 and 2021. In Table 2, N represents the number of patents in the training and test sets. L indicates the label count, $\bar{L}$ stands for the average number of IPC labels of a document. $\hat{L}$ represents the average number of documents per label. The subscripts of 4,6,8 represent respectively IPC's subclass, group, and subgroup levels (4, 6, and 8 correspond to the number of characters used to encode the class). We then compare the performance of the XML (Extreme Multi-label Learning) approaches with other popular NLP methods on our INPI-CLS as well as on the English patent classification benchmark USPTO-2M[12] with 1.9 million training data and 48,000 test data.

We are releasing all relevant code and our French patent classification dataset as open source. The dataset may be used for research purposes and is available under specific licensing requirements detailed in the GitHub repository. [2]

section	title	abstract	description	claims
# items	296 270	295 421	296 216	291 539
# tokens (average)	11	111	4202	725

Table 1: Description of our French corpus INPI_fr

Dataset	N	L_4	$\bar{L_4}$	$\hat{L_4}$	L_6	$\bar{L_6}$	$\hat{L_6}$	L_8	$\bar{L_8}$	$\hat{L_8}$
Train	268254	638	1.73	420.46	6788	2.21	39.52	48932	2.73	5.48
Test	28017	583	1.77	48.06	4351	2.20	6.44	19593	2.64	1.43

Table 2: Basic Statistics of INPI-CLS dataset

2 EXPERIMENTS AND RESULTS

The details of the selected model are listed below:

Logistic Regression The one-vs-all approach is implemented to train a binary logistic regression classifier for each label. We use TF-IDF as input features after applying the snowball stemmer from NLTK and eliminating stop words from the first 1000 words of the input text.

FastText text classification [9] FastText applies a shallow neural network on a hidden variable represented by the average of n-gram character embeddings of input, where the ngram character embeddings are trained under supervision specifically for text classification. We initialize the token representations by the embedding matrix pre-trained on Wikipedia[3] and train a linear classifier for multi-label text classification.

[1] French National Institute of Industrial Property https://www.inpi.fr/fr

[2] https://github.com/ZoeYou/Patent-Classification-2022

[3] https://fasttext.cc/docs/en/pretrained-vectors.html

Bert [4] Just as in the PatentBert[11] experiments, we fine-tune Bert on patent classification. We test the `bert-large` instead of `bert-base` to allow for a comparison with Bert for Patents [23] of the same architecture.

Bert for Patents [23] The model was trained from scratch on more than 100 million English patent documents of USPTO, it leveraged `bert-large` architecture, and built a patent-specific custom tokenizer to hold longer tokens. We took their officially released checkpoint[4] and fine-tuned it on USPTO-2M[12]. [5]

XML-CNN [13] Based on CNN-Kim[10], XML-CNN applies a dynamic maximum pooling to accommodate longer texts and extract location information. It adds a hidden bottleneck layer between the pooling and the output layer, which learns a better representation of the document and improves the prediction accuracy.[6]

Parabel [19] As one of the baseline tree-based algorithms of XML approaches, Parabel firstly learns a balanced binary tree of labels by recursively dividing the label nodes into two balanced clusters until the number of labels in each cluster is less than a given value, and then trains a probabilistic hierarchical multi-label model that generalizes hierarchical softmax to a multi-label setup.[7]

AttentionXML [26] AttentionXML compresses the binary partitioned label tree of [18] into shallower and wider tree to better handle larger label size. A bi-LSTM with multi-label attention mechanism is trained for each level of the tree with the first 500 words of raw text as input. The word representation layers are initialized by Glove[8] for English and French FastText trained on Wikipedia for French patents.

