# Taxonomy for Patent Classification: A Step Towards Intelligent Patent Analytics

Elham Motamedi[1,*], Inna Novalija[2] and Luis Rei[3]

[1] University of Primorska, Koper, Slovenia
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3] Jožef Stefan Institute, Ljubljana, Slovenia

**Abstract**

In this study, we proposed a knowledge taxonomy for patents, called KnowMap, which aligns with the CPC schema and reduces the number of classes to 83 at the lowest hierarchical level. We classified patents into these fine-grained classes within a multi-label setting, fine-tuning a distilled version of the RoBERTa model for this purpose. We employed two sampling techniques: (i) random sampling and (ii) conditional random sampling, and found that conditional random sampling led to less pronounced class imbalance, resulting in more generalisable outcomes. Additionally, our results showed higher F1-Macro scores for minority classes, which will be further explored in future work.

**Keywords**

Knowledge Taxonomy, Knowledge Tracking, Patent Classification, Hierarchical Classification, Multi-label Classification

## 1. Introduction

Exploring and leveraging patent-related data is a key task in both scientific and industrial domains. Patent analytics offers a comprehensive view of emerging innovative technologies across various fields. Consequently, business and research initiatives, including European projects, depend on analysing and enhancing patent datasets with specialised innovation-related taxonomies.

One such initiative, the enRichMyData project [1], provides an open software toolbox with practical, robust, and scalable components. This toolbox supports organisations in enriching their data with reference information they may not fully understand and aids data providers in making their data reusable and accessible for data enrichment processes.

In this paper, we propose a novel hierarchical knowledge taxonomy that aligns with the widely used Cooperative Patent Classification (CPC) schema. The CPC classification system organises patents into hierarchical taxonomies, which helps streamline internal processes and enhances the efficiency of search queries. In the first level of the CPC hierarchy, there are nine sections, which are divided into classes, subclasses, groups, and subgroups. Each level of this hierarchy can have several codes ending in approximately 250,000 classification labels [2]. Our taxonomy merges several class entities within the CPC schema based on the scope of the knowledge field and the number of patents associated with each class. This approach addresses the challenge of reducing the large number of class entities in the CPC schema in a way that differs from previous works and provides a benchmark taxonomy for future research. In this study, we also classified patents into the fine-grained classes defined by our proposed taxonomy in a multi-label setting.

*EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands.*
*Corresponding author.
elham.motamedi@ijs.si (E. Motamedi); inna.koval@ijs.si (I. Novalija); luis.rei@ijs.si (L. Rei)
0000-0003-0127-4997 (E. Motamedi); 0000-0001-8587-1638 (L. Rei)

## 2. Related Work

Patent documents contain various types of information, including text [3]. The textual content of a patent is divided into several sections, such as the title, abstract, claim, and description [2]. The title and abstract are shorter than the description but still provide relevant information for classification. Li et al. [4] evaluated various lengths of the abstract and title, finding that using the first 100 words of title and abstract resulted in the best classification performance in their study.

Various classification systems exist for organising patents [5]. In this work, we focus on the CPC schema. Kamateri et al. [2] discussed several potential challenges that artificial intelligence technologies face in patent classification. One such challenge is the extensive number of class labels. As an example, the CPC has around 250,000 labels.

Patent classification is a multi-label classification problem since every patent can belong to several knowledge fields [6, 7]. Given the large number of classes at the lowest level of the taxonomy tree, the performance of automatic models in predicting such fine-grained categories is limited [4, 8, 9]. Several previous studies have focused on higher levels of the hierarchy, limiting classification to broader categories such as sections, classes, or subclasses within the taxonomy [10]. Bekamiri et al. [10] fine-tuned the SBERT model to predict labels at the subclass level (i.e., 663 class labels) using a multi-label formulation. Aroyehun et al. [11] similarly truncated the IPC hierarchy at the subclass level and predicted these labels by transferring knowledge from two higher levels (section and class) to the lower level (subclass). While it remains valuable to use an automatic model that can narrow down applications to higher levels of the taxonomy tree, this approach has limitations. One such limitation is that the choice of target class labels does not depend on the scope of the knowledge area. More established and expansive areas may benefit from directing experts to detailed groups, while less developed areas may be adequately served by broader classifications.

