Classification of Patents Into Knowledge Fields: Using a Proposed Knowledge Mapping Taxonomy (KnowMap)

Elham Motamedi
elham.motamedi@upr.si
University of Primorska
Koper, Slovenia

Inna Novalija
inna.koval@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Luis Rei
luis.rei@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Various platforms, including patent systems and repositories like GitHub and arXiv, support knowledge dissemination across domains. As knowledge increasingly spans multiple disciplines, there is a need to track innovations that intersect various fields. Despite available data, a comprehensive knowledge taxonomy for effectively tracking innovations across domains is lacking. Developing such a taxonomy and employing automated classification methods will enhance the ability to track shared knowledge.

In this work, we first developed a knowledge taxonomy based on the CPC schema. We formulated the classification of textual data into defined knowledge fields as a multi-label problem. Then, we evaluated the effectiveness of the classification models by fine-tuning pre-trained transformer language models. The multi-label framework enables the tracking of knowledge trends at the intersection of various disciplines.

Keywords

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

1 Introduction

According to the World Intellectual Property Organisation (WIPO), a patent is an exclusive right granted for an invention, providing legal protection to the inventor while simultaneously benefiting society by making the invention publicly accessible [1]. Each year, patent offices receive numerous patent applications that need to be processed [13].To ensure the novelty of patent applications, inventors should also be able to search existing patents. Organising patents with unique codes in a hierarchical structure aids efficient retrieval and aligns with natural human navigation, starting from broad categories and narrowing down to specifics[21]. Among these hierarchical structures, the CPC system is widely recognised [6]. The CPC codes are organised as a taxonomy, meaning that each entity in the lower level is the detail group of the parent. A patent can be assigned to one or more labels by the experts in patent offices [8, 18]. 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 [11]. An example of the hierarchical structure of CPC code is provided in Tab. 1.

Table 1: Example of a sequence of codes across different levels of the CPC hierarchy

CPC	Code	Title
Section	H	Electricity
Class	H03	Electronic circuitry
Subclass	H03C	Modulation
Group	H03C3/00	Angle modulation
Subgroup	H03C3/005	Circuits for asymmetric modulation

The CPC schema's top level has only nine sections, but the number of groups increases substantially at lower levels. In this study, we created a knowledge field taxonomy by merging CPC's detailed classes into a more abstract representation. This taxonomy not only serves as a framework for knowledge representation but also offers a benchmark for patent classification systems. While some studies address the issue of numerous class labels by excluding less-represented classes or truncating hierarchies [24], a consistent benchmark taxonomy has been lacking. Since our proposed knowledge taxonomy aligns with the CPC schema, it is able to provide a benchmark for future studies, facilitating the comparison of different models.

In summary, our paper's contribution is the proposal of a knowledge field taxonomy, KnowMap, which aligns with the widely used CPC schema. The KnowMap merged several class labels within the CPC schema based on the scope of the knowledge field and the number of patents associated with each class. The KnowMap taxonomy is available online [2]. In this study, we also performed a classification task to categorise patents into the fine-grained classes defined by our proposed taxonomy.

2 Related Work

Patent documents contain various types of information, including text, diagrams, plots, and references to other patents or scientific publications [20]. The textual content of a patent is divided into several sections, such as the title, abstract, claim, and description [11]. The title and abstract are shorter than the description but still provide relevant information for classification. Li et al. [15] 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 [6]. In this work, we focus on the CPC schema. The hierarchical representations help organise patents and facilitate efficient searching. Kamateri et al. [11] 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 IPC contains approximately 86,000 classes, while the CPC has around 250,000.

Patent classification is a multi-label classification problem since every patent can belong to several knowledge fields [18,

[1] https://www.wipo.int/portal/en/

[2] https://github.com/elmotamedi/KnowMap-Taxonomy

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.19

10]. Given the large number of classes at the lowest level of the taxonomy tree, the performance of automatic models in predicting such granular categories is limited. Various models have been used to classify patents in a multi-label setting, ranging from classical machine learning models to deep learning models [15, 5, 8]. 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 [3]. Bekamiri et al. [3] fine-tuned the SBERT model to predict labels at the subclass level (i.e., 663 class labels) using a multi-label formulation. They achieved F1-score of 66%, outperforming previous studies that used the same datasets. Aroyehun et al. [1] 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), achieving a precision score of 0.53. While it remains valuable for patent office experts to use an automatic model that can narrow down applications to higher levels of the taxonomy tree, this approach has limitations and challenges. One such challenge 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 taxonomy and classified patents into fine-grained classes by fine-tuning pre-trained models. Below, we outline the methods and materials used.