LightXML [8] LightXML applies multiple pre-trained language models. For each model, it concatenates the representations of the `[CLS]` in the last five hidden state as text representation, then trains a label recalling network to dynamically sample negative samples followed by a label ranking network to separate positive from negative labels.[9]

We tested all models on both English and French datasets, except for Bert and Bert for patent, two language models trained on the English corpus. We use ensemble approach for Parabel and AttentionXML with the number of ensemble being 3. For LightXML, we use three different encoders for ensemble. The encoders used for USPTO-2M are

- `bert-base-uncased`[4]
- `roberta-base`[14]
- `xlnet-base-cased`[25]

and

- `camembert-base`[15]
- `bert-base-multilingual-cased`[17]
- `xlm-roberta-base`[2]

for the INPI French patent corpus.

We employ the rank-based metrics Precision@K (P@k(%); k = 1, 3, 5) as evaluation metric following prior Multi-label text classification works. P@K are calculated for each test document and then averaged over all the documents. Due to space limitations, we only show the two main results that we test on the English Benchmark USPTO-2M and our new French dataset INPI-CLS (title+abstract as classifiers' input).

Table 3 demonstrates that LightXML achieves the best results on USPTO-2M, and Bert for Patents achieves comparable performance on it. Compared to the results obtained from [11, 22], we can conclude that we achieve state-of-the-art performance on USPTO-2M with LightXML. It is worth noting that Bert for patent is a large-scale language model specifically pre-trained on patent text from scratch. Bert is very time and resource intensive to train, and we may not be able to find a training corpus of the same size for non-English languages. Yet, the same performance can easily be achieved or even exceeded based on LightXML using ensemble learning with several other off-the-shelf language models including some blocks specifically designed for the XML task. This gives the possibility to obtain higher patent classification performance in languages that do not have as much patent data as English (e.g. French).

For our proposed French patent classification dataset INPI-CLS, LightXML is vastly outperforming the others on both subclass and group levels. LightXML's outstanding performance is attributed to its powerful feature extraction from multiple layers of different transformer encoders and its negative sampling approach on dynamically selecting negative labels from easy to difficult.

Model	P@1	P@3	P@5
Logistic Regression	74.63	41.66	28.82
FastText	73.89	40.55	28.02
bert-large	83.77	46.27	31.37
Bert for Patents	84.31	46.73	31.73
XML-CNN	57.00	31.22	22.08
Parabel	74.43	41.49	28.50
AttentionXML	82.49	45.15	30.82
LightXML	**84.43**	**46.81**	**31.91**

Table 3: Overall Performance on IPC subclass on USPTO-2M (title + abstract)

Model	subclass			group		
	P@1	P@3	P@5	P@1	P@3	P@5
Logistic Regression	65.87	37.63	26.02	49.12	30.32	22.06
FastText	53.76	30.64	21.31	36.21	22.32	16.35
XML-CNN	43.43	25.50	18.23	17.74	10.20	6.96
Parabel	65.13	36.87	25.32	48.93	30.61	22.28
AttentionXML	72.54	40.68	27.63	54.83	33.78	24.49
LightXML	**76.45**	**42.82**	**29.05**	**60.60**	**36.95**	**26.65**

Table 4: Overall Performance on IPC subclass and group on INPI-CLS (title + abstract)

The reasons why the same model performs better on USPTO-2M are 1) USPTO-2M has a much larger dataset for training, almost ten times larger than the French dataset, and 2) by calculating the KL-divergence of the label distributions of the training and test data, we find that the label distributions of the training and test

[4]BERT-for-patents GitHub repository.
[5]The hyperparameters for fine-tuning the two previous language models on patent classification are set as follows: max_sequence_length = 128; epoch = 4; batch_size = 32; learning_rate (Adam) = $3e^{-5}$; binary cross-entropy loss.
[6]We used the code provided by the authors with default values for hyperparameters from https://github.com/siddsax/XML-CNN.
[7]The scripts we utilize are from the Omikuji project. We change CBOW to TF-IDF for better label representation and leave all other hyperparameters as default.
[8]Glove 840B,300d from https://nlp.stanford.edu/projects/glove/
[9]For AttentionXML and LightXML, we used codes provided by the online extreme classification repository.

data are closer for USPTO-2M than that for INPI-CLS. Therefore, we assert that our dataset is "more difficult" to classify.