## 3. Methods and Materials

In this work, we developed a knowledge field taxonomy using CPC schema labels. We also classified patents into KnowMap's fine-grained classes by fine-tuning some pre-trained models.

### 3.1. Data Acquisition and Pre-processing

We used the Google Patents Public Datasets on BigQuery [1] and applied preprocessing and sampling techniques. The dataset contains various information, with the abstract offering a brief overview of the patent's novelty and the description providing more detail. For classification, we concatenated the title, abstract, and description, filtering out documents with fewer than 100 words, as prior studies suggest this improves classifier performance [4].

In developing the taxonomy, we considered both the shared knowledge across fields and the distribution of documents within each defined class. To have sufficiently abstract classes, we set a threshold for the minimum number of patents in each detailed group at the lowest level of the hierarchy. Prior to counting the documents in each class, we applied a deduplication step as part of the preprocessing to remove duplicate and near-duplicate texts, which may refer to the same patent [12, 13, 14].

Deduplication was performed using Locality Sensitive Hashing (LSH) [15, 16, 17]. In particular, we used MinHash to approximate the similarities between the documents. Each document was first transformed into a set of n-grams (i.e., in our case 1-grams, 2-grams, and 3-grams). LSH then grouped documents with similar signatures into the same buckets, ensuring that only documents within the same bucket were compared in detail. A Jaccard similarity threshold of 0.9 was set, meaning documents with a similarity score greater than 0.9 were considered duplicates. After deduplication, we generated a dataset sample using two techniques: (i) random sampling and (ii) conditional random sampling,

[1] https://github.com/google/patents-public-data

which included documents in the sample only if their class had fewer than 20,000 documents. Random sampling resulted in 1,092,991 samples, and conditional random sampling, resulted in 1,244,469 samples.

### 3.2. Knowmap Taxonomy Generation

We developed the KnowMap taxonomy by refining the CPC hierarchy and its class entities to create a more abstract representation of patents. Starting from the highest level, we manually merged groups at each level based on shared knowledge and document counts. While all major CPC sections were retained at the first level, groups with fewer than 40,000, 20,000, and 9,000 documents were merged at levels 2, 3, and 4, respectively.

### 3.3. Patent Classification Method and Experimental Setup

We formulated the classification problem as a multi-label problem, in which each document is assigned to one or multiple knowledge fields. In this study, we aimed to classify the patents into the fine-grained classes in the lowest level of the proposed taxonomy (i.e., 83 classes). We used the pre-trained language model *distilroberta-base*, a distilled version of RoBERTa [18, 19]. To adapt this model for our classification task, we fine-tuned it by adding a classification head using the *AutoModelForSequenceClassification* class from the *Hugging Face* library [2]. This classification head processes the hidden state of the first token through a fully connected dense layer. Given that our task is multi-label classification, we applied a sigmoid function to the output logits for each class to obtain probabilities. The implementation of classification method is available online [3].

For model training, we used a learning rate of 4e-5 with a linear scheduler, a weight decay of 0.1, and trained for up to 5 epochs with early stopping. The best checkpoint was selected to prevent overfitting, based on validation accuracy. The sampled datasets were split into training, validation, and test sets with ratios of 0.8, 0.1, and 0.1, respectively. To maintain the class distribution across these sets, we used stratified splitting [4] proposed by Sechidis et al. [20].

## 4. Results and Analysis

In this section, we first present the KnowMap taxonomy and then evaluate the performance of classifiers in categorising patents into fine-grained classes of the taxonomy.

### 4.1. KnowMap Taxonomy

Following the methodology described in Sec. 3.2, we established a hierarchy with the root node as *level 0* and *level 4* as the lowest level. There are nine classes at *level 1* and 83 classes at the lowest level of the hierarchy.

Our hierarchical labels, including merged classes and document counts, are available online [5]. The taxonomy retains the nine CPC sections at the first level, while subsequent levels include merged CPC groups, all detailed in the shared online source.

### 4.2. Classification Results

In this study, we classified patents into the fine-grained classes at the lowest level of the hierarchical taxonomy, which includes 83 labels. To gain further insights into the datasets generated by the two sampling techniques, we analysed the number of samples per class in each dataset. Fig. 1 illustrates this information through box plots for both sampling techniques. The plots highlight the first quartile (Q1), median, third quartile (Q3), minimum, and maximum values for each sampling method.