3.1 Patent Collection and Preprocessing

The dataset used in our experiments is the Google Patents Public Datasets on BigQuery [3]. Each patent has several pieces of information, including the publication number, application number, CPC code, title, abstract, and detailed description. We have expanded the dataset to include the titles associated with each CPC code from Espacenet. [4]. In this study, we focused on the textual data. We generated the input text by concatenating the title, followed by the abstract, and then the description. We included only those documents where the concatenated text is at least 100 words long. Previous studies have examined various lengths of textual data and found that using the first 100 words often results in higher performance for classification tasks [15].

To create a hierarchical structure where we have enough documents among leaf-node labels (i.e., avoiding scenarios where one group contains only a few hundred documents while others contain hundreds of thousands as an example), we needed to count the number of documents which fall into the defined categories. As a preprocessing step before counting, we performed de-duplication, which involved removing duplicate and near-duplicate textual data [4, 12, 14].

Due to the large size of the dataset, we employed MinhHash Locality Sensitive Hashing (LSH) as a deduplication method to efficiently identify similar documents [7, 9, 22]. Specifically, we used MinHash to approximate the Jaccard similarities between sets of n-grams within the documents. MinHash is particularly advantageous for large datasets because it supports parallel computation, enhancing scalability [2]. We set the similarity threshold at 0.9, meaning that documents with a Jaccard similarity of 90% or higher were considered duplicates. To generate the hash signatures in MinHash, we used 128 permutations. For the n-gram representation, we used a range of 1 to 3, incorporating 1-grams, 2-grams, and 3-grams.

3.2 Refining Hierarchical Structure Through Group Merging

The hierarchical structure of the CPC groups was refined at each level of the tree. We started with nine sections at the top level (i.e., *level 1*), which were preserved. At subsequent levels (i.e., *level 2* to *level 4*), groups were merged by manual analysis based on shared knowledge and the number of documents. Groups with relatively few documents (i.e., groups with fewer than 40,000 for *level 2*, 20,000 for *level 3*, and 9,000 for *level 4*) were combined with other groups at the same level that shared similar knowledge. As an example, at the subclass level of the CPC hierarchy, "A01B" (i.e., Soil working) and "A01C" (i.e., Planting, Sowing, Fertilising) represent related steps in agricultural practices, as both are foundational processes in land preparation and management. We merged them into a single group labelled "Soil working and planting," resulting in 162,567 patents in this category. The refinement continued until the fine-grained classes contained at least 9,000 documents.

3.3 Text Classification

We formulated the classification problem as a multi-label problem, in which each document can be assigned to 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). To balance performance and computational cost given the large size of the dataset, We used the pre-trained language models *distilroberta-base*, a distilled version of RoBERTa [16, 19], and *all-MiniLM-L6-v2*, a version of MiniLM fine-tuned for semantic similarity [22, 17]. The pre-trained models were fine-tuned for the downstream task by adding a classification head. The classification head takes the hidden state of the first token from the model and processes it through a fully connected dense linear layer, followed by a dropout layer for regularisation and a tanh activation function for non-linearity. Since our task is multi-label classification, the output logits for each class are converted into probabilities using a sigmoid function.

For model training, we used a learning rate of 4e-5 with a linear scheduler and a weight decay of 0.1. To prevent overfitting, the best checkpoint was selected based on evaluation metrics on the validation set. We trained the model for up to 5 epochs with early stopping criteria based on validation accuracy. The dataset, consisting of 1,092,991 samples randomly selected after deduplication, was split into training, validation, and test sets with ratios of 0.8, 0.1, and 0.1, respectively. To preserve the ratio of samples per class in training, validation, and test sets, we used stratified splitting [5].