Different combinations of document parts were tested on our proposed French patent corpus and it was experimentally demonstrated that the combination of title and description achieves the best results (compared to abstract, claims, description and title+abstract). More precisely, when the input constraints are loose (much larger than 128 subwords), there is an improvement of about 4% to 8% on precision@1. However, for methods using pre-trained language models with max_sequence_length set to 128, the precision improvement using title+description compared to title+abstract is less than 2%.

We perform the error analysis by examining the single-label AUC and confusion matrix at $k = 1$. We conclude that weaker models perform worse in learning to classify those labels with less training examples (the AUC of the classifier corresponding to that IPC label is lower). Also, all models have a tendency to mistake "long-tail" labels for those more frequent labels.

3 ONGOING AND FUTURE WORK

Our current focus is on classifying labels with fewer patent examples by using label descriptions or correlations between labels as input information as in [3, 16] and using propensity scored metrics [7] to better evaluate the "long-tailed" labels.

REFERENCES

[1] Juho Bai, Inwook Shim, and Seog Park. 2020. MEXN: Multi-Stage Extraction Network for Patent Document Classification. *Applied Sciences* 10, 18 (2020).

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[3] Kunal Dahiya, Ananye Agarwal, Deepak Saini, Gururaj K, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. SiameseXML: Siamese Networks meet Extreme Classifiers with 100M Labels. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 2330–2340.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Arousha Haghighian Roudsari, Jafar Afshar, Charles Lee, and Wookey Lee. 2020. Multi-label Patent Classification using Attention-Aware Deep Learning Model. 558–559.

[6] Jason Hepburn. 2018. Universal Language Model Fine-tuning for Patent Classification. In *Proceedings of the Australasian Language Technology Association Workshop 2018*. Dunedin, New Zealand, 93–96.

[7] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme Multi-Label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 935–944.

[8] Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. 2021. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. *arXiv preprint arXiv:2101.03305* (2021).

[9] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. arXiv:1607.01759 [cs.CL]

[10] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification.

[11] Jieh-Sheng Lee and Jieh Hsiang. 2019. PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model. *CoRR* abs/1906.02124 (2019). arXiv:1906.02124

[12] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. DeepPatent: Patent Classification with Convolutional Neural Networks and Word Embedding. *Scientometrics* 117, 2 (nov 2018), 721–744.

[13] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep Learning for Extreme Multi-label Text Classification. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017).

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint:1907.11692* (2019).

[15] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of ACL*.

[16] Anshul Mittal, Noveen Sachdeva, Sheshansh Agrawal, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. ECLARE: Extreme Classification with Label Graph Correlations. *CoRR* abs/2108.00261 (2021). arXiv:2108.00261

[17] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502* (2019).

[18] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press.

[19] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising. *WWW '18: Proceedings of the 2018 World Wide Web Conference*, 993–1002.

[20] Subhash Pujari, Annemarie Friedrich, and Jannik Strötgen. 2021. *A Multi-task Approach to Neural Multi-label Hierarchical Patent Classification Using Transformers*. 513–528.

[21] Julian Risch, Samuele Garda, and Ralf Krestel. 2020. Hierarchical Document Classification as a Sequence Generation Task. In *Proceedings of the Joint Conference on Digital Libraries (JCDL)*. 147–155.

[22] Arousha Haghighian Roudsari, Jafar Afshar, Wookey Lee, and Suan Lee. 2021. PatentNet: multi-label classification of patent documents using deep learning based language understanding. *Scientometrics* (2021).

[23] Rob Srebrovic and Jay Yonamine. 2020. *Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery*. Technical Report. Global Patents, Google, https://services.google.com/fh/files/blogs/bert_for_patents_white_paper.pdf.

[24] Pingjie Tang, Meng Jiang, Bryan Ning Xia, Jed W Pitera, Jeffrey Welser, and Nitesh V Chawla. 2020. Multi-label patent categorization with non-local attention-based graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9024–9031.

[25] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]

[26] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems* 32 (2019).