[2] https://huggingface.co/
[3] https://github.com/elmotamedi/KnowMap-Taxonomy
[4] https://github.com/trent-b/iterative-stratification?tab=readme-ov-file#multilabelstratifiedkfold
[5] https://github.com/elmotamedi/KnowMap-Taxonomy

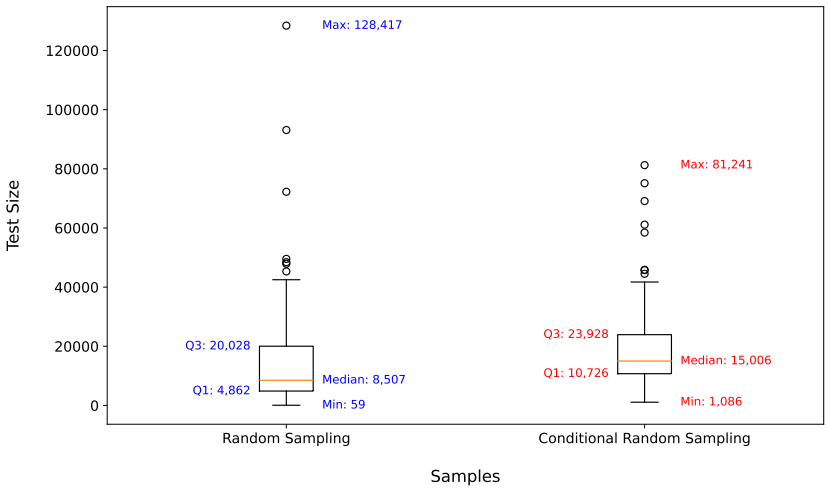



**Figure 1:** Distribution of sample counts per class for test sets generated by the two sampling techniques: (i) random sampling and (ii) conditional random sampling.

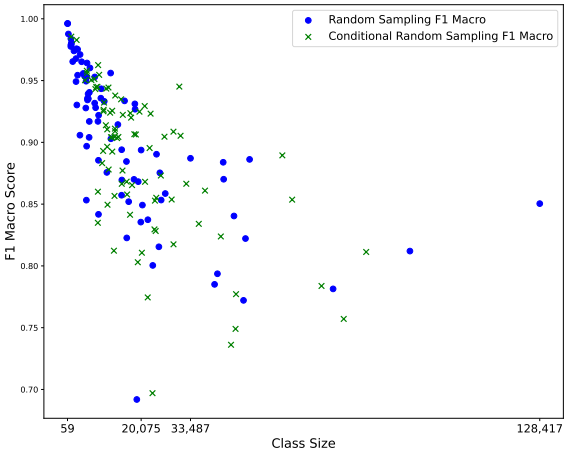
Based on Fig. 1, random sampling resulted in a broader range of document counts per class, with a minimum of 59 samples and a maximum of 128,417 samples per class. The conditional random sampling technique produced a narrower range, with a minimum of 1,086 samples and a maximum of 81,241 samples per class. For our analysis, we categorise classes into three groups: small classes (those in the first quartile), medium classes (those in the second and third quartiles), and large classes (those above the third quartile). With this categorisation, conditional random sampling appears to offer more balanced class distributions compared to random sampling, potentially enhancing the generalisability of classification models trained on this dataset. We present the classification results on both the validation and test sets, applied to the datasets generated by the two sampling techniques in Tab. 1.

**Table 1**
Classification Results for two sampling techniques: (i) random sampling and (ii) conditional random sampling.

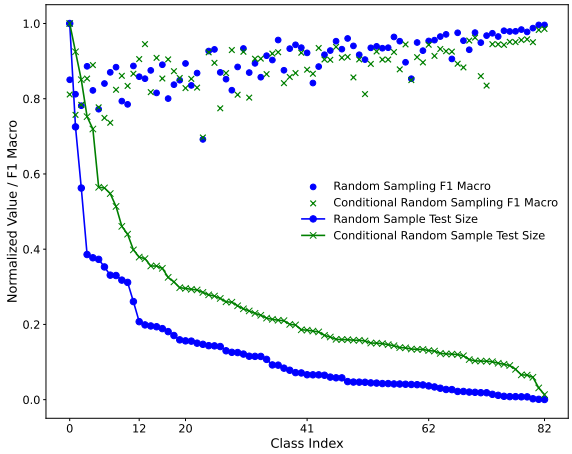
| Metric | Dataset (Random Sampling) | Dataset (Conditional Random Sampling) |
|---|---|---|
| Micro-F1 (Val) | 0.76 | 0.77 |
| Macro-F1 (Val) | 0.86 | 0.83 |
| Micro-F1 (Test) | 0.77 | 0.77 |
| Macro-F1 (Test) | 0.90 | 0.88 |