3.4 Classification Evaluation

The F1-score is a common metric for classification tasks. We report both Micro-F1, averaged across all instances, and Macro-F1, averaged across all classes.

4 Results and Analysis

In this section, the results are presented in two parts. First, we present our proposed KnowMap taxonomy. Then, we report the

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

[4] https://worldwide.espacenet.com/

[5] https://github.com/trent-b/iterative-stratification?tab=readme-ov-file#multilabelstratifiedkfold

performance of classifiers in categorising patents into the fine-grained classes of this taxonomy.

4.1 The Proposed Knowledge Mapping Taxonomy (KnowMap)

The taxonomy, along with the associated CPC sections, classes, subclasses, groups, and subgroups are provided in the shared online source. An example of detailing the knowledge field of *soil working and planting* within the broader knowledge field of *human necessities* is illustrated in Fig. 1.

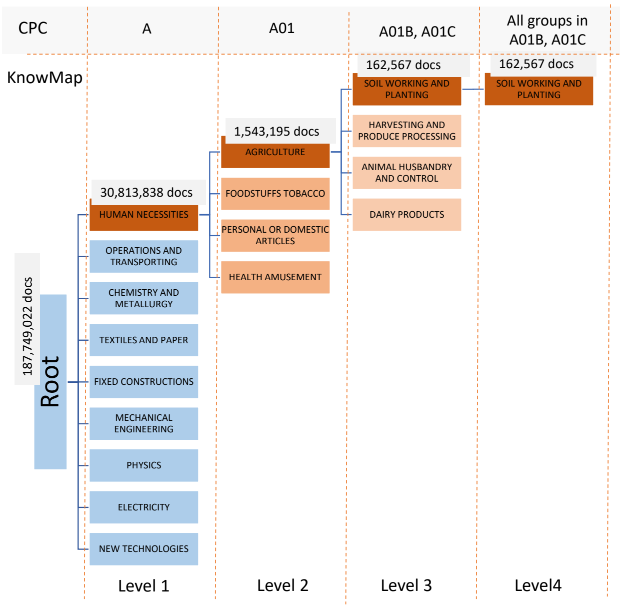

Figure 1: An example of a branch extension in KnowMap from the root to the lowest level, showing the association of KnowMap classes with corresponding CPC classes at each level.

4.2 Classification Results

The classification task in this study was to classify patents into 83 fine-grained classes within our proposed KnowMap taxonomy. The dataset comprised 1,092,991 documents, which were split into the train, validation, and test sets with a ratio of 0.8, 0.1, and 0.1 respectively. We preserved the ratio of samples per class in all three sets with stratified splitting. The average number of documents in the train set, validation set, and test sets are presented in Tab. 2.

Table 2: Overview of sample metrics: total number of samples, average number of samples per class, and normalised average number of samples per class across training, validation, and test sets.

Set	Total	Avg/ class	Normalised Avg
Train	1,092,991	132,202	0.012
Val	874,372	16,476	0.012
Test	218,619	15,543	0.012

Table 3: Classification Results

Metric	RoBERTa	SBERT
Micro-F1 (Val)	0.76	0.76
Macro-F1 (Val)	0.86	0.86
Micro-F1 (Test)	0.77	0.76
Macro-F1 (Test)	0.90	0.90

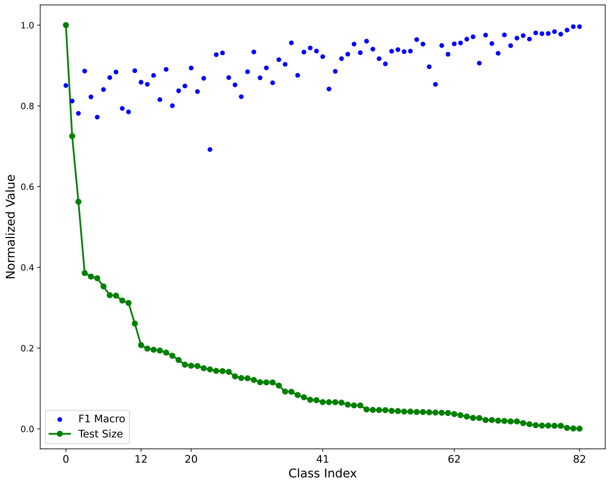

Figure 2: Normalised test size along with F1 Macro scores for each class. The x-axis represents class indices. The y-axis shows normalised values for test size and F1 Macro scores (blue dots).

We demonstrated the experimental results on the two classification models *RoBERTa* and *SBERT* in Tab. 3.

As observed from the results, the Macro-F1 score is higher than the Micro-F1 score, which may indicate 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 along with the normalised number of documents for each class in the test set. We used normalised values to allow both F1 scores and class sizes to be displayed in a single figure, facilitating better comparison.

The plot shows that the Macro-F1 score is higher for minority classes than for majority classes, also indicating that random sampling led to an unbalanced dataset. The imbalanced sample likely caused the higher Macro-F1 score relative to Micro-F1, reflecting poorer performance in the majority classes. Future work will focus on using balancing techniques when sampling to address this issue and enhance model performance.

When looking more closely at the lowest F1-Macro scores, we found that the bottom 10 classes were all leaves under the *chemistry and metallurgy* section. Moreover, the highest F1-Macro scores (0.996) were achieved by the two classes in the *textiles and paper* section, followed by all 17 leaves from the *physics* section. We suspect this performance difference may be due to greater variation in the textual data of *chemistry and metallurgy* class compared to *physics* and *textiles and paper*, leading to more variation between the training and test sets. Analysing this variation in detail remains a task for future work. Additionally, we believe future work could benefit from adapting the classifier to a hierarchical structure, prioritising correct predictions at higher levels before refining predictions at the leaf level. In our current approach, the classifier does not account for the hierarchy and predicts all leaves directly.

5 Discussion and Conclusions

In this work, we proposed a knowledge field taxonomy, KnowMap, which aligns with the widely used CPC schema. The taxonomy consists of 83 groups at the lowest level, with fine-grained classes containing a minimum of 9,000 samples from the original Google Patents Public Dataset after preprocessing. KnowMap serves as a benchmark taxonomy, addressing a gap in the existing literature.

From the preprocessed original dataset, we randomly selected 1,093,151 samples to fine-tune pre-trained RoBERTa and SBERT models for downstream tasks. However, the random sampling resulted in an unbalanced dataset, which contributed to higher Macro-F1 scores compared to Micro-F1 scores. To enhance classification results, we plan to create a balanced dataset from the original data. Additionally, we aim to use larger models than those used in this study to further improve the fine-tuning process.

6 Future Work

Several knowledge platforms, such as news sites and GitHub, host various types of information shared online. In future work, we aim to incorporate these sources to extend and enhance the knowledge taxonomy's coverage. For example, the All Science Journal Classification (ASJC), which organises research publications by subject area, can be used to identify alignments with the existing taxonomy. This taxonomy alignment can then be further analysed to determine whether to merge or split classes at various levels. Beyond patents, we plan to evaluate the classifier on other data, using domain adaptation methods to transfer knowledge from the labelled patent domain to those with limited or no labels. Large language models (LLMs) could further aid in evaluating the classifier's performance across different domains. Recent research has shown the potential of LLMs to augment or even replace human-labeled training data with labels generated by these models [23].

Moreover, we plan to enhance the classification task by balancing the dataset using balancing techniques for multi-label problems and leveraging larger pre-trained models. we will also closely examine the different knowledge fields to better understand the variations in classifier performance across them.

Acknowledgements

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] Segun Taofeek Aroyehun, Jason Angel, Navonil Majumder, Alexander Gelbukh, and Amir Hussain. 2021. Leveraging label hierarchy using transfer and multi-task learning: A case study on patent classification. *Neurocomputing*, 464, 421–431. DOI: 10.1016/j.neucom.2021.07.057.

[2] Mehmet Aydar and Serkan Ayvaz. 2019. An improved method of locality-sensitive hashing for scalable instance matching. *Knowledge and Information Systems*, 58, 2, 275–294. ISBN: 1011501811995. DOI: 10.1007/s10115-018-1199-5.