As observed from the results, the Macro-F1 score is higher than the Micro-F1 score, which may show that the model performs better for minority classes compared to majority classes. To gain more insights into these results, we generated a plot (see Fig.2), showing the F1 scores compared to the number of documents in each class.

The plot shows that the Macro-F1 score is higher for minority classes compared to majority classes for both sampling techniques. The gap between the line plots for random sampling and conditional random sampling (Fig. 2b) highlights the presence of larger classes in the dataset created by conditional random sampling. To provide further insights into the F1-macro scores for small, medium, and large classes across each sample, we have created box plots summarising these scores for each class group and sampling method. The minimum, median, and maximum F1-macro scores for each class group are presented in Fig. 3. Although the increasing trend in F1-macro scores for smaller classes is still visible, it is less pronounced compared to the random sampling technique.

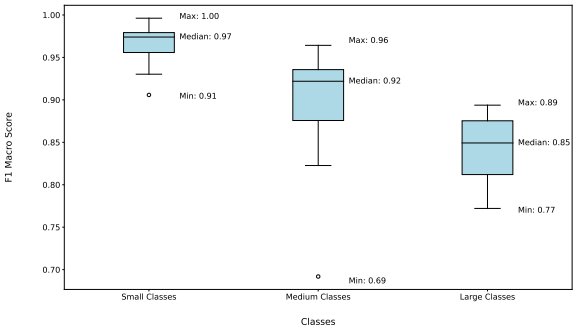



(a) The number of documents in each class vs. F1-macro

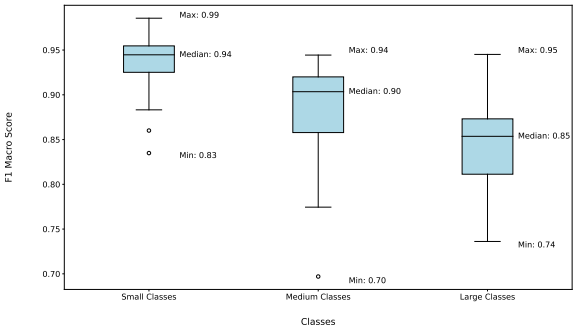



(b) The normalised number of documents + F1-macro for each class

**Figure 2:** Relation of F1 scores and class sizes in the test set for two different sampling methods.




(a) Dataset compiled from random sampling




(b) Dataset compiled from conditional random sampling

**Figure 3:** Box plots of F1-Macro scores for small, medium, and large classes under both random and conditional random sampling techniques.

## 5. Discussion and Conclusions

In this work, we proposed a knowledge taxonomy that aligns with the CPC schema, reducing the number of classes to 83 at the lowest level while ensuring a minimum number of documents for each class in the studied dataset.

We created two datasets from the preprocessed original data using two sampling techniques: (i) random sampling and (ii) conditional random sampling. The conditional random sampling technique resulted in class entities with a minimum of 1,086 samples, substantially more than the minimum sample size achieved through random sampling. This suggests that the results from conditional random sampling may be more generalisable compared to those from random sampling.

In terms of performance, classifiers showed comparable results with both sampling techniques. Both datasets were unbalanced, with the imbalance being less pronounced in the dataset created through conditional random sampling. The classification results exhibited higher F1-Macro scores compared to F1-Micro scores, likely due to the unbalanced nature of the datasets. We conjecture that the lower F1-Macro scores for larger classes may result from the varied nature of documents within those classes, possibly due to imprecise patent assignments in the CPC system or the broader scope of these knowledge fields. Our future research will focus on analysing the classes that the classifier struggles with.

To improve classification performance, we plan to address the dataset imbalance using techniques specifically designed for multi-label classification with long-tailed distributions. Additionally, we aim to explore the use of a larger or alternative pre-trained model to potentially enhance classification results.

## Acknowledgments

This work was supported by the Slovenian Research and Innovation Agency under grant agreements CRP V2-2272, V5-2264, CRP V2-2146 and the European Union through enrichMyData EU HORIZON-IA project under grant agreement No 101070284.

## References

[1] enRichMyData consortium, enrichmydata project, https://enrichmydata.eu, ????

[2] E. Kamateri, M. Salampasis, E. Perez-Molina, Will AI solve the patent classification problem?, World Patent Information 78 (2024) 102294. URL: https://doi.org/10.1016/j.wpi.2024.102294. doi:10.1016/j.wpi.2024.102294.

[3] M. Suzgun, L. Melas-Kyriazi, S. K. Sarkar, S. D. Kominers, S. M. Shieber, The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications, in: 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks, NeurIPS, 2023, pp. 1–39. arXiv:2207.04043.

[4] S. Li, J. Hu, Y. Cui, J. Hu, DeepPatent: patent classification with convolutional neural networks and word embedding, Scientometrics 117 (2018) 721–744. doi:10.1007/s11192-018-2905-5.

[5] J. C. Gomez, M. F. Moens, A survey of automated hierarchical classification of patents, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8830 (2014) 215–249. doi:10.1007/978-3-319-12511-4_11.

[6] A. H. Roudsari, J. Afshar, C. C. Lee, W. Lee, Multi-label patent classification using attention-aware deep learning model, in: Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020, 2020, pp. 558–559. doi:10.1109/BigComp48618.2020.000-2. arXiv:arXiv:1910.01108.

[7] G. Jung, J. Shin, S. Lee, Impact of preprocessing and word embedding on extreme multi-label patent classification tasks, Applied Intelligence 53 (2023) 4047–4062. doi:10.1007/s10489-022-03655-5.

[8] C. J. Fall, A. Törcsvári, K. Benzineb, G. Karetka, Automated categorization in the international patent classification, ACM SIGIR Forum 37 (2003) 10–25. doi:10.1145/945546.945547.

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

[10] H. Bekamiri, D. S. Hain, R. Jurowetzki, PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT, Technological Forecasting and Social Change 206 (2024) 123536. doi:10.1016/j.techfore.2024.123536.

[11] S. T. Aroyehun, J. Angel, N. Majumder, A. Gelbukh, A. Hussain, Leveraging label hierarchy using transfer and multi-task learning: A case study on patent classification, Neurocomputing 464 (2021) 421–431. doi:10.1016/j.neucom.2021.07.057.

[12] G. Costa, A. Cuzzocrea, G. Manco, R. Ortale, Data De-duplication : A Review Data De-duplication : A Review, Learning structure and schemas from documents (2011). doi:10.1007/978-3-642-22913-8.

[13] N. Kandpal, E. Wallace, C. Raffel, Deduplicating Training Data Mitigates Privacy Risks in Language Models, in: International Conference on Machine Learning, Baltimore, volume 162, 2022, pp. 10697–10707.

[14] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, N. Carlini, Deduplicating Training Data Makes Language Models Better, Proceedings of the Annual Meeting of the Association for Computational Linguistics 1 (2022) 8424–8445. doi:10.18653/v1/2022.acl-long.577. arXiv:2107.06499.

[15] B. Gyawali, L. Anastasiou, P. Knoth, Deduplication of scholarly documents using locality sensitive hashing and word embeddings, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association, 2020, pp. 894–903.

[16] O. Jafari, P. Maurya, P. Nagarkar, K. M. Islam, C. Crushev, A Survey on Locality Sensitive Hashing Algorithms and their Applications, ACM Computing Surveys (2021). arXiv:2102.08942.
[17] M. Aydar, S. Ayvaz, An improved method of locality-sensitive hashing for scalable instance matching, Knowledge and Information Systems 58 (2019) 275–294. doi:10.1007/s10115-018-1199-5.
[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019).
[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arxiv 2019, arXiv preprint arXiv:1910.01108 (2019).
[20] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, in: D. Gunopulos, T. Hofmann, D. Malerba, M. Vazirgiannis (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 145–158. doi:10.1007/978-3-642-23808-6_10.

## A. Online Resources

The sources referenced in this paper, including the proposed taxonomy and the classification implementation, are available at:

- KnowMap taxonomy and classification implementation
- Multi-label stratified K-fold implementation
- Google Patents Public Datasets on BigQuery
- Hugging Face library