[3] Hamid Bekamiri, Daniel S. Hain, and Roman Jurowetzki. 2024. PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT. *Technological Forecasting and Social Change*, 206, June, 123536. DOI: 10.1016/j.techfore.2024.123536.

[4] Gianni Costa, Alfredo Cuzzocrea, Giuseppe Manco, and Riccardo Ortale. 2011. Data De-duplication : A Review Data De-duplication : A Review. *Learning structure and schemas from documents*, January. ISBN: 9783642229138. DOI: 10.1007/978-3-642-22913-8.

[5] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. 2003. Automated categorization in the international patent classification. *ACM SIGIR Forum*, 37, 1, 10–25. DOI: 10.1145/945546.945547.

[6] Juan Carlos Gomez and Marie Francine Moens. 2014. 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, 215–249. DOI: 10.1007/978-3-319-12511-4_11.

[7] Bikash Gyawali, Lucas Anastasiou, and Petr Knoth. 2020. 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, 894–903.

[8] Arousha Haghighian Roudsari, Jafar Afshar, Wookey Lee, and Suan Lee. 2022. PatentNet: multi-label classification of patent documents using deep learning based language understanding. *Scientometrics*, 127, 1, 207–231. DOI: 10.1007/s11192-021-04179-4.

[9] Omid Jafari, Preeti Maurya, Parth Nagarkar, Khandker Mushfiqul Islam, and Chidambaram Crushev. 2021. A Survey on Locality Sensitive Hashing Algorithms and their Applications. *ACM Computing Surveys*. eprint: 2102.08942.

[10] Guik Jung, Junghoon Shin, and Sangjun Lee. 2023. Impact of preprocessing and word embedding on extreme multi-label patent classification tasks. *Applied Intelligence*, 53, 4, 4047–4062. DOI: 10.1007/s10489-022-03655-5.

[11] Eleni Kamateri, Michail Salampasis, and Eduardo Perez-Molina. 2024. Will AI solve the patent classification problem? *World Patent Information*, 78, June, 102294. DOI: 10.1016/j.wpi.2024.102294.

[12] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating Training Data Mitigates Privacy Risks in Language Models. In *International Conference on Machine Learning, Baltimore* number 1. Vol. 162, 10697–10707.

[13] Jong Wook Lee, Won Kyung Lee, and So Young Sohn. 2021. Patenting trends in biometric technology of the Big Five patent offices. *World Patent Information*, 65, March, 102040. DOI: 10.1016/j.wpi.2021.102040.

[14] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 8424–8445. eprint: 2107.06499. DOI: 10.18653/v1/2022.acl-long.577.

[15] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117, 2, 721–744. ISBN: 1119201829. DOI: 10.1007/s11192-018-2905-5.

[16] Yinhan Liu et al. 2019. Roberta: a robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[18] Arousha Haghighian Roudsari, Jafar Afshar, Charles Cheolgi Lee, and Wookey Lee. 2020. Multi-label patent classification using attention-aware deep learning model. In *Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020*, 558–559. ISBN: 9781728160344. eprint: arXiv:1910.01108. DOI: 10.1109/BigComp48618.2020.000-2.

[19] Victor Sanh, L Debut, J Chaumond, and T Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arxiv 2019. *arXiv preprint arXiv:1910.01108*.

[20] Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. 2023. 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* number NeurIPS, 1–39. eprint: 2207.04043.

[21] Christoph Trattner, Philipp Singer, Denis Helic, and Markus Strohmaier. 2012. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *ACM International Conference Proceeding Series*, 0–7. ISBN: 9781450312424. DOI: 10.1145/2362456.2362474.

[22] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33, 5776–5788.

[23] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24) Article 303. Association for Computing Machinery, Honolulu, HI, USA, 21 pages. ISBN: 9798400703300. DOI: 10.1145/3613904.3641960.

[24] Junghwan Yun and Youngjung Geum. 2020. Automated classification of patents: A topic modeling approach. *Computers and Industrial Engineering*, 147, July, 106636. DOI: 10.1016/j.cie.2020.106